UNIVERSITE MONTPELLIER 1

THESE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITE MONTPELLIER 1

Discipline : *Bioinformatique*
Formation Doctorale : *Interface Chimie-Biologie*
Ecole Doctorale : *Sciences Chimiques et Biologiques pour la Santé*

Présentée et soutenue publiquement par

Xavier BROCHET

Le 18 décembre 2008

Titre :

CONCEPTION ET INTEGRATION D'UN SYSTEME D'INFORMATION DEDIE A L'ANALYSE ET A LA GESTION DES SEQUENCES REARRANGEES DES RECEPTEURS D'ANTIGENES AU SEIN D'IMGT® : APPLICATION A LA LEUCEMIE LYMPHOÏDE CHRONIQUE

JURY

M. Gérard Lefranc, Professeur, Université Montpellier 2, Président du jury

Mme Marie-Paule Lefranc, Professeur, Université Montpellier 2, Directeur de thèse

M. Frédéric Davi, Enseignant chercheur, Hôpital Pitié Salpêtrière et Université Pierre et Marie Curie, Rapporteur

M. Patrice Marche, Directeur de recherche INSERM, Université Joseph Fourier, Rapporteur

Mme. Lina Yip Sonderegger, Maître assistant, Université de Genève, Examinateur

Mme Véronique Giudicelli, Ingénieur d'études, Université Montpellier 2, Examinateur

UNIVERSITE MONTPELLIER 1

THESE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITE MONTPELLIER 1

Discipline : *Bioinformatique*
Formation Doctorale : *Interface Chimie-Biologie*
Ecole Doctorale : *Sciences Chimiques et Biologiques pour la Santé*

Présentée et soutenue publiquement par

Xavier BROCHET

Le 18 décembre 2008

Titre :

CONCEPTION ET INTEGRATION D'UN SYSTEME D'INFORMATION DEDIE A L'ANALYSE ET A LA GESTION DES SEQUENCES REARRANGEES DES RECEPTEURS D'ANTIGENES AU SEIN D'IMGT® : APPLICATION A LA LEUCEMIE LYMPHOÏDE CHRONIQUE

JURY

M. Gérard Lefranc, Professeur, Université Montpellier 2, Président du jury

Mme Marie-Paule Lefranc, Professeur, Université Montpellier 2, Directeur de thèse

M. Frédéric Davi, Enseignant chercheur, Hôpital Pitié Salpêtrière et Université Pierre et Marie Curie, Rapporteur

M. Patrice Marche, Directeur de recherche INSERM, Université Joseph Fourier, Rapporteur

Mme. Lina Yip Sonderegger, Maître assistant, Université de Genève, Examinateur

Mme Véronique Giudicelli, Ingénieur d'études, Université Montpellier 2, Examinateur

REMERCIEMENTS

Je tiens à remercier tout d'abord les rapporteurs de cette thèse, Frédéric Davi et Patrice Marche. Je remercie également les autres membres du jury Lina Yip Sonderegger et Véronique Giudicelli, pour leur examen critique de mon travail et le Professeur Gérard Lefranc pour ces conseils avisés et la présidence de ce jury.

Je voudrais exprimer toute ma reconnaissance à mon directeur de thèse, le Professeur Marie-Paule Lefranc, pour le temps qu'elle m'a accordé, son aide et son soutien tout au long de ce travail de thèse.

Un grand merci à tous les membres de l'équipe IMGT actuelle et à celles et ceux qui sont partis entre-temps qui m'ont accueilli avec tant de gentillesse et m'ont accompagné au cours de ces trois dernières années: Bellahcene Fatena, Duroux Patrice, Ginestoux Chantal, Giudicelli Véronique, Jabado-michaloud Joumana, Lane Jérôme, Ehrenmann François, Folch Géraldine, Gemrot Elodie, Regnier Laëtitia, Wu Yan, Lucas Odile, Servier Emmanuel et Garapati Phani vijay. Merci pour leur soutien, leur amitié et pour leur bonne humeur.
Je voudrais particulièrement remercier Véronique, qui m'a aidé et conseillé tout au long de cette thèse, même pendant ses jours de vacances et ses week-end. Merci également à Chantal pour son aide lors de la réalisation des superbes illustrations du manuscrit.

Je remercie également les institutions qui m'ont apportées leur soutien financier durant ces trois années : le Ministère de l'Enseignement Supérieur et de la Recherche et le CNRS.

Mes affectueuses pensées vont à mes parents qui m'ont soutenu tout au long de mes études (qui furent longues), à Justine ma compagne des bons moments et de ceux plus difficiles, comme ceux de la fin d'une thèse, ainsi qu'à mes frérots et à tous mes amis qui m'ont soutenu et encouragé.

Merci à tous pour votre soutien et votre affection qui m'ont permis de mener à bien cette thèse.

PUBLICATIONS

Brochet, X., Lefranc, M.-P. and Giudicelli, V. IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. Nucl. Acids Res., 36, W503-508 (2008). PMID: 18503082

Duroux, P., Kaas, Q., **Brochet, X.**, Lane, J., Ginestoux, C., Lefranc, M.-P. and Giudicelli, V IMGT-Kaleidoscope, the Formal IMGT-ONTOLOGY paradigm. Biochimie, 90, 570-583 (2008). PMID: 17949886

Lefranc, M.-P., Giudicelli, V., Ginestoux, C., Jabado-Michaloud, J., Folch, G., Wu, Y., Bellahcene, F., Gemrot, E., **Brochet, X.**, Lane, J., Regnier, L., Ehrenmann, F., Lefranc, G. and Duroux, P. IMGT®, the international ImMunoGeneTics information system®. Nucleic Acids Res (2008).

TABLE DES MATIERES

INTRODUCTION

Le développement intensif de techniques expérimentales performantes, en particulier le séquençage de l'ADN, aboutit à l'accumulation de nombreuses données concernant la séquence, la structure ou encore la fonction des gènes et des protéines. Plus généralement, ces développements ont permis l'augmentation rapide et continuelle de la masse d'informations disponible dans le domaine des sciences fondamentales et biomédicales. Parallèlement, le développement des réseaux et du matériel informatique (capacité de stockage, vitesse d'accès aux données) a permis la création d'outils performants et conviviaux pour gérer et simplifier l'accès aux informations pour la communauté scientifique.

Dès lors, le problème est de faire face à ces volumes sans cesse croissants de données biologiques diverses (sur les gènes, les ARN, les protéines) provenant d'organismes les plus variés, et de réaliser que l'accumulation et l'accès à un grand nombre de données n'ont d'intérêt que si l'on peut les interpréter et les exploiter afin d'en déduire de nouvelles connaissances. La bioinformatique est la discipline qui permet de faire la liaison entre la biologie et l'informatique. Elle joue un rôle prépondérant dans le développement des biotechnologies et de la recherche, dans les sciences biologiques et médicales. Les outils bioinformatiques s'imposent pour stocker les données de séquences, les trier, les analyser, les classer, les comparer, faire émerger des liens, identifier des corrélations ou prédire la structure et la fonction des molécules. Ils sont devenus incontournables aux traitements, à l'analyse des volumes importants de données essentielles à la compréhension des êtres vivants, de leur évolution, des pathologies et la découverte de nouveaux traitements.

Cependant, la qualité de l'information ne peut être restituée qu'au niveau de structures dites 'spécialisées' (bases de données, outils d'analyses) capables de prendre en compte les spécificités des données, de les gérer, de les analyser afin de produire des informations conformes aux règles dans un domaine précis de la biologie. Malgré l'existence de notions communes en biologie, chaque domaine a ses propres règles pour décrire les données [1]: il est donc indispensable de créer les structures nécessaires à la gestion des données spécialisées dans un domaine précis pour préserver et améliorer la qualité de l'information distribuée.

Parmi tous les domaines de la biologie qui engendrent le plus d'informations, le système immunitaire est l'un des plus complexes et des plus importants pour la recherche médicale. Il

a pour fonction de protéger l'hôte, de prévenir et de vaincre les infections et les cancers. Alors qu'une immunité naturelle ou innée est présente chez les invertébrés et les vertébrés, l'immunité acquise ou adaptative, qui permet la reconnaissance spécifique et la mémorisation des agents pathogènes, est rencontrée exclusivement chez les vertébrés. Cette spécificité et cette mémoire immunitaire, caractérisent l'immunité adaptative. L'immunogénétique étudie la génétique de ces molécules impliquées dans la réponse du système immunitaire adaptatif, parmi lesquelles se trouvent les immunoglobulines (IG) ou anticorps, les récepteurs des cellules T (TR) et les protéines du complexe majeur d'histocompatibilité (MHC). Au cours de la synthèse des IG et TR, des mécanismes complexes de réarrangements somatiques [2] engendrent une grande diversité dans les séquences d'IG et TR [3, 4].

Jusqu'à présent, de tels mécanismes ne sont connus que pour la synthèse des IG et TR. Par ailleurs, des phénomènes d'hypermutations somatiques se produisent spécifiquement au niveau des régions codant les domaines variables des IG. Le nombre de séquences potentiellement produites pour un seul individu, compte tenu de tous ces mécanismes de diversité, est estimé à 10^{12}, ce qui permet d'obtenir des IG et des TR susceptibles de répondre potentiellement à tous les antigènes. Les IG, sont présentes à la surface des lymphocytes B ou sécrétées par les plasmocytes, alors que les TR sont exprimés à la surface des lymphocytes T [3, 4]. La connaissance et la description de la structure des récepteurs d'antigènes sont primordiales en recherche fondamentale (évolution du système immunitaire, étude des répertoires), dans le développement de l'ingénierie des anticorps et des anticorps thérapeutiques [5] (single chain Fragment variable (scFv), phage displays, librairies combinatoires, anticorps chimériques et humanisés) et en recherche clinique (rôle dans la pathologie, tests diagnostiques et pronostiques). Les lymphocytes B et les immunoglobulines membranaires sont impliqués dans de nombreuses maladies du système immunitaire, non seulement lorsqu'il s'agit de défauts des réponses humorales ou d'hyperproduction de tel ou tel isotype (HyperIgM et HyperIgE, par exemple), mais également dans des syndromes lymphoprolifératifs B (lymphomes, leucémie aiguë lymphoïde ou LAL, leucémie lymphoïde chronique ou LLC etc…). L'étude des récepteurs d'antigènes a permis des progrès considérables dans la compréhension des propriétés immunobiologiques des cellules leucémiques, en particulier de la LLC: par exemple, le degré de mutation des IG du clone malin en rapport avec l'état clinique des patients [6] et la mise en évidence d'IG très similaires, stéréotypées, impliquant un rôle essentiel de l'antigène dans la leucémogenèse [7, 8].

L’extrême diversité des séquences des IG et TR rend difficile leur description rigoureuse par des outils classiques d’analyse de séquences d’ADN ou de gestion des données dans des bases généralistes. Il est nécessaire de disposer d’outils, de structures et de règles de standardisation spécifiques pour rendre compte de cette complexité.

IMGT®, the international ImMunoGeneTics information system® (système d'information international en ImMunoGénéTique) [9] (publication 3) a été créé par Marie-Paule Lefranc en 1989 à Montpellier. IMGT® est spécialisé dans la gestion des données de séquences et de structures 3D des IG, TR et MHC des vertébrés [10]. L'exactitude et la cohérence des données d’IMGT® sont fondées sur IMGT-ONTOLOGY [11-13] (publication 2), la première ontologie en immunogénétique et immunoinformatique. L’IMGT-ONTOLOGY permet la gestion des connaissances en immunogénétique en se fondant sur sept axiomes, «Identification», «DESCRIPTION», «CLASSIFICATION», «NUMEROTATION», «LOCALIZATION», «ORIENTATION» et «OBTENTION», qui postulent que chaque objet, chaque processus et leurs relations doivent être identifiés, décrits, classés, numérotés, localisés, orientés de façon à standardiser toutes les informations. Cette standardisation a permis de décrire l’organisation modulaire des récepteurs et des chaînes des IG, TR et MHC, et les structures des domaines protéiques: V-DOMAIN [14], C-DOMAIN [15], et G-DOMAIN [16]. La numérotation unique IMGT de chaque domaine permet pour la première fois de comparer de manière standardisée les séquences et les structures quelque soit le récepteur, le type de chaîne ou l’espèce.

Les standards IMGT® basés sur IMGT-ONTOLOGY, ont été approuvés par l’Organisation Mondiale de la Santé et par l’Union Internationale des Sociétés Immunologiques (OMS-IUIS), le comité de la nomenclature des IG et TR [17, 18] et constituent le cadre de référence pour mise en place des structures de gestion et d’analyse des données des récepteurs d’antigènes.

Les objectifs de cette thèse étaient de concevoir et d'intégrer au sein d’IMGT® une nouvelle composante dans le domaine médical, avec la mise en place d’outils capables d’analyser et de gérer les données de séquences de patients atteints de pathologies, afin de caractériser à plus long terme les processus et les mécanismes moléculaires impliqués. Le projet a été initié dans le cadre d’une collaboration internationale avec des équipes cliniques spécialisées dans la leucémie lymphoïde chronique et impliquées dans l’European Research Initiative on CLL (ERIC).

Le premier objectif était la création de nouvelles fonctionalités dans l'analyse des séquences réarrangées d'IG et TR et leur intégration à l'outil d'analyse IMGT/V-QUEST afin de fournir à la communauté scientifique un système standardisé et convivial en ligne. Dans un premier temps, après une étude de l'algorithme des modules existants dédiés à la détermination des gènes et allèles, j'ai réécrit et intégré le programme en un seul module de même langage JAVA, apportant de ce fait une simplification dans la gestion des tâches et des améliorations dans les performances de l'outil. De nouvelles fonctionnalités ont été apportées au cours de la réécriture de l'outil, afin de fournir une analyse détaillée et complète de la séquence: description et localisation des mutations de la V-REGION (transversion ou transition, mutation silencieuse ou non), classification des mutations des acides aminés (AA) selon leur hydropathie, leur volume, et leurs caractéristiques physicochimiques, identification de la position des 'hot spots' dans le gène et l'allèle germline le plus proche. Le programme IMGT/V-QUEST est maintenant capable de détecter les insertions et/ou les délétions potentielles qui se produisent lors de la synthèse des IG et des TR, et de fournir une évaluation de la fonctionnalité des séquences analysées (productives ou non productives). Il peut également analyser jusqu'à 50 séquences en un seul lot et fournit, selon le choix de l'utilisateur, une analyse individuelle des séquences, ou une analyse multiple avec alignements des séquences qui expriment le même gène variable V et le même allèle. Enfin, il offre une interface paramétrable.

Le second objectif était la création de la base de données IMGT/CLL-DB, dédiée d'une part à la gestion des données concernant les patients atteints de LLC, et d'autre part aux résultats standardisés d'IMGT/V-QUEST issus de l'analyse détaillée des IG exprimées à la surface des lymphocytes B de ces patients. Le résultat de l'intégration d'IMGT/V-QUEST dans IMGT/CLL-DB fournit un système d'information dédié à la LLC mais dont les composantes sont génétiques. Chaque séquence d'IG stockée dans la base de données est ainsi reliée aux résultats d'analyse de l'outil IMGT/V-QUEST, aux informations relatives aux patients (description de la pathologie, résultats des analyses, de la cytogénétique…), et aux informations relatives aux échantillons biologiques prélevés (tissu prélevé, date du prélèvement etc…). Un programme a été développé pour gérer les nouvelles entrées dans la base, sous contrainte des règles de contrôle et de protection des données, avec la mise en place d'un vocabulaire contrôlé défini en collaboration avec les partenaires du projet. Finalement, une interface web a été développée donnant accès à l'ensemble des informations d'IMGT/CLL-DB aux partenaires du projet. Cette interface permet la recherche des séquences selon de nombreux critères et affiche les résultats sous forme de tableaux à onglets

qui permettent de visualiser les données par thème. Cette interface donne également accès à des outils d'analyse interactifs qui permettent d'obtenir et de visualiser les alignements de séquences exprimant le même gène V, les données sous différents formats, ou de visualiser l'ensemble des résultats d'IMGT/V-QUEST sous le format standard.

Le **chapitre 1** présente le système immunitaire adaptatif, les règles de description des IG et TR, les règles de numérotation de leurs domaines V et les mécanismes de la diversité du répertoire des IG. Le **chapitre 2** présente dans les grandes lignes (diagnostic, facteurs de pronostic…), de la leucémie lymphoïde chronique (LLC), pathologie étudiée par les cliniciens partenaires du projet. Le **chapitre 3** décrit l'outil d'analyse des séquences nucléotidiques réarrangées des IG et des TR, IMGT/V-QUEST, et l'algorithme mis en place afin de caractériser et de décrire les séquences réarrangées. Les résultats issus de cette analyse seront également présentés. Le **chapitre 4** présente le système d'information composé de l'outil IMGT/V-QUEST et de la base de données IMGT/CLL-DB, dédié à la gestion des séquences associées aux données concernant les patients atteints de LLC. Nous verrons ensuite les structures et les procédures que nous avons mises en place pour la gestion des données, les règles de contrôles et de standardisation grâce à un vocabulaire contrôlé. Enfin nous décrirons l'interface web donnant accès aux informations de la base de données.

CHAPITRE 1

Le système immunitaire adaptatif et les récepteurs d'antigènes

Le système immunitaire protège l'hôte contre les infections et ses propres cellules devenues tumorales ou 'étrangères' au soi. La défense de l'hôte nécessite différents systèmes de reconnaissance et une grande variété de mécanismes effecteurs pour détecter et détruire des pathogènes très variés dans les diverses localisations où ils se présentent. L'immunité innée (ou naturelle) est constituée des barrières naturelles anatomiques, physiologiques, phagocytaires et inflammatoires. Elle constitue la première ligne de défense mais elle ne dispose pas de la capacité de reconnaître certains pathogènes, ni d'installer un état d'immunité spécifique qui préviendrait les récidives. L'immunité adaptative (ou spécifique), observée uniquement chez les vertébrés, est basée sur la sélection clonale de lymphocytes B et T dotés de récepteurs de l'antigène, immunoglobulines (IG) ou anticorps et les récepteurs T (TR), hautement diversifiés, qui permettent au système immunitaire de reconnaître de manière spécifique les antigènes. Dans cette réponse immunitaire adaptative, les lymphocytes spécifiques de l'antigène prolifèrent et se différencient en cellules effectrices aptes à éliminer les pathogènes. Cette réponse adaptative peut non seulement éliminer les pathogènes, mais en même temps, par la sélection clonale, générer des lymphocytes mémoires, ce qui autorise une réponse plus rapide et plus efficace en cas de réinfection. La régulation des réponses immunitaires est l'objet principal de la recherche médicale en immunologie, en vue soit de les supprimer ou de les atténuer si elles sont intempestives, soit de les stimuler dans la prévention de maladies infectieuses. La création du système d'information IMGT® a permis de définir et de mieux comprendre les caractéristiques des IG et des TR, principaux acteurs de la reconnaissance antigénique.

Dans ce chapitre, nous rappellerons brièvement les deux composantes, l'immunité humorale et cellulaire, du système immunitaire adaptatif, nous décrirons ensuite les mécanismes à l'origine de la diversité des IG, protéines membranaires des lymphocytes B ou sécrétées par les plasmocytes, qui jouent un rôle essentiel dans la LLC, pathologie utilisée comme modèle dans le cadre de ma thèse, de IMGT/CLL-DB. Enfin, nous indiquerons les règles définies, au sein d'IMGT® pour la description de ces protéines, et les règles de numérotation de leurs domaines variables (V) basées sur les concepts d'IMGT-ONTOLOGY [11-13].

1.1 Immunité humorale et cellulaire

Le système immunitaire adaptatif est souvent décrit par ses deux composantes, l'immunité humorale (sécrétion d'anticorps) et l'immunité cellulaire (cytolyse des cellules infectées ou cancéreuses) qui font intervenir différentes cellules et molécules (Figure 1.1) et défendent l'organisme contre les pathogènes extracellulaires et intracellulaires.

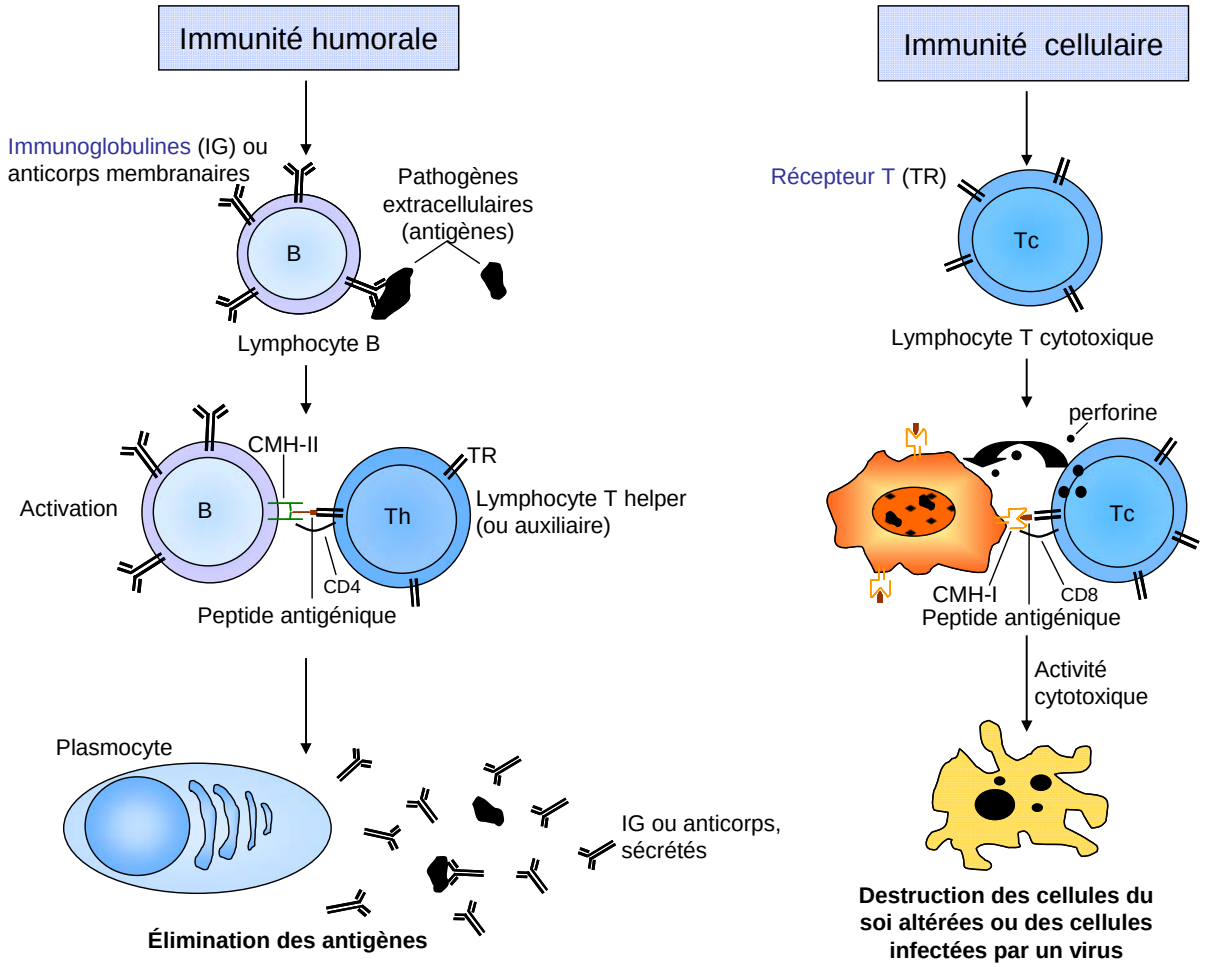


Figure 1.1: Les deux composantes du système immunitaire adaptatif: immunité humorale et cellulaire. Les cellules de la lignée B comprennent les lymphocytes B et les plasmocytes. Les cellules de la lignée T comprennent les lymphocytes T, qui peuvent être des lymphocytes T helper ou auxiliaire (Th), ou des lymphocytes T cytotoxiques (Tc). Les protéines caractéristiques de la réponse immunitaire adaptative comprennent les immunoglobulines (IG), les récepteurs T (TR), le complexe majeur d'histocompatibilité de classe I (CMH-I) et de classe II (CMH-II) (avec la permission de IMGT®, http://www.imgt.org).

L'immunité humorale agit contre les pathogènes (bactéries et virus) circulant dans le sang et la lymphe et est basée sur la reconnaissance spécifique de déterminants antigéniques (ou épitopes) par les sites anticorps (ou paratopes) des domaines variables (V) des IG. Les IG existent en tant que protéines membranaires à la surface des lymphocytes B ou sont sécrétées par les plasmocytes, cellules qui représentent le stade de différenciation terminal des cellules

B (Figure 1.1). Les lymphocytes B se développent à partir de cellules-souches dans la moelle osseuse. Durant la synthèse des IG, des mécanismes de réarrangements de l'ADN permettent la génération d'une énorme diversité de lymphocytes B (10^{12} chez l'homme), le facteur limitant étant le nombre de lymphocytes B génétiquement programmé pour une espèce donnée. Toutes les IG exprimées à la surface d'un lymphocyte B sont identiques et ont la particularité d'avoir la même spécificité de reconnaissance d'un antigène. Les lymphocytes B matures circulent alors dans la lymphe et gagnent les organes lymphoïdes secondaires (ganglions lymphatiques, rate). Dans les organes lymphoïdes secondaires, un lymphocyte B qui reconnaît un antigène pour lequel il est spécifique, est activé et prolifère et après contact avec un lymphocyte T spécifique, se différencie soit en plasmocyte qui sécrète des IG, soit en lymphocyte B mémoire (Figure 1.1). Les anticorps sécrétés par les plasmocytes neutralisent le pouvoir infectieux des pathogènes en se liant à leurs antigènes de surface, qui interfère avec leur capacité à se fixer aux cellules de l'hôte (anticorps neutralisants). Les anticorps peuvent entraîner également une destruction de l'agent pathogène, par le complément (complement-dependent cytotoxicity ou CDC) ou par une cellule cytotoxique (antibody-dependent cellular cytotoxicity ou ADCC) (IMGT®, http://www.imgt.org). Enfin en recouvrant les pathogènes, les anticorps favorisent la phagocytose par les macrophages (opsonisation). Les lymphocytes B mémoires issus de lymphocytes B déjà sélectionnés et ayant subi l'expansion clonale et la commutation de classe possèdent à leur surface des IG membranaires. Ces lymphocytes B mémoires ont une durée de vie beaucoup plus longue que les plasmocytes, et pourront être activés et se différencier en plasmocytes lors d'une nouvelle rencontre avec le même antigène.

L'immunité cellulaire est chargée de la défense de l'organisme vis-à-vis des cellules infectées par des agents pathogènes intracellulaires (virus) ou des cellules cancéreuses. Les lymphocytes T sont issus des cellules-souches de la moelle osseuse (comme les lymphocytes B) mais qui se différencient ensuite dans le thymus. Les mécanismes de synthèse des TR, semblables à ceux des IG sont basés sur des réarrangements de l'ADN, qui génèrent une grande diversité combinatoire de TR et de lymphocytes T (potentiellement 10^{12} chez l'homme). Chaque lymphocyte T exprime des TR d'une seule et même spécificité. Les lymphocytes T sont sélectionnés par une double sélection négative et positive qui permet premièrement d'éliminer les lymphocytes T fortement autoréactifs spécifiques des peptides du soi, et deuxièmement de sélectionner les lymphocytes T qui reconnaissent des peptides du non soi. L'interaction TR/peptide aboutit à l'expansion clonale du lymphocyte T impliqué dans la reconnaissance spécifique de l'agent pathogène et à la différenciation des clones en

lymphocytes T effecteurs, cytotoxiques (Tc) ou auxiliaires ou helper (Th) et en lymphocytes T mémoire. Au sein d'une cellule, les protéines subissent une dégradation par le protéasome et les peptides de 8 à 10 acides aminés issus de cette protéolyse sont ensuite transportés à la surface de la cellule pour être présentés par l'intermédiaire du CMH-I. Ainsi, les cellules saines présentent à leur surface des peptides du soi qui ne déclenchent pas de réaction immunitaire. Au contraire, les cellules étrangères, les cellules tumorales ou infectées par un virus ou un autre agent pathogène présentent à leur surface des peptides du non soi, qui entraînent de manière spécifique une activation des lymphocytes T cytotoxiques (Tc) qui reconnaissent de manière spécifique ce complexe peptide-MHC-I (pMHC-I) et qui les détruisent. Les cellules T auxiliaires ou helper (Th) sécrètent des cytokines qui stimulent la réaction immunitaire auprès des autres cellules. Elles contribuent notamment à l'activation des cellules présentatrices d'antigènes (CPA professionnels) lesquelles comprennent les cellules dendritiques, les macrophages et les lymphocytes B. Ceux-ci prolifèrent et se différencient en plasmocytes dans les organes lymphoïdes secondaires. On parle alors de réponse humorale T-dependante (Figure 1.1), par opposition à la réponse humorale T-indépendante qui ne requiert pas l'aide des Th, et dans laquelle les lymphocytes B se différencient en plasmocytes sans contact au préalable avec un Th. Au sein d'une cellule CPA, les protéines exogènes (antigènes extracellulaires) sont dégradées dans les vésicules d'endocytose en peptides de 10 à 15 acides aminés, lesquels sont présentés à la surface de la cellule par l'intermédiaire du CMH-II, dont le G-DOMAIN forme un sillon ou un 'groove' où se loge le peptide. C'est le complexe pMHC-II qui est reconnu de manière spécifique par le TR des Th $CD4^+$.

1.2 Synthèse des chaînes d'immunoglobuline

Les immunoglobulines (IG ou anticorps) sont exprimées en surface des cellules B matures et des cellules B mémoires ou sont sécrétées par les plasmocytes. Les différentes étapes de la différenciation des cellules souche hématopoïétiques en cellules B matures qui expriment des IgM et des IgD, se produisent dans la moelle osseuse indépendamment de l'antigène (figure 1.2). Les étapes finales de la différenciation, des cellules B matures en cellules mémoires ou en plasmocytes qui expriment ou sécrètent des immunoglobulines de diverses classes ou de sous-classes se produisent dans le centre germinatif des organes lymphoïdes secondaires, et sont tributaires de l'antigène (Figure 1.2).

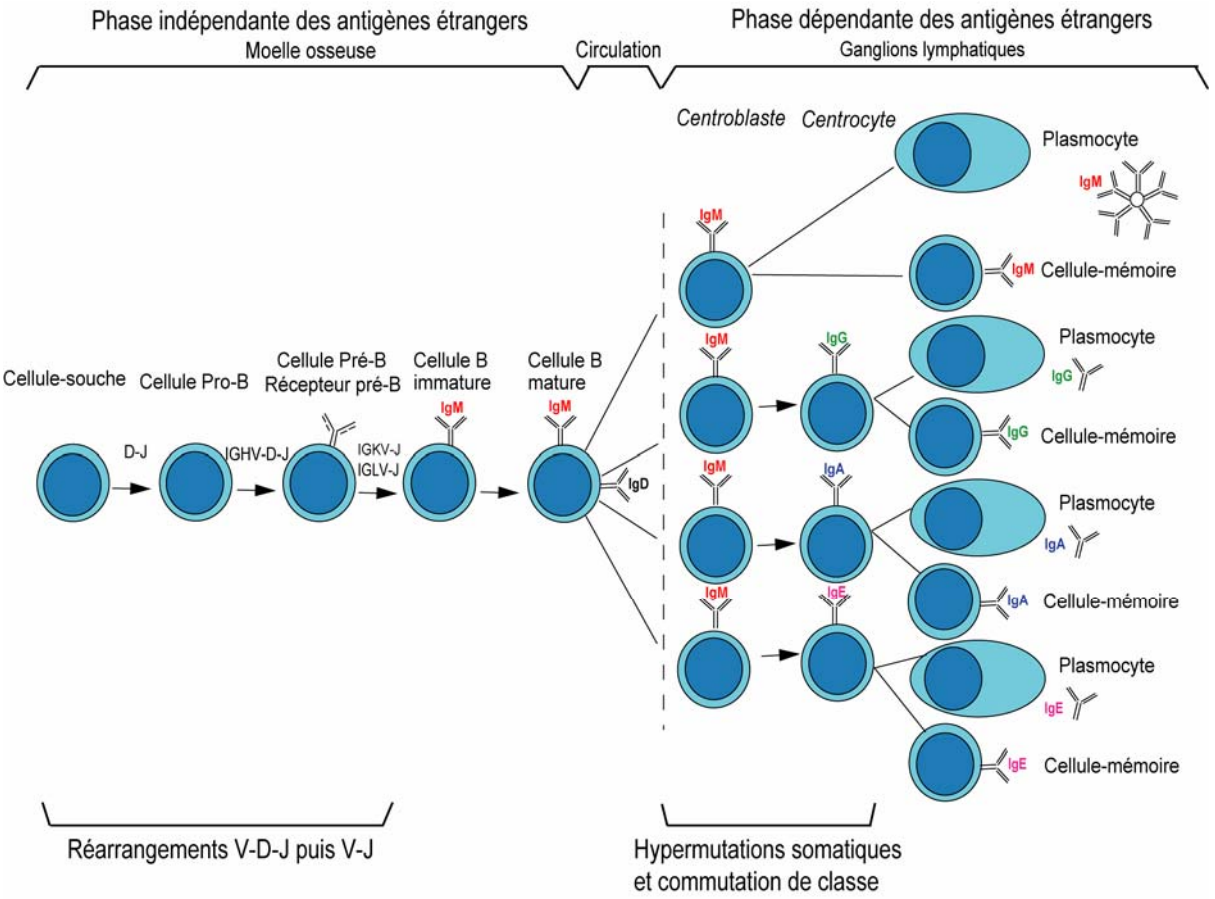


Figure 1.2: Différenciation des lymphocytes B. La différenciation des lymphocytes B comprend deux phases : une phase indépendante des antigènes étrangers, de la cellule souche hématopoiétique jusqu'au lymphocyte B mature, dans la moelle osseuse, et une phase dépendante des antigènes étrangers, du lymphocyte B mature au plasmocyte et au lymphocyte B mémoire, dans les centres germinatifs des organes lymphoïdes secondaires (rate, ganglions lymphatiques). Cette seconde phase requiert généralement une coopération entre les lymphocytes B et T [3] (avec la permission de M.-P. et G. Lefranc, IMGT®, http://www.imgt.org).

Les immunoglobulines se composent de deux chaînes lourdes identiques, associées à deux chaînes légères identiques, kappa ou lambda (Figure 1.3). Chez les humains, les gènes qui codent les chaînes lourdes, les chaînes légères kappa et les chaînes légères lambda, sont localisés dans les locus IGH, IGK et IGL respectivement sur les chromosomes 14 (14q32.33), 2 (2p11.2) et 22 (22q11.2). La synthèse des chaînes lourdes et des chaînes légères des immunoglobulines requiert le réarrangement de trois types de gènes, variables (V), de diversité (D) et de jonction (J), au niveau de l'ADN, dans les locus IG durant la différenciation des cellules B [2, 19, 20]. Chronologiquement, la synthèse des chaînes lourdes mu précède celle des chaînes légères.

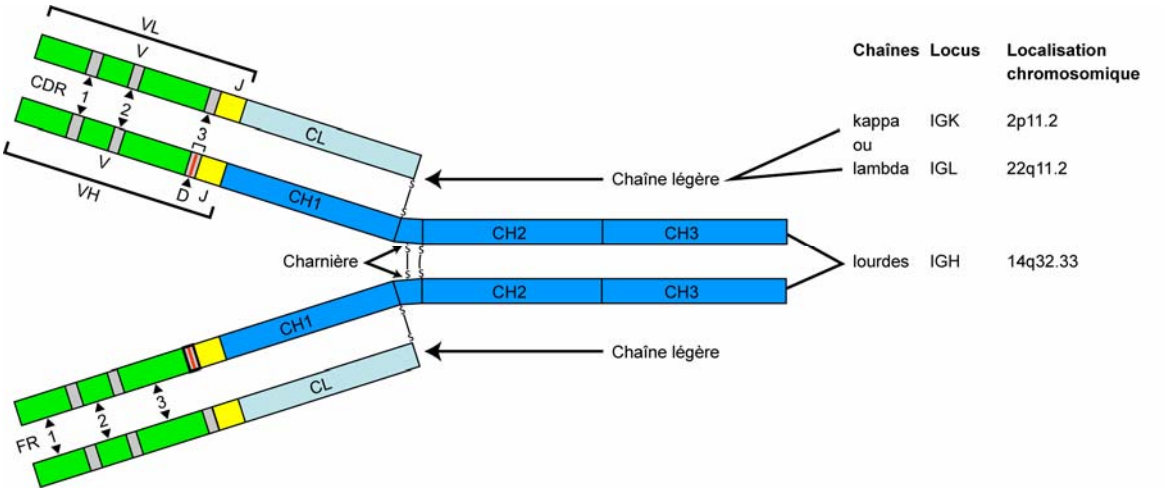


Figure 1.3: Représentation schématique d'une molécule d'IgG1 humaine sécrétée. Le domaine variable d'une chaîne lourde est codé par trois gènes réarrangés (un gène IGHV, un gène IGHD et un gène IGHJ). Le domaine variable d'une chaîne légère, ou V-J-REGION, est codé par deux gènes réarrangés (un gène IGKV réarrangé à un gène IGKJ pour une chaîne kappa, un gène IGLV réarrangé à un gène IGLJ pour une chaîne lambda). Les trois régions hypervariables ou complementarity determining regions (CDR) déterminent le site de reconnaissance et de liaison à l'antigène dans la structure tridimensionnelle. La région constante de la chaîne lourde, codée par des gènes IGHC, comprend 3 ou 4 domaines constants (domaines CH1, CH2, CH3 pour la région constante des IgG, des IgA et des IgD, 4 domaines CH1 à CH4 pour les IgE et les IgM). La région charnière située entre les domaines CH1 et CH2 des IgG est codée par 1 exon (cas des IgG1, IgG2 et IgG4) ou plusieurs exons, le plus souvent 4 (cas des IgG3). La région constante, ou C-REGION, de la chaîne légère est codée par le gène IGKC (cas des chaînes kappa) ou l'un des gènes IGLC (cas des chaînes lambda), et comprend un seul domaine constant (CL) [3] (avec la permission de M.-P. et G. Lefranc, IMGT® http://www.imgt.org).

1.2.1 Synthèse des chaînes lourdes mu

1.2.1.1 Réarrangement D-J et V-D-J dans le locus IGH

Le locus des chaînes lourdes (IGH) comprend des gènes variables (V), de diversité (D), de jonction (J) et constants (C). Le domaine variable de la chaîne lourde, ou V-D-J-REGION, est généré par le réarrangement au niveau de l'ADN de trois gènes: un gène IGHV, un gène IGHD et un gène IGHJ. Il se fait en deux temps, le premier correspond au réarrangement d'un gène D à un gène J avec délétion de l'ADN intermédiaire (excision d'une boucle d'ADN) (Figure 1.4), et le second correspond au réarrangement d'un gène V au D-J précédemment réarrangé pour générer la séquence réarrangée IGHV-D-J (Figure 1.5 et Figure 1.6). La séquence réarrangée IGHV-D-J est transcrite avec le gène IGHM en un pré-messager IGHV-D-J-M (ou IGHV-D-J-Cmu). Le gène IGHM code les quatre domaines CH1 à CH4 de la région constante de la chaîne lourde mu. Les séquences d'ARN correspondant aux introns et aux gènes J non utilisés sont alors excisées lors de l'épissage du prémessager et l'on obtient un ARN messager mature qui comprend les régions codantes épissées et les régions 5' et 3' non traduites. L'ARN messager est ensuite traduit en une chaîne polypeptidique par les ribosomes. Le peptide signal L est éliminé par une peptidase après pénétration de la chaîne

polypeptidique dans la cavité du réticulum endoplasmique, et une chaîne lourde mu est alors produite.

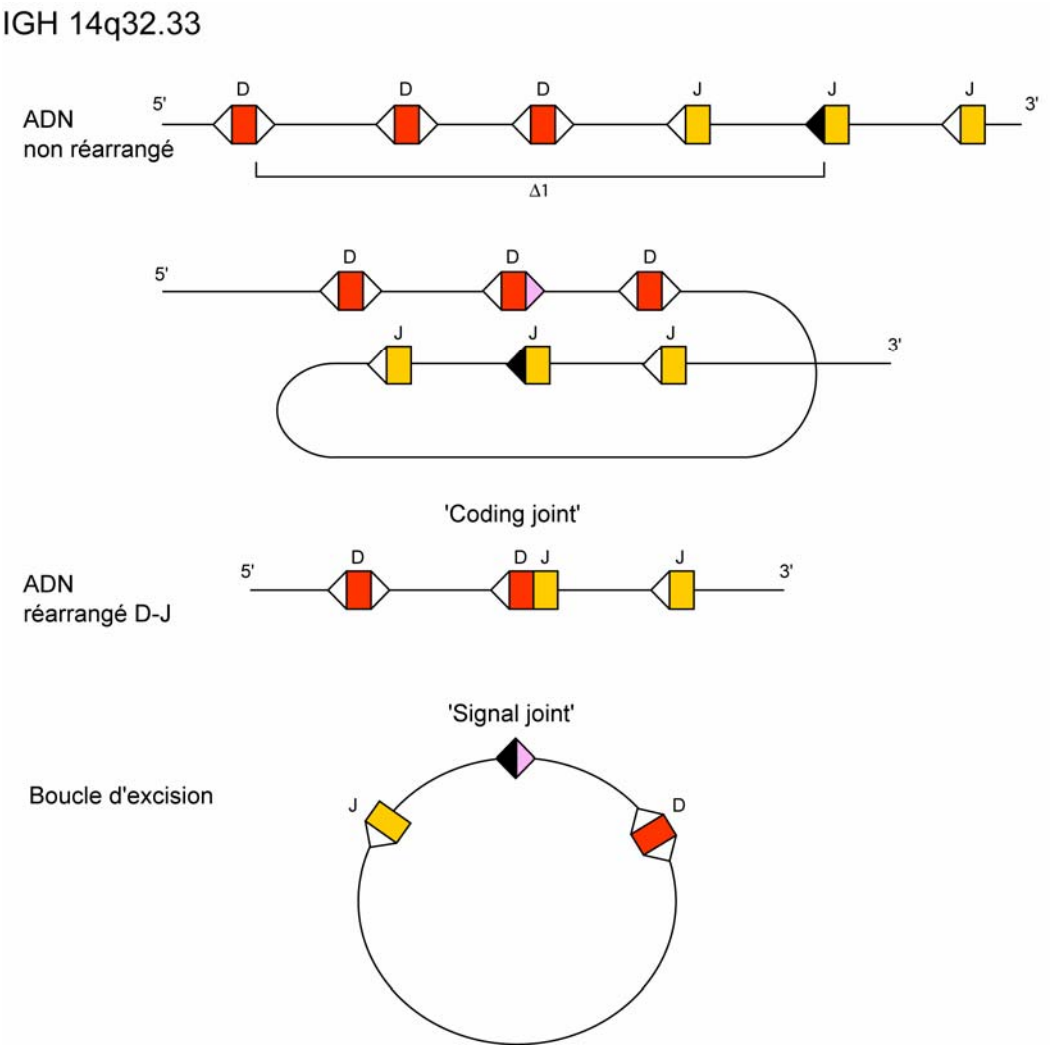


Figure 1.4: Réarrangement dans le locus IGH d'un gène D à un gène J avec délétion de l'ADN intermédiaire (excision d'une boucle d'ADN) (avec la permission de IMGT®, http://www.imgt.org).

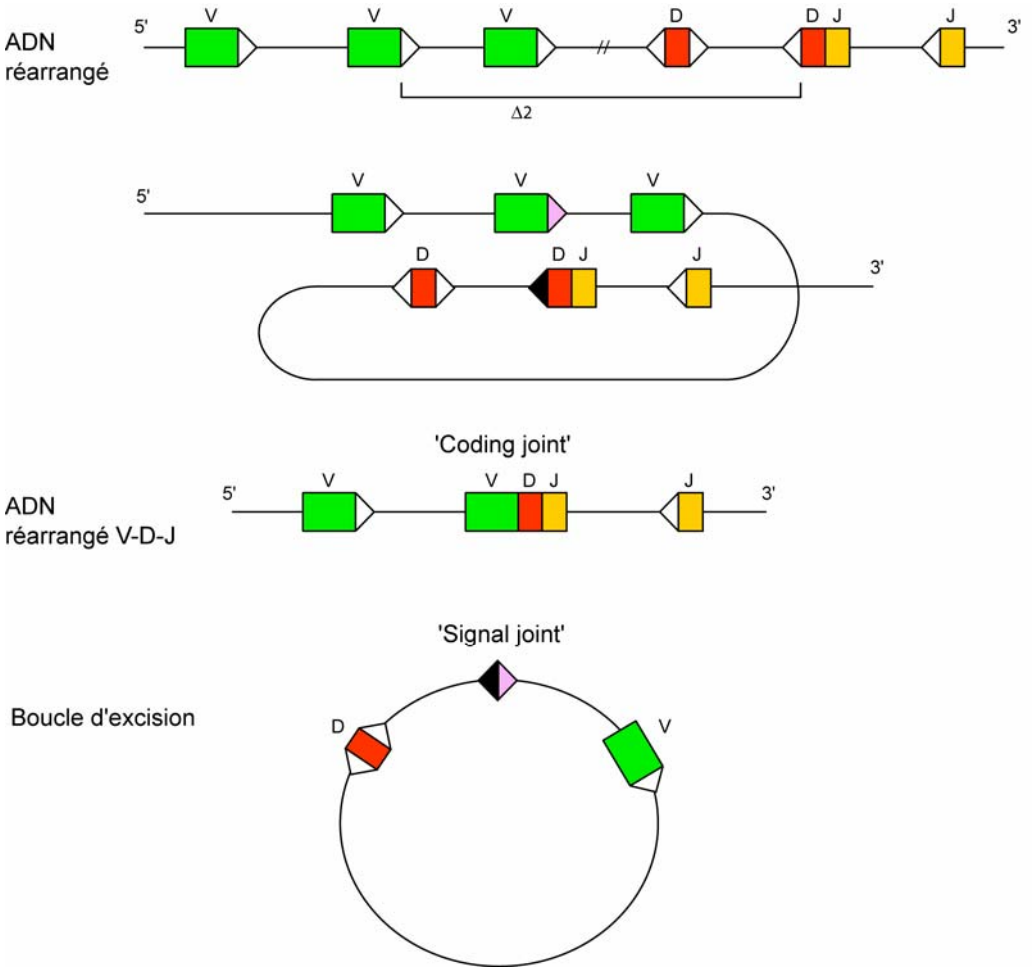


Figure 1.5: Réarrangement dans le locus IGH d'un gène V au D-J précédemment réarrangé pour générer la séquence réarrangée IGHV-D-J (avec la permission de IMGT®, http://www.imgt.org).

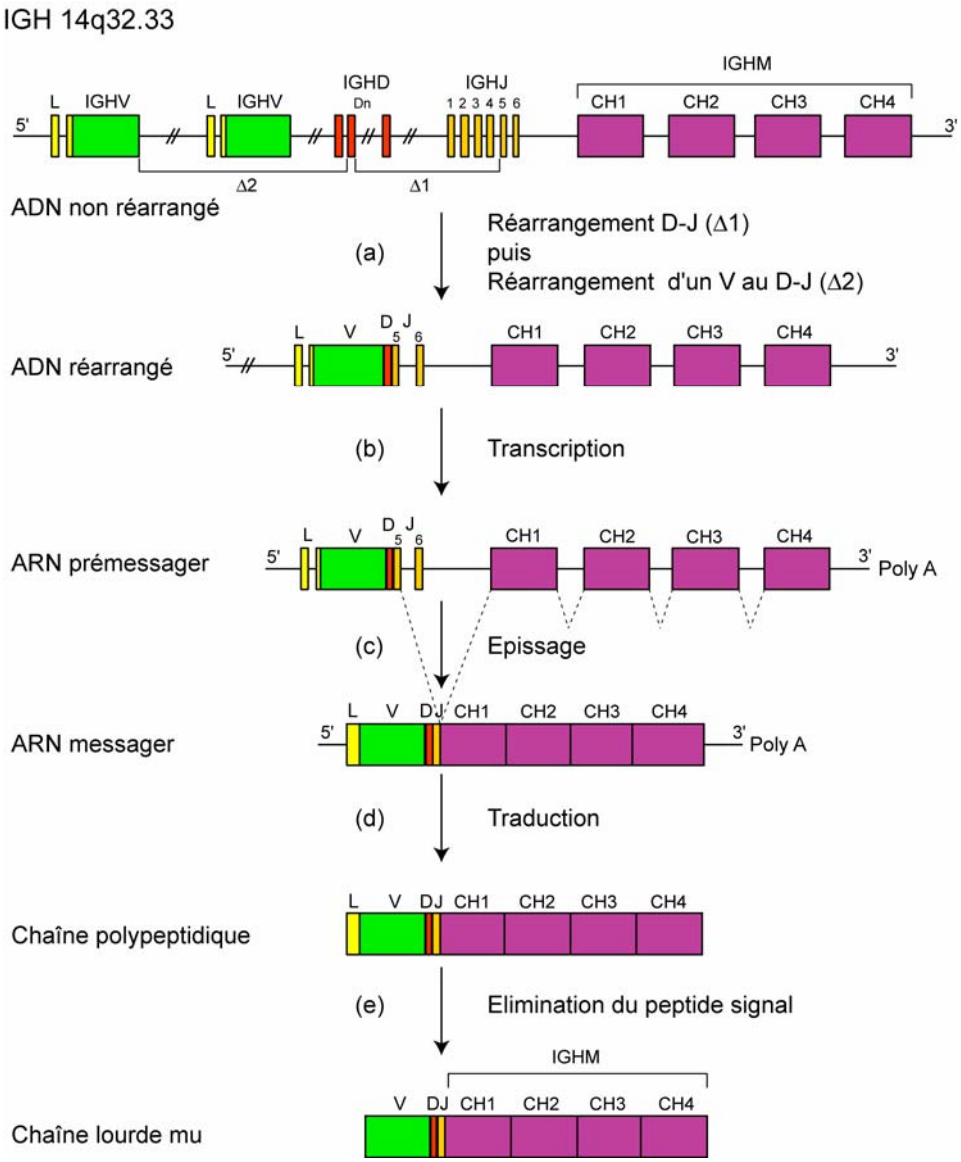


Figure 1.6: Synthèse d'une chaîne lourde mu d'immunoglobuline. (a) Au niveau de l'ADN, lors d'une première étape, l'un des gènes IGHD est joint à l'un des gènes IGHJ avec délétion de l'ADN intermédiaire, pour créer une séquence D-J partiellement réarrangée. Dans une deuxième étape, un des gènes IGHV est joint au D-J préalablement réarrangé, avec délétion de l'ADN intermédiaire pour générer un ensemble IGHV-D-J complètement réarrangé. (b) La séquence réarrangée IGHV-D-J est transcrite avec le gène IGHM en un ARN prémessager IGHV-D-J-M (ou IGHV-D-J-Cmu). (c) Les séquences d'ARN correspondant aux introns et aux gènes IGHJ non utilisés sont alors excisées lors de l'épissage de l'ARN prémessager et l'on obtient un ARN messager mature qui comprend les régions codantes épissées et les régions 5' et 3' non traduites. (d) L'ARN messager est ensuite traduit en une chaîne polypeptidique par les ribosomes. (e) Le peptide signal est éliminé par une peptidase après pénétration de la chaîne polypeptidique dans la cavité du réticulum endoplasmique. Une chaîne lourde mu est produite. Dans l'ADN et L'ARN prémessager, L (pour leader) correspond à L-PART1 et L-PART2, dans l'ARN messager et la chaîne polypeptidique L correspond à la L-REGION [3] (avec la permission de M.-P. et G. Lefranc, IMGT®, http://www.imgt.org).

1.2.2 Synthèse des chaînes légères lambda et kappa

1.2.2.1 Réarrangement V-J dans les locus IGK et IGL

Le locus kappa (IGK) et le locus lambda (IGL) comprennent des gènes variables (V), des gènes de jonction (J) et des gènes constants (C). Le domaine variable de la chaîne légère (kappa ou lambda) ou V-J-REGION est généré par le réarrangement au niveau de l'ADN de deux gènes: un gène V et un gène J, avec délétion de l'ADN intermédiaire pour créer une séquence réarrangée IGKV-J dans le locus IGK (Figure 1.7) ou IGLV-J dans le locus IGL.

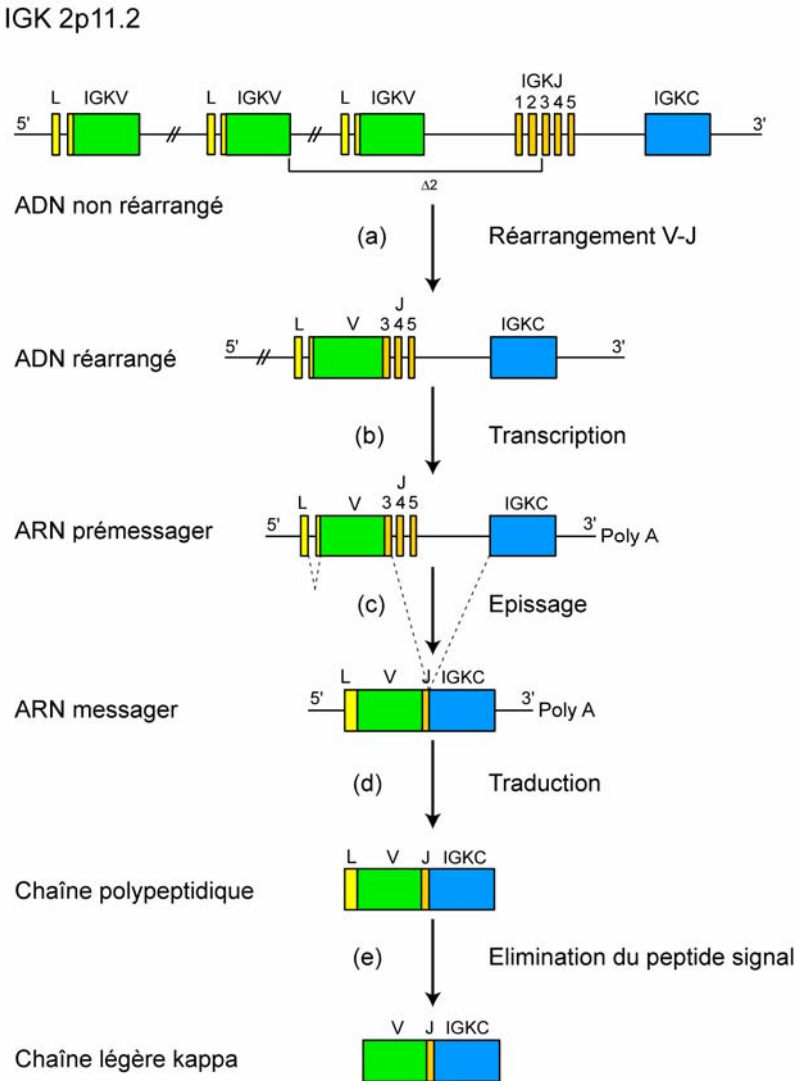


Figure 1.7: Synthèse d'une chaîne légère kappa d'immunoglobuline. (a) Au niveau de l'ADN, l'un des gènes IGKV est réarrangé à l'un des 5 gènes IGKJ avec délétion de l'ADN intermédiaire, pour créer un ensemble IGKV-J. (b) La séquence réarrangée IGKV-J est transcrite avec le gène IGKC en un ARN prémessager IGKV-J-C. (c) Les séquences d'ARN correspondant aux introns et aux gènes IGKJ non utilisés sont alors excisées lors de l'épissage de l'ARN prémessager, et l'on obtient un ARN messager mature qui comprend les régions codantes épissées et les régions 5' et 3' non traduites. (d) L'ARN messager est ensuite traduit en une chaîne polypeptidique par les ribosomes. (e) Le peptide signal L est éliminé par une peptidase après pénétration de la chaîne polypeptidique dans la cavité du réticulum endoplasmique et une chaîne légère kappa mature est produite. Dans l'ADN et l'ARN prémessager, L (pour leader) correspond à L-PART1 et L-PART2, dans l'ARN messager et la chaîne polypeptidique, L correspond à la L-REGION [3] (avec la permission de M.-P. et G. Lefranc, IMGT®, http://www.imgt.org).

La séquence réarrangée IGKV-J (ou IGLV-J) est transcrite avec le gène IGKC (ou un des gènes IGLC) en un ARN pré-messager IGKV-J-C (ou IGLV-J-C). L'unique gène IGKC, ou l'un des gènes fonctionnels IGLC, avec leur unique exon, code respectivement le seul domaine de la région constante des chaînes kappa ou lambda. Les séquences d'ARN correspondant aux introns (et pour le locus IGK aux gènes IGKJ non utilisés, pour le locus IGL aux gènes IGLJ non utilisés) sont alors excisées par épissage de l'ARN prémessager, et l'on obtient un ARN messager mature qui comprend les régions codantes épissées et les régions 5' et 3' non traduites. L'ARN messager est ensuite traduit en une chaîne polypeptidique par les ribosomes. Le peptide signal L est éliminé par une peptidase après pénétration de la chaîne polypeptidique dans la cavité du réticulum endoplasmique et une chaîne légère mature (kappa ou lambda) est alors produite.

1.2.3 Origine de la diversité des domaines variables des immunoglobulines

La diversité du domaine variable des chaînes d'immunoglobuline résulte principalement de la diversité combinatoire, de la diversité des jonctions (diversité fonctionnelle V-J des chaînes légère, de la N-diversité des jonctions V-D-J des chaînes lourdes), et des hypermutations somatiques [3]. De plus, au sein d'une IG, l'association des domaines variables d'une chaîne lourde et d'une chaîne légère, pour former le site de reconnaissance de l'anticorps crée un degré supplémentaire de diversité.

1.2.3.1 Diversité combinatoire

La diversité combinatoire est créée par les réarrangements V-D-J des chaînes lourdes et V-J des chaînes légères. Les réarrangements somatiques IGKV-J, IGLV-J et IGHV-D-J exigent la présence de motifs spécifiques dans la séquence d'ADN, appelés signaux de recombinaison (RS). Ils sont localisés à l'extrémité 3' des gènes V, à l'extrémité 5' des gènes J et de part et d'autre des gènes D (Figure 1.8) [21].

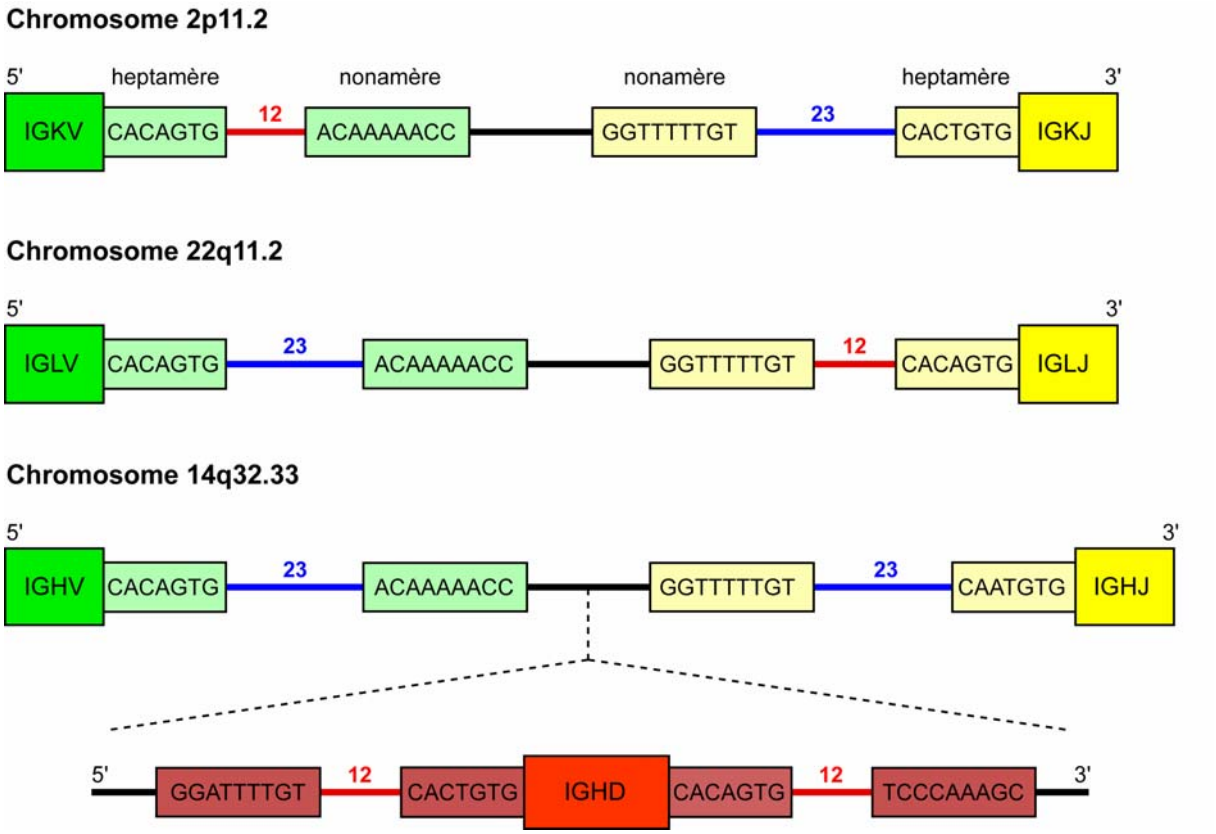


Figure 1.8: Exemples de signaux de recombinaison de gènes V, D et J humains. Bien que les séquences heptamères et nonamères soient bien conservées, il y a des différences entre elles. Les longueurs des espaceurs sont de 12±1 ou 23±1 pb (observation décrite sous le nom de règle 12/23 [22]) [3] (avec la permission de M.-P. et G. Lefranc, IMGT®, http://www.imgt.org).

Ces signaux de recombinaison reconnus par l'enzyme recombinase (protéine RAG1 et RAG2 [23, 24]), sont constitués de deux motifs hautement conservés, un heptamère palindromique et un nonamère riche en A et T, séparés par un espaceur de 12 ± 1 ou 23 ± 1 nucléotides non conservés. Chaque gène V dans l'ADN est suivi par un heptamère de séquence consensus 'CACAGTG' et un nonamère de séquence consensus 'ACAAAAACC', séparés par un intervalle de 12 nucléotides pour le locus IGK, 23 nucléotides pour le locus IGL et pour le locus IGH. De même, les gènes J sont précédés d'un nonamère de consensus 'GGTTTTTGT' et d'un heptamère de consensus 'CACTGTG', séparés par 23 nucléotides pour le locus IGK, de 12 nucléotides pour le locus IGL et de 23 nucléotides pour le locus IGH. Des motifs heptamère-nonamère sont présents de part et d'autre des gènes D séparés par un espaceur de 12 nucléotides non conservé (Figure 1.8 et Figure 1.9).

La distance qui sépare les nonamères des heptamères (12 ± 1 ou 23 ± 1 pb) correspond à 1 ou 2 tours d'hélice de l'ADN. De plus, les signaux heptamère-nonamère en 3' des gènes V sont des séquences complémentaires des signaux heptamère-nonamère en 5' des gènes J, et les séquences des heptamères présentent une structure palindromique. Les réarrangements efficaces ont lieu lorsqu'ils font intervenir deux signaux de jonction dont l'un a un intervalle

de 12±1 nucléotides et l'autre de 23±1 nucléotides entre l'heptamère et le nonamère, règle dite 12/23 [22]. Ceci explique que les gènes d'un même type au sein d'un même locus, ne réarrangent pas normalement entre eux. Ainsi, dans le locus IGH, un gène D réarrange avec un gène J, puis un gène V réarrange avec le D-J précédemment réarrangé (Figure 1.6). Le répertoire potentiel résulte de la diversité combinatoire qui dépend du nombre de gènes V, D, J et de leur fonctionnalité.

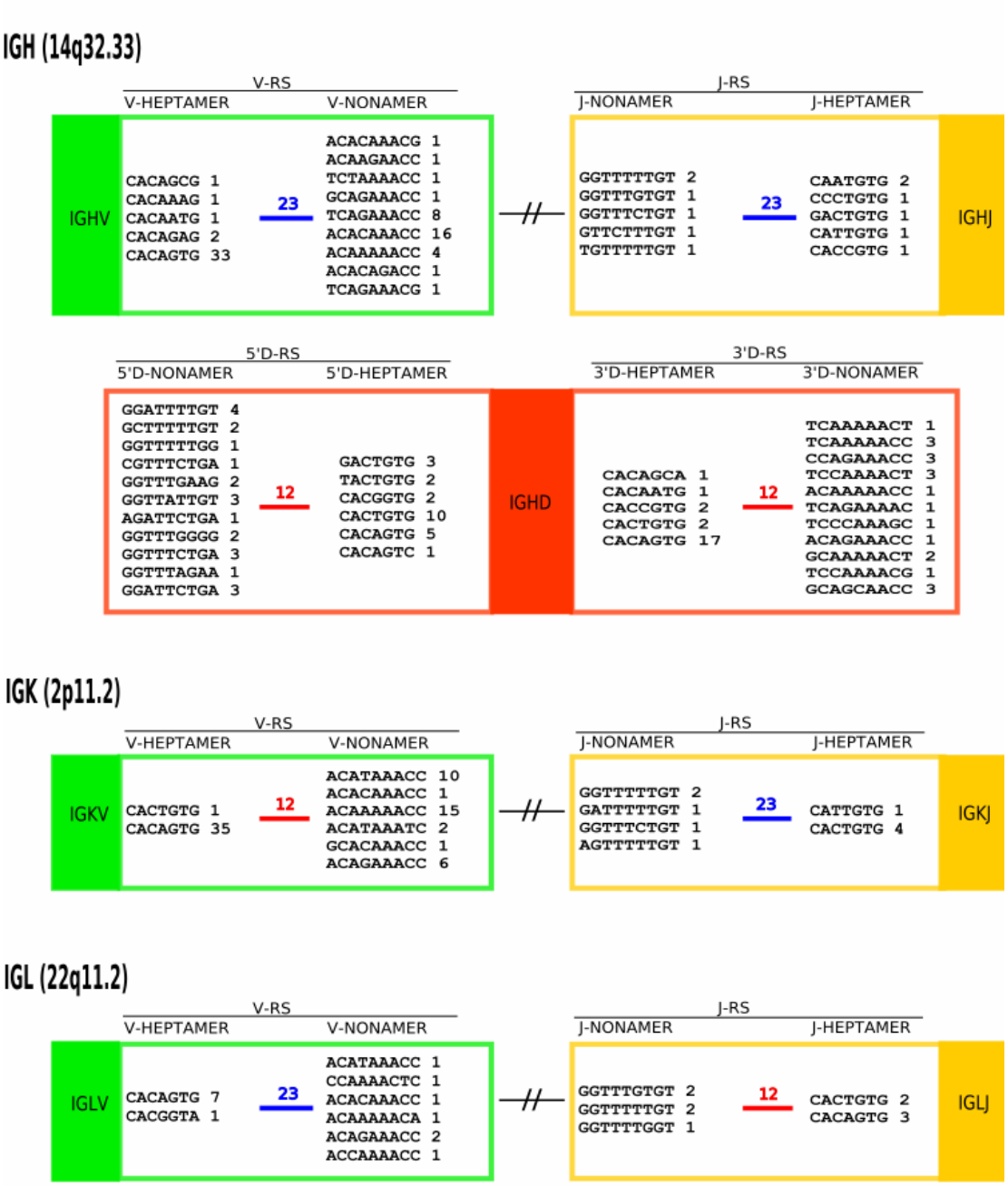


Figure 1.9: Séquence des signaux de recombinaison (RS) pour les locus IGH, IGK et IGL humains. Seules les séquences des RS des allèles fonctionnels *01 sont montrées. Les nombres à droite de chaque heptamère et nonamère indiquent le nombre de gènes où la séquence RS a été trouvée (avec la permission de IMGT®, http://www.imgt.org).

1.2.3.2 Diversité des jonctions

La diversité des jonctions est représentée par la diversité jonctionnelle V-J des chaînes légères kappa qui crée une variabilité de l'acide animé en position 115 du CDR3-IMGT réarrangé [21, 25], et par la N-diversité (N, pour nucléotides) essentiellement observée au jonctions V-D-J des chaînes lourdes d'immunoglobuline et qui représente la source principale de la diversité des CDR3 [26]. En 1982, Alt et Baltimore ont proposé un modèle qui explique le mécanisme de la N-diversité [26]. Elle résulte de l'excision de nucléotides par une exonucléase aux extrémités des gènes V, D et J lors du réarrangement, suivie de l'addition de nucléotides au hasard par la terminaldeoxynucleotidyltransferase TdT [27]. Cette addition de nucléotides implique préférentiellement des nucléotides indépendamment de toute matrice. Si les extrémités des régions codantes restent intactes (pas d'activité exonucléase), l'on peut observer adjacents à ces régions codante des P-nucléotides [28] qui résultent de l'ouverture dissymétrique de l'épingle à cheveux (ou hairpin) formée aux extrémités des régions codantes au cours des réarrangements V-J ou V-D-J [29]. Les P-nucléotides sont ainsi de courtes séquences de 1 à 3 nucléotides palindromiques de l'extrémité codante intacte de la V-REGION, de la D-REGION ou de la J-REGION réarrangée.

1.2.3.3 Mécanisme des réarrangements

Le réarrangement V-(D)-J est assuré par l'intervention successive de nombreuses protéines. Les protéines RAG1 et RAG2 (Recombination Activating Gene) spécifiques de la lignée lymphoïde, forment un complexe majeur dans le processus de réarrangement. D'autres enzymes interviennent pendant les réarrangements V-D-J et V-J. Ce sont les enzymes recrutées pour la réparation de l'ADN double brin qui appartiennent à la voie de réparation des extrémités non homologues (Non-Homologous End Joining ou NHEJ). Les réarrangements V-(D)-J se produisent en 5 étapes.

La formation du complexe RAG1-RAG2 est responsable de la reconnaissance et de la liaison aux séquences RS sur les deux gènes qui réarrangent et dont la chromatine est ouverte comme l'a démontré la présence de transcrits germline. Il y a alors création d'une boucle d'ADN maintenue par le complexe (Figure 1.10A). Le complexe RAG1-RAG2 initialise la coupure de l'ADN en deux étapes (Figure 1.10A étape a et b). A la suite de la coupure de l'ADN double brin, il se forme d'une part une boucle d'excision par jonction des RS (signal joint) qui est élimineé. Cette boucle d'ADN comprend l'ADN intermédiaire entre les 2 gènes qui réarrangent. Sur les brins codants, la coupure crée pour chacun des gènes, un groupe 3′-OH

(hydroxyle) entre la fin de la partie codante du gène et l'heptamère du RS. Cette étape est suivie au niveau des brins codants d'une réaction de transestérification (activité phosphodiester par attaque de l'hydroxyle sur la liaison opposée du brin antiparallèle), qui forme une épingle à cheveux ('hairpin') aux extrémités des deux gènes qui réarrangent (extrémités codantes) (Figure 1.10A étape b).

L'étape suivante (Figure 1.10B étape c) fait intervenir le complexe Ku70/Ku80 qui reconnaît et se fixe aux extrémités codantes hairpin, et recrute la protéine kinase dépendante de l'ADN (DNA-PK). Ce complexe recrute ensuite le facteur Artémis qui, après avoir été phosphorylé par la sous-unité catalytique de la DNA-PK (DNA-PKcs), ouvre les structures hairpin par son activité endonucléase [30, 31]. L'ouverture de la structure hairpin libère ainsi les extrémités codantes et peut, selon la localisation du site de coupure, engendrer la formation de nucléotides P qui constituent un court palindrome, si ces nucléotides ne sont pas excisés par l'exonucléase.
Deux formes distinctes résultant d'un épissage alternatif ont été décrites pour la TdT: une forme longue TdT-L et une forme courte. Il a été suggéré que la TdT-L serait responsable de l'activité exonuclase [32, 33] (Figure 1.10B étape d). Cependant cette observation est sujette à caution [34]. L'addition de novo de nucléotides N de manière aléatoire (souvent g ou c) sans brin matrice (Figure 1.10B étape e) se fait grâce à la forme courte de la TdT. Les régions de N-diversité sont différentes d'une cellule B à l'autre, ce qui crée une diversité considérable au niveau des jonctions responsables de la spécificité des sites anticorps. Ces régions N représentent de véritables signatures des lymphocytes B et des clones qui en résultent après activation.
Enfin, le complexe XRCC4/ Cernunnos/DNA-Ligase IV prend en charge la ligation et la réparation de l'ADN réarrangé (Figure 1.10B étape f).

La transcription de la séquence réarrangée aboutit à la production d'un ARN prémessager comprenant la séquence réarrangée IGHV-D-J avec le gène IGHM ou Cmu. Au départ, toutes les cellules B produisent une chaîne mu dans la moelle osseuse et expriment IgM (le gène IGHM est celui qui est situé le plus près de l'ensemble réarrangé V-D-J dans le locus IGH). L'association du même domaine variable avec les autres régions constantes pour former des IgG et IgE a lieu après l'activation de la cellule B par l'antigène dans les organes lymphoïdes secondaires (rate et ganglions lymphatiques).

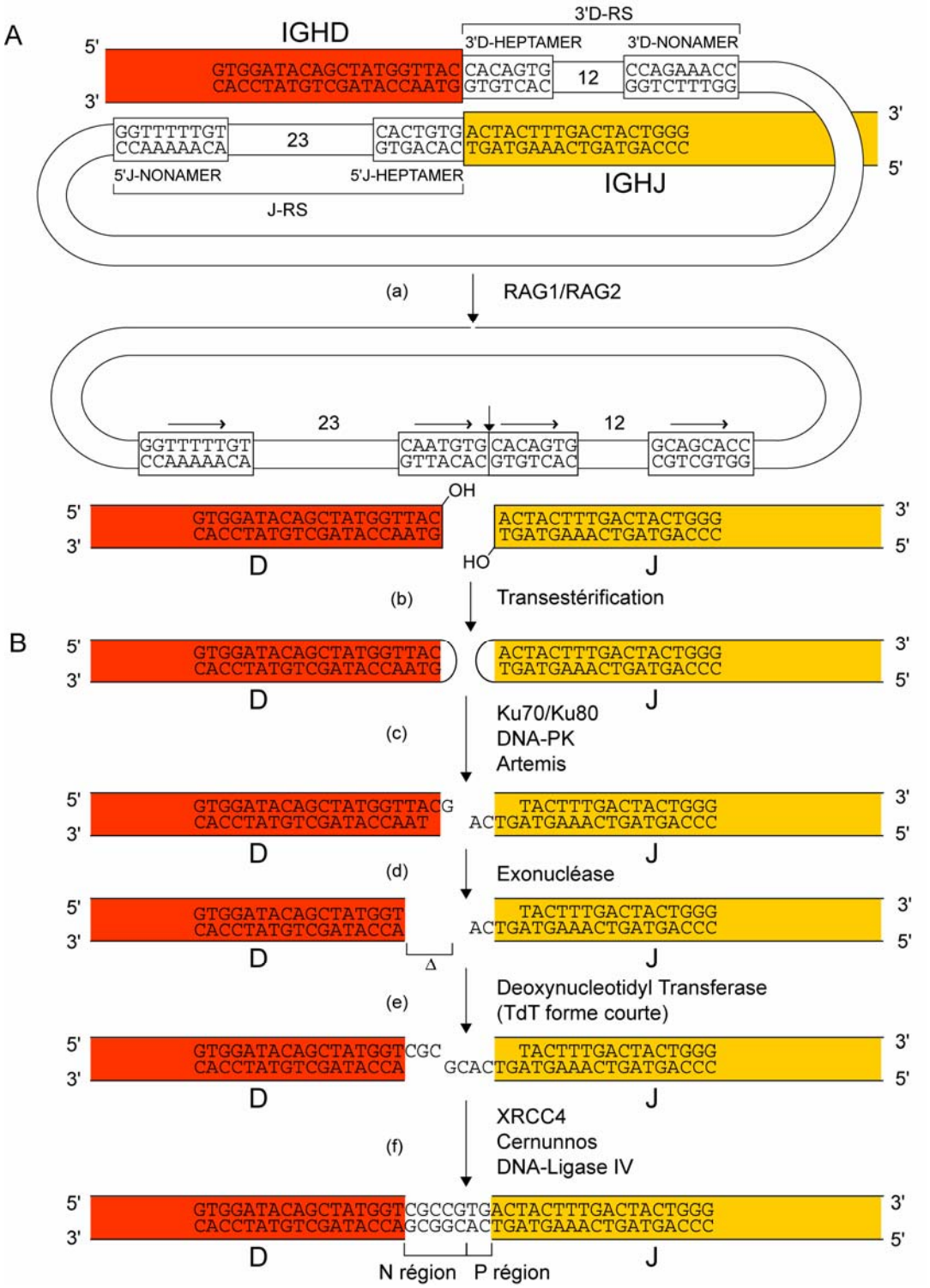


Figure 1.10: Schéma des événements moléculaires à l'origine d'un réarrangement entre un gène D et un gène J dans le locus IG. (a) Le complexe RAG1/RAG2 reconnaît les sites RS et coupe l'ADN germline entre l'heptamère et la région codante. Une épingle à cheveux est formée à l'extrémité codante. (b) L'épingle à cheveux coupée par un complexe protéique (Ku70/80, DNA-PK, Artémis). Selon la position du site de coupure, on obtient une extrémité à bouts francs ou des P nucléotides. (c) Une exonucléase élimine des nucléotides aux extrémités codantes. (d) La TdT (forme courte) ajoute des nucléotides N préférentiellement des g et c. (e) Enfin, vient la ligation et la réparation de l'ADN réarrangé (XRCC4, Cernunnos, DNA-Ligase) (avec permission de IMGT®, http://imgt.org).

1.2.3.4 Hypermutations somatiques

Les hypermutations somatiques (HMS) apparaissent durant la maturation des cellules B dans les centres germinatifs des organes lymphoïdes secondaires (rate et ganglions lymphoïdes). Elles affectent spécifiquement les gènes réarrangés V-J et V-D-J des IG et représentent un mécanisme majeur pour la génération de la diversité domaines variables de des anticorps [35] [36]. Ce processus d'hypermutation somatique implique l'introduction, à un taux élevé, de mutations ponctuelles dans les gènes réarrangés V-D-J des chaînes lourdes et V-J des chaînes légères. Les autres gènes exprimés dans les cellules B ne sont pas modifiés par les mécanismes d'hypermutation somatique. Les HMS se produisent à une fréquence estimée de 10^{-3} par base dans une cellule en division, ce qui est à peu près 10^{6} fois plus fréquent que le taux de mutations spontanées dans les autres cellules.

Les HMS commencent à environ 150 pb en aval du promoteur des gènes V des IG, et s'étendent sur une distance de 1 à 2 kb. La fréquence des mutations atteint son maximum dans les gènes réarrangés V-J ou V-D-J et est inversement proportionnelle à la distance par rapport au promoteur. Ceci suggère que le promoteur joue un rôle important dans la localisation des mutations dans les domaines V. En effet, il a été montré que les HMS sont fortement liées à la transcription, la délétion du promoteur dans des transgènes IGH provoquant une diminution des HMS [37]. De plus, le taux de mutations est directement corrélé à la quantité de transcrits dans le locus [38].

Les hypermutations somatiques ciblent préférentiellement des motifs particuliers: (a/t)a et g(c/t)(a/t) et leurs motifs en inverse complémentaire: t(a/t) et (a/t)(a/g)c. Les transitions (c>t, t>c, a>g, g>a) sont plus fréquentes que les transversions (par exemple c>a, c>g, a>c, g>c). Les deux brins de la séquence peuvent être affectés. Les mutations qui engendrent des changements d'acides aminés dans la séquence protéique des IG sont observées le plus souvent dans les CDR. Ces mutations permettent après sélection, une amélioration de l'affinité du site de liaison à l'antigène. Les mutations silencieuses sont observées plus généralement dans les FR, préservant ainsi la structure du domaine.

L'étape initiale du processus d'hypermutations somatiques est induite par l'activation de la cytidine deaminase (AID): son activation permet la déamination des cytosines (c) en uracile (u) sur un des deux brins de l'ADN, ce qui engendre un mésappariement (u/g).

Comme, l'uracile n'est pas un composant naturel de l'ADN, ce mésappariement doit être réparé. Selon le mécanisme de réparation mis en jeu et le profil des mutations introduites (Figure 1.11) sera différent [39].

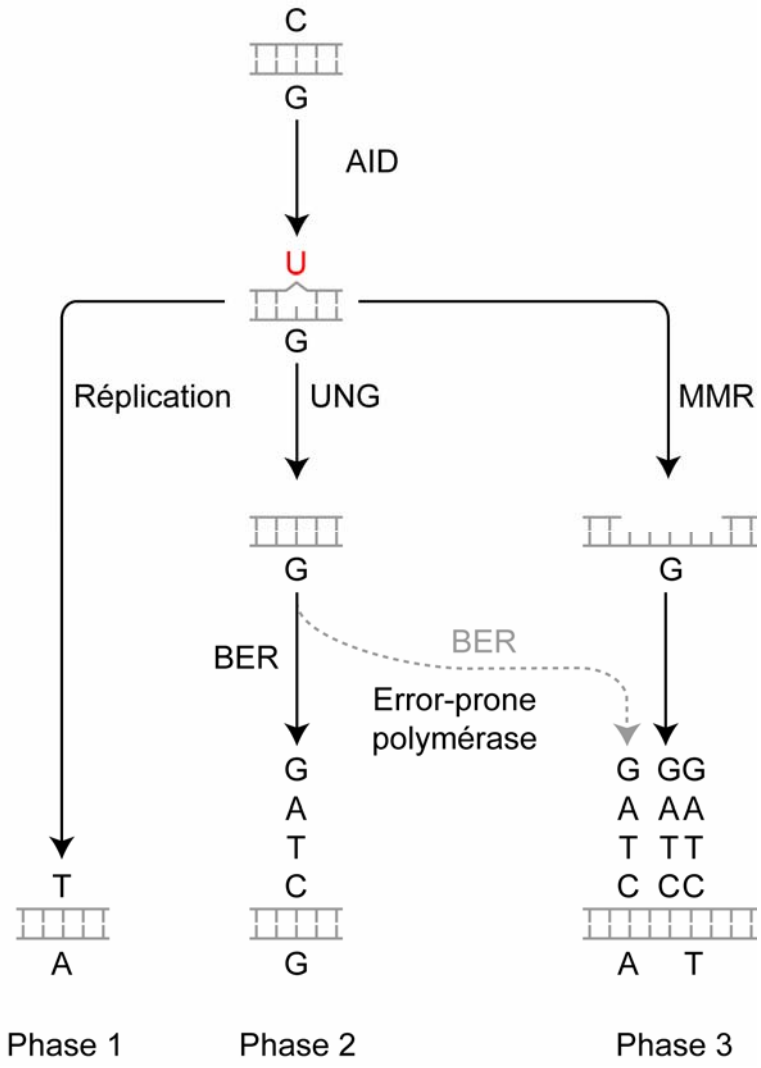


Figure 1.11: Modèle expliquant les hypermutations somatiques d'après [39]. La cytidine deaminase (AID) induit la déamination des cytosines (C) en uracile (U) en rouge. Le mode de traitement de l'uracile par les mécanismes de réplication et/ou de réparation, déterminera le devenir et le profil des mutations introduites selon 4 processus. Avec UNG pour l'uracile glycosilase, BER pour Base Excision Repair, MMR pour mismatch repair.

- Lors de la réplication, l'uracile est considéré comme une thymine par les ADN polymérases, il y a alors création d'une mutation de type transition c > t ou dans le brin opposé g>a, après réplication (phase 1).
- La mutation (l'uracile) peut également être éliminée par l'uracile glycosilase (UNG) créant ainsi un site abasique. Ce dernier peut être la cible d'une endonucléase APE1, endonucléase majeure de la voie 'Base Excision Repair ' (BER), qui clive l'ADN (phase 2). Il existe également une voie annexe, appelée long patch BER, qui est responsable de plusieurs mutations de bases a et t.
- Il peut également impliquer le système de réparation des mésappariements nucléotidiques (mismatch repair ou MMR), par lequel les mésappariements 'u>g' sont

reconnus par le complexe MSH2/MSH6 et suivis par excision de nucléotides entourant la base uracile (u) par l'exonucléase 1. Ce mécanisme est également responsable des mutations des bases (a/t) situées à proximité des paires de bases 'c/g'.

Finalement lors des voies BER et MMR (respectivement les phase 2 et 3), les cassures de l'ADN sont réparées par les polymérases translésionnelles (telles que Polθ et Rev1) qui insérent un des quatre nucléotides, aboutissant à la reconstitution de la cytosine initiale, à une transition, ou à une transversion [40-42].

1.2.3.4.1 Délétions et insertions

Le mécanisme d'hypermutation somatique peut entraîner très rarement des insertions ou des délétions dans les séquences réarrangées d'IG. De telles modifications dans un réarrangement peuvent tout de même donner des séquences productives, si les insertions ou délétions sont des multiples de trois nucléotides, ce qui permet de préserver le cadre de lecture de la séquence, et si aucun codon stop n'est généré. Les mécanismes exacts qui engendrent de tels événements ne sont pas connus.

1.2.4 Co-expression des chaînes lourdes membranaires mu et delta

Au cours de sa différenciation, le lymphocyte B immature devient un lymphocyte B mature qui exprime simultanément des IgM et des IgD membranaires (Figure 1.3). Les domaines variables des chaînes lourdes mu et delta sont identiques et sont codés par les mêmes gènes réarrangés IGHV-D-J. L'expression de l'isotype IgD diffère de l'expression des autres isotypes en ce que son expression dépend d'un mécanisme d'épissage et non du mécanisme de commutation de classe comme les autres isotypes. De plus, l'IgD est co-exprimée avec l'IgM à la surface des cellules B matures naïves (seul cas où deux isotypes différents d'IG sont exprimés par la même cellule).

Les gènes IGHM et IGHD sont situés à proximité dans le locus IGH. Les cellules B qui expriment IgM et IgD produisent deux ARN prémessager différents; dans l'un, la transcription se termine après le gène IGHM; dans l'autre, les deux gènes IGHM et IGHD sont transcrits et la transcription se termine après le gène IGHD. Les transcrits se terminant après le gène IGHM subissent un épissage pour produire un ARNm de chaîne mu; les transcrits se terminant après le gène IGHD subissent un épissage qui élimine les exons du gène IGHM pour produire un ARNm de chaîne delta. Il n'est pas exclu que ce long prémessager soit aussi utilisé pour produire un ARNm de chaîne mu. Par une régulation des

sites de terminaison des prémessagers et de l'épissage, le lymphocyte B mature naïf exprime ainsi l'IgM et l'IgD [43-45].

1.2.5 Expression des chaînes gamma, epsilon et alpha et commutation de classe

Chez l'homme, les IG sont réparties en 5 classes ou isotypes IgM, IgD, IgG, IgA et IgE, les IgG et IgA étant elles-mêmes divisées en sous-classes IgG1, IgG2, IgG3, IgG4, IgA1 et IgA2 [3] qui diffèrent par leurs propriétés physico-chimiques, leurs fonctions biologiques [46] et les propriétés effectrices des Fc des régions constantes [47]. La cellule B mature qui entre dans les ganglions lymphatique exprime des IgM et des IgD. Après stimulation antigénique, les cellules B prolifèrent et peuvent se différencier pour produire d'autres isotypes: c'est la commutation de classes ou switch. Le changement de classe ou commutation de classe se produit dans les ganglions lymphatiques, lors de la maturation du lymphocyte B. Le lymphocyte B activé ayant reconnu un antigène entre en contact avec un lymphocyte T (interaction entre le CD40 de la cellule B et le CD40L exprimé à la surface de la cellule T à la suite de son activation) (Figure 1.12).

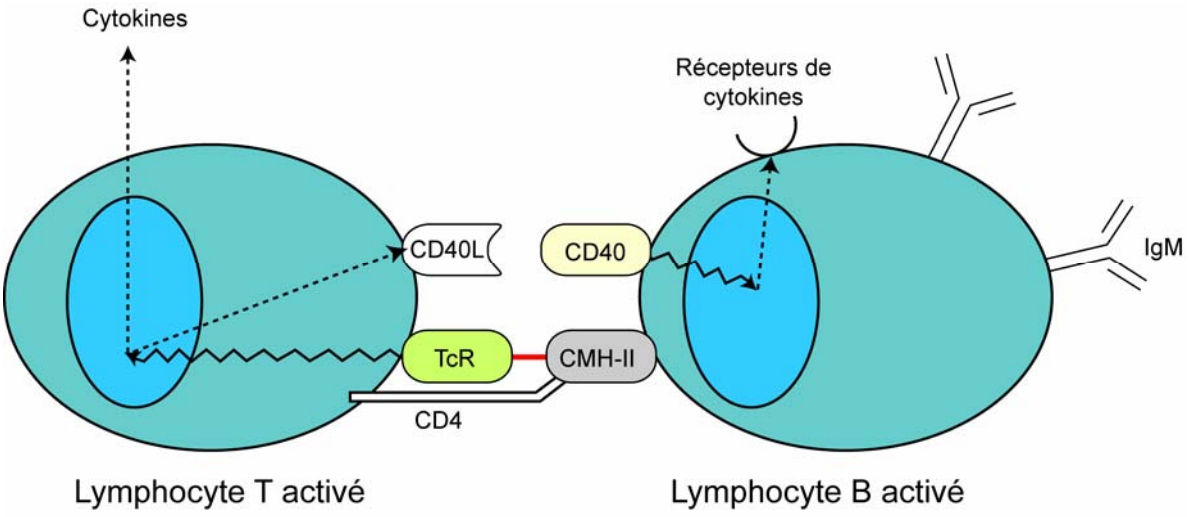


Figure 1.12: Coopération cellulaire B et T. La cellule B reconnaît la cellule T par l'interaction MHC-II – CD4. La cellule T reconnaît la cellule B par la reconnaissance de manière spécifique du complexe TR – peptide/MHC-II), la cellule devient « T activé ». La cellule « T activé » synthétise CD40L et des cytokines. La cellule B reconnaît la cellule « T activé » (CD40- CD40L) et devient une cellule « B activé ». Enfin la cellule « B activé » synthétise des récepteurs de cytokines, et subit une commutation de classe (SWITCH).

Les cellules B passent d'une production d'IgM et d'IgD à la synthèse d'IgG (IgG1, IgG2, IgG3, ou IgG4) ou d'IgA (IgA1 ou IgA2) ou d'IgE (Figure 1.13). Ce processus, qui permet le changement de la région constante de la chaîne lourde tout en maintenant l'expression de la même spécificité anticorps et en renforçant même son affinité, est appelé la «commutation de

classe» (CSR) ou "switch". Elle est rendue possible par l'existence à environ 2 kb en 5' de chacun des gènes IGHC, de séquences particulières. La participation de ces régions au mécanisme de commutation de classe leur a valu l'appellation de séquences S (‘switch’). Ces signaux, d'environ 2 kb, sont composés de 20 à 80 bases répétées en tandem. A l'intérieur de ces motifs, plusieurs courtes séquences répétées ggggt et gagct et, près du site de commutation tggg et tgag sont observées.

La commutation de classe implique la recombinaison de la séquence Smu avec la séquence S d’un autre gène IGHC, par exemple avec une séquence Sgamma dans le cas d’une commutation d’IgM à IgG, ce qui entraîne la délétion des gènes IGHC situés entre Smu et le Sgamma du gène IGHG utilisé (Figure 1.13). Ceci se produit par la formation de boucles de délétion [48-52]. Par exemple, dans le cas d’une commutation d’IgM à IgG1 (Figure 1.13), les gènes IGHM, IGHD et IGHG3 sont délétés.

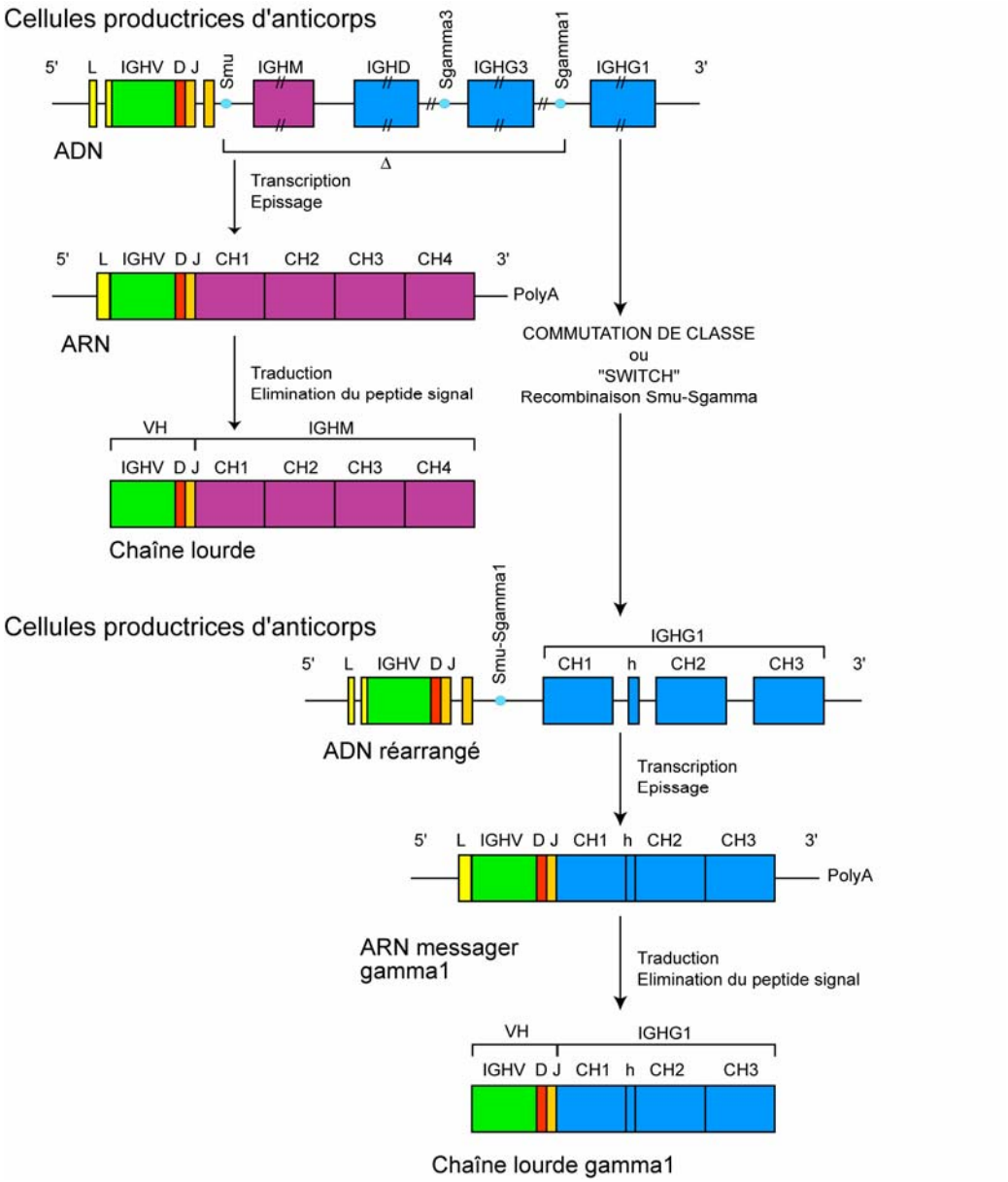


Figure 1.13: Commutation de classe IgM-IgG: recombinaison Smu-Sgamma. Dans la cellule productrice d'anticorps IgM, le réarrangement V-D-J a eu lieu et tous les gènes IGHC sont présents. Lors de la commutation de classe, un deuxième réarrangement se produit, mais cette fois entre deux séquences S et avec délétion des gènes IGHC situés en amont (5') du gène IGHC utilisé. h (pour "hinge") = région charnière [3] (avec la permission de M.-P. et G. Lefranc, IMGT®, http://www.imgt.org).

1.2.6 Expression des chaînes delta de cellules IgM$^-$ IGD$^+$

Seule une minorité de plasmocytes normaux et de rares cellules B malignes expriment exclusivement des IgD (cellule B IgM$^-$ IgD$^+$). Leur faible fréquence a été expliquée par le manque de séquences switch reconnaissables entre Cmu et Cdelta. Cependant, une région, indiquée comme un sigma delta, contient de façon relativement élevée des répétitions pentamériques avec une région extrêmement riche en g et semble fonctionner comme un

switch rudimentaire menant à l'expression de chaînes delta par les cellules B et les plasmocytes dans les centres germinatifs [53-55].

1.2.7 Immunoglobulines membranaires et sécrétées

Les IG membranaires et les IG sécrétées diffèrent par leur région C terminale. D'une manière générale, les chaînes lourdes des IG membranaires à la surface des lymphocytes B ont une région C-terminale hydrophobe qui les maintient ancrées dans la membrane plasmique, tandis que les IG sécrétées par les plasmocytes ont une extrémité C-terminale hydrophile [56]. Les IG membranaires et sécrétées résultent d'un épissage alternatif du transcrit primaire des chaînes lourdes (Figure 1.14).

1.2.7.1 Chaînes mu membranaires et sécrétées

La région C-terminale des chaînes mu membranaires est codée par deux petits exons M1 et M2 localisés à environ 2 kb en 3' de l'exon CH4 [57], M1 code 39 acides aminés, alors que M2 code seulement 2 acides aminés. Ces 41 acides aminés représentent l'ensemble du dispositif d'ancrage des chaînes mu membranaires qui comprend une région extracellulaire de 13 acides aminés entre le domaine CH4 et la membrane, une région transmembranaire hydrophobe de 25 acides aminés et une courte région cytoplasmique de 3 acides aminés. La région C-terminale de la chaîne mu sécrétée comprend 20 acides aminés codés par l'extrémité 3' de l'exon CH4 (désignée par CH-S) [3].

L'expression d'une chaîne mu membranaire suppose l'utilisation d'un site polyA situé en 3' de l'exon M2 et lors de la maturation de l'ARN prémessager, l'utilisation d'un site d'épissage interne situé à l'intérieur de l'exon CH4, à la limite 5' du CH-S. Cet épissage permet d'éliminer le CH-S hydrophile, son codon stop ainsi que toute la région comprise entre CH4 et M1, ce qui permet de joindre le CH4 aux exons M1 et M2. Lors de la synthèse d'une chaîne mu sécrétée, c'est le site polyA situé à 103 bp de l'extrémité 3' de l'exon CH4 et le codon stop à l'extrémité 3' de l'exon CH4 qui sont utilisés (Figure 1.14). Une même cellule peut donc présenter les deux précurseurs ARN mu et l'expression relative d'une chaîne mu membranaire et sécrétée dépend d'un contrôle au niveau de la sélection du site polyA utilisé [58].

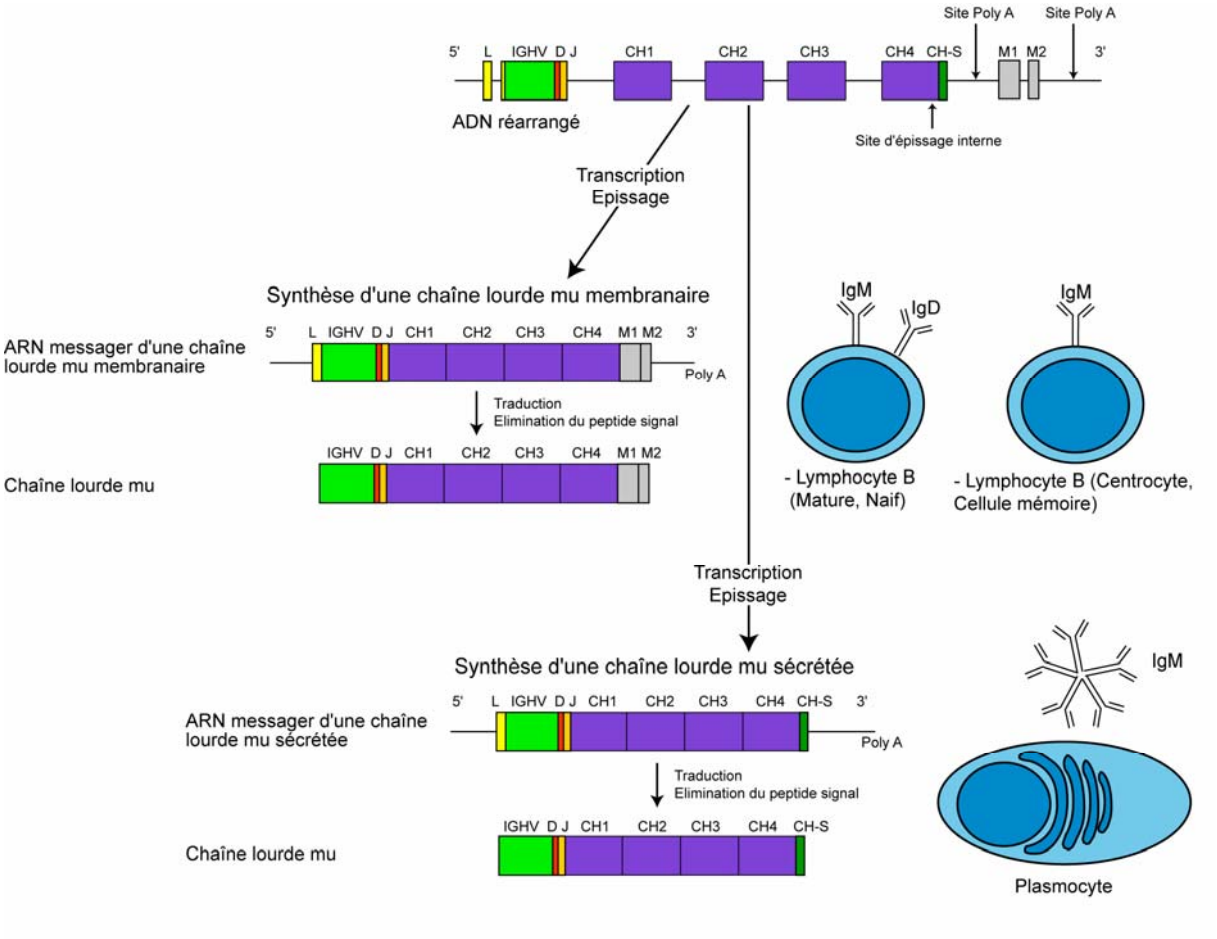


Figure 1.14: Synthèse d'une chaîne lourde mu membranaire ou sécrétée [3] (avec la permission de M.-P. et G. Lefranc, IMGT®, http://www.imgt.org).

1.2.7.2 Chaînes delta membranaires et sécrétées

La région constante de la chaîne delta est codée par le gène IGHD. L'organisation de la région 3' des gènes IGHD diffère par la présence d'un petit exon CH-S indépendant, localisé à 1,9 kb en 3' de l'exon CH3, et qui code 9 acides aminés dans les chaînes delta sécrétées. Les exon M1 et M2 sont localisés respectivement à 0,8 kb et 1,1 kb en 3' du CH-S et codent les régions transmembranaire et cytoplasmique. L'exon M1 code 53 acides aminés alors que l'exon M2 code 2 acides aminés. L'expression des chaînes delta membranaires et sécrétées dépend du site poly A utilisé: pour la synthèse d'une chaîne delta membranaire le site poly A utilisé est localisé en 3' de l'exon M2 et pour la synthèse d'une chaîne delta sécrétée le site poly A utilisé est localisé en 3' du CH-S [59] [3].

1.2.7.3 Chaînes gamma, alpha, epsilon membranaires et sécrétées

L'expression des chaînes gamma, alpha, epsilon membranaires et sécrétées suivent le même mécanisme que celui décrit pour la chaîne mu, le CH-S fait partie de l'exon CH3 ou CH4, selon le gène d'IGHC [3].

1.2.8 Régulation dans le temps des réarrangements V-D-J

Les réarrangements des gènes d'IG se déroulent dans la moelle osseuse au cours du développement des cellules B. Les progéniteurs identifiables au stade le plus précoce du développement des cellules B sont appelés cellules pro-B (Figure 1.15). Le réarrangement D-J a lieu à la suite de l'expression de la protéine recombinase RAG. Il est suivi d'un deuxième réarrangement entre un gène V et D-J. Le réarrangement du locus IGK a lieu ensuite, suivi si nécessaire par le réarrangement du locus IGL [60, 61]. Ces différentes étapes de la synthèse des IG permettent de distinguer les différents stades de la différenciation des cellules B à travers le réarrangement des gènes des IG et l'expression des marqueurs de surface membranaires.

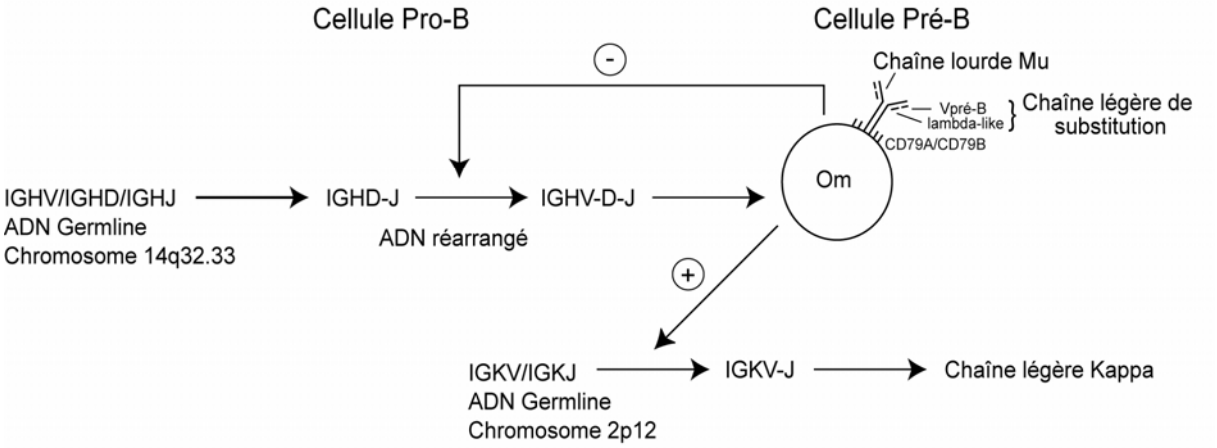


Figure 1.15: Réarrangement des gènes IGH et IGK dans les cellules Pro-B et Pré-B. (avec la permission de M.-P. et G. Lefranc, [3], IMGT®, http://www.imgt.org).

Chaque lymphocyte B somatique, diploïde, a deux copies de chacun des locus des IG (IGH, IGK et IGL). Les deux copies de chaque locus, héritées l'une de la mère et l'autre du père, sont situées sur des chromosomes homologues. Le locus IGH subit un réarrangement avant les locus des chaînes légères. Dans les cellules B en cours de différenciation (cellule pro-B), les réarrangements peuvent s'effectuer sur les deux chromosomes homologues 14. Le réarrangement entre un gène D et un gène J puis le deuxième réarrangement d'un gène V au

D-J réarrangé concerne initialement un seul des locus IGH. Bien que les gènes IGHD aient généralement 2 cadres de lecture, par suite des modifications au niveau de la jonction, il y a approximativement une chance sur trois pour que le cadre de lecture du gène J réarrangé soit conservé. Dans le cas de réarrangements improductifs sur l'un des deux chromosomes, le réarrangement s'effectue alors sur le deuxième chromosome. Si aucun réarrangement V-D-J productif n'est produit, la cellule B meurt par apoptose. Ainsi, il y a une perte considérable des précurseurs des cellules B dans la moelle, bien que les mécanismes dits de 'sauvetage' par des réarrangements secondaires existent (voir ci-dessous).

Les réarrangements IGHV-D-J conduisent à la synthèse d'une chaîne mu intracellulaire. A ce stade, la protéine mu est associée de façon transitoire à une pseudo-chaîne légère ou chaîne légère de substitution, composée en fait de deux protéines codées par le gène Vpré-B (VPREB) et le gène lambda-like. Ces protéines appartiennent à la superfamille IgSF et sont constituées respectivement d'un domaine V-like et d'un domaine C-like, la chaîne lambda-like étant reliée à la chaîne mu par un pont disulfure.
L'ensemble de la chaîne mu associée à la pseudo-chaîne légère et du co-récepteur hétérodimère CD79A/CD79B, définit le récepteur pré-B [62, 63]. Ce récepteur pré-B joue un rôle crucial car il représente un premier contrôle de la maturation du lymphocyte B. Sa stimulation par un ligand, la galectine 1 du stroma de la moelle osseuse environnante, induit l'activation et la prolifération des cellules pré-B [63]. Le récepteur pré-B inhibe le réarrangement du locus IGH du deuxième chromosome 14 et active réarrangement des gènes du locus IGK sur le chromosome 2. Ce mécanisme permet d'assurer qu'une cellule B ne produit qu'une seule chaîne lourde (exclusion allélique). Cependant, il existe de très rares cellules qui expriment deux chaînes lourdes distinctes, ceci étant dû à des échecs du mécanisme d'exclusion allélique.
Il existe une chronologie des réarrangements des gènes des locus de chaînes légères. En effet si les premiers réarrangements du locus IGK sont improductifs, le locus IGK sur le second chromosome 2 réarrange à son tour. Si ces réarrangements IGK sont improductifs, les locus IGL réarrangent à leur tour, sur l'un des deux chromosomes 22. Après la formation d'un réarrangement IGK ou IGL productif, la chaîne légère synthétisée est associée avec la chaîne lourde mu pour constituer une IgM. L'IgM et le co-récepteur formé par les protéines CD79A et CD79B constituent le récepteur des cellules B (BcR) (Figure 1.16).

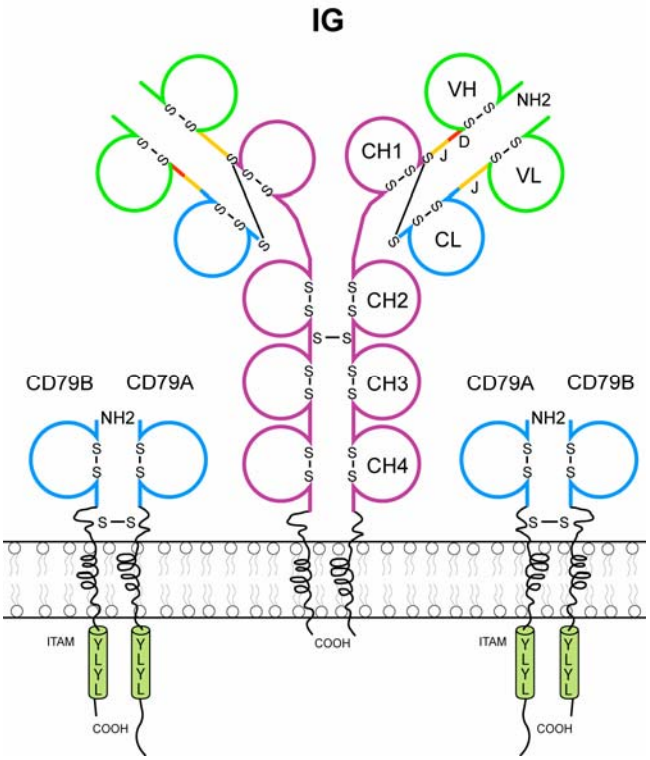


Figure 1.16: Le récepteur des lymphocyte B (B cell receptor ou BcR). Le BcR est composé de l'immunoglobuline membranaire, ici une IgM, à la surface du lymphocyte B, associée au co-récepteur hétérodimère CD79A/CD79B. Le co-récepteur assure la transmission du signal lorsque l'IG membranaire se lie à un antigène. CD79A et CD79B contiennent dans leur région cytoplasmique des motifs spécifiques appelés ITAM, riches en tyrosines dont la phosphorylation permet le recrutement de molécules de signalisation qui appartiennent à au moins deux familles de protéines tyrosine kinases (PTK), la famille Syk et la famille Tec, et assurent la transmission du signal (avec la permission de IMGT®, http://www.imgt.org).

L'apparition à la surface cellulaire du lymphocyte B d'un BcR fonctionnel constitue un deuxième contrôle de la maturation du lymphocyte B. En effet, le BcR induit l'arrêt du réarrangement des gènes codant les chaînes légères si le premier réarrangement IGK est productif: inhibition du second allèle IGK (exclusion allélique) et du réarrangement du locus IGL (exclusion isotypique). Ce mécanisme permet d'exprimer une seule et unique chaîne légère. Dans le cas où le premier réarrangement du locus IGK est improductif, une tentative est faite sur le deuxième allèle, et si de nouveau aucun réarrangement productif n'est obtenu, les gènes V et J du locus IGL sont à leur tour réarrangés. Si aucun réarrangement productif de chaîne légère n'est obtenu, la cellule mourra, à moins qu'elle ne soit sauvée par un réarrangement secondaire.

1.2.9 La différenciation des cellules B

Dans la moelle osseuse, les cellules B immatures IgM^+ sont soumises à une sélection négative à l'origine de leur tolérance vis-à-vis des molécules du soi. Un lymphocyte B immature qui se lie à un antigène du soi avec une forte affinité, soit meurt par apoptose, soit réactive la recombinase RAG, entraînant des réarrangements secondaires des locus IGK et IGL, et à moindre degré du locus IGH. Ceci permet au lymphocyte B de générer une nouvelle chaîne légère (et éventuellement une chaîne lourde) et donc de changer la spécificité du récepteur pour l'antigène (processus appelé receptor editing) [64].

L'étape finale de la maturation dans la moelle osseuse correspond à la co-expression des IgM et des IgD. Le lymphocyte IgM^+ IgD^+ est un lymphocyte B mature naïf, capable de répondre à l'antigène dans les tissus lymphoïdes périphériques (Figure 1.3). La cellule B activée prolifère et se différencie en cellules sécrétrices d'anticorps (plasmocytes) et en cellules mémoire. Deux types de réponses distinctes existent selon le type d'antigène et l'existence ou non de contact avec les lymphocytes Th. Dans le cas d'une réponse indépendante des lymphocytes T, la réponse est rapide et représente la première ligne de défense de l'immunité adaptative. Les cellules B matures naïves poursuivent leur différenciation dans la zone marginale des ganglions lymphatiques. Les cellules B stimulées par les polysaccharides généralement présents sur les pathogènes tels que les bactéries encapsulées, se différencient en plasmocytes ou en cellules B mémoire. Dans le cas d'une réponse dépendante des lymphocytes Th, les cellules B matures naïves poursuivent leur différenciation à l'intérieur de ces centres germinatifs qui sont des structures spécialisées des follicules. Les cellules B prolifèrent rapidement et se différencient en centroblastes. Au cours de cette prolifération, un taux élevé d'hypermutations somatiques sont introduites dans les régions V-J et V-D-J. Les centroblastes migrent alors dans la zone claire, où elles deviennent des centrocytes et stoppent leur division. Ces centrocytes sont alors sélectionnés selon l'affinité de leurs BcR, par contact avec des antigènes à la surface des cellules folliculaires dendritiques. Les cellules ayant un BcR de faible affinité meurent par apoptose, tandis que les cellules de forte affinité reçoivent un signal de survie (processus de maturation d'affinité). Les centrocytes ainsi sélectionnés pour leur affinité pour un antigène spécifique prolifèrent. Ils entrent en contact avec les lymphocytes Th et subissent le phénomène de la commutation de classe (CSR) qui permet l'expression des isotypes IgG, IgA, et IgE. Finalement ils se différencient en plasmocytes et en cellules B mémoire.

Au niveau moléculaire, les gènes d'IG sont ainsi soumis dans les ganglions lymphatiques à 3 types de modifications: 1) le phénomène d'hypermutations somatiques (HMS) au niveau des régions V-J et V-D-J, 2) la commutation de classe (CSR), et 3) la production d'IG sécrétées. Les hypermutations somatiques et la commutation de classe sont initiées par la même enzyme, la cytidine déaminase (AID). Elle est présente dans les lymphocytes B activés. Il faut noter que, chacun de ces événements peut se produire indépendamment l'un de l'autre.

1.3 Organisation et localisation chromosomique des locus et des répertoires potentiels.

1.3.1 Le locus humain IGH

Le locus IGH humain est localisé sur le chromosome 14 [65], à la bande 14q32.33 du bras long [66, 67]. Le locus IGH (Figure 1.17) comprend de 123 à 129 gènes IGHV [68-73] localisés sur une distance de plus de 900 kilobases (kb), dont 38 à 46 sont fonctionnels (Tables 1.1, 1.2) et sont répartis en 6 à 7 sous-groupes. Le locus IGH comprend également 27 gènes IGHD [74-77], dont 23 sont fonctionnels, disposés en tandem sur une distance de 9 kb, tandis que le gène IGHD7-27 est situé à 100 pb (paires de base) en 5' des gènes IGHJ. Il y a 9 gènes IGHJ [77, 78] localisés sur une distance de 8 kb, dont 6 gènes sont fonctionnels. Finalement, le locus IGH comprend 11 gènes IGHC [57, 79-89] situés sur une distance de 300 kb et dont un est pseudogène et un est ORF. Les gènes IGHC à l'exception du gène IGHGP qui est ORF, sont précédés d'une séquence switch qui joue un rôle important dans la commutation de classe.

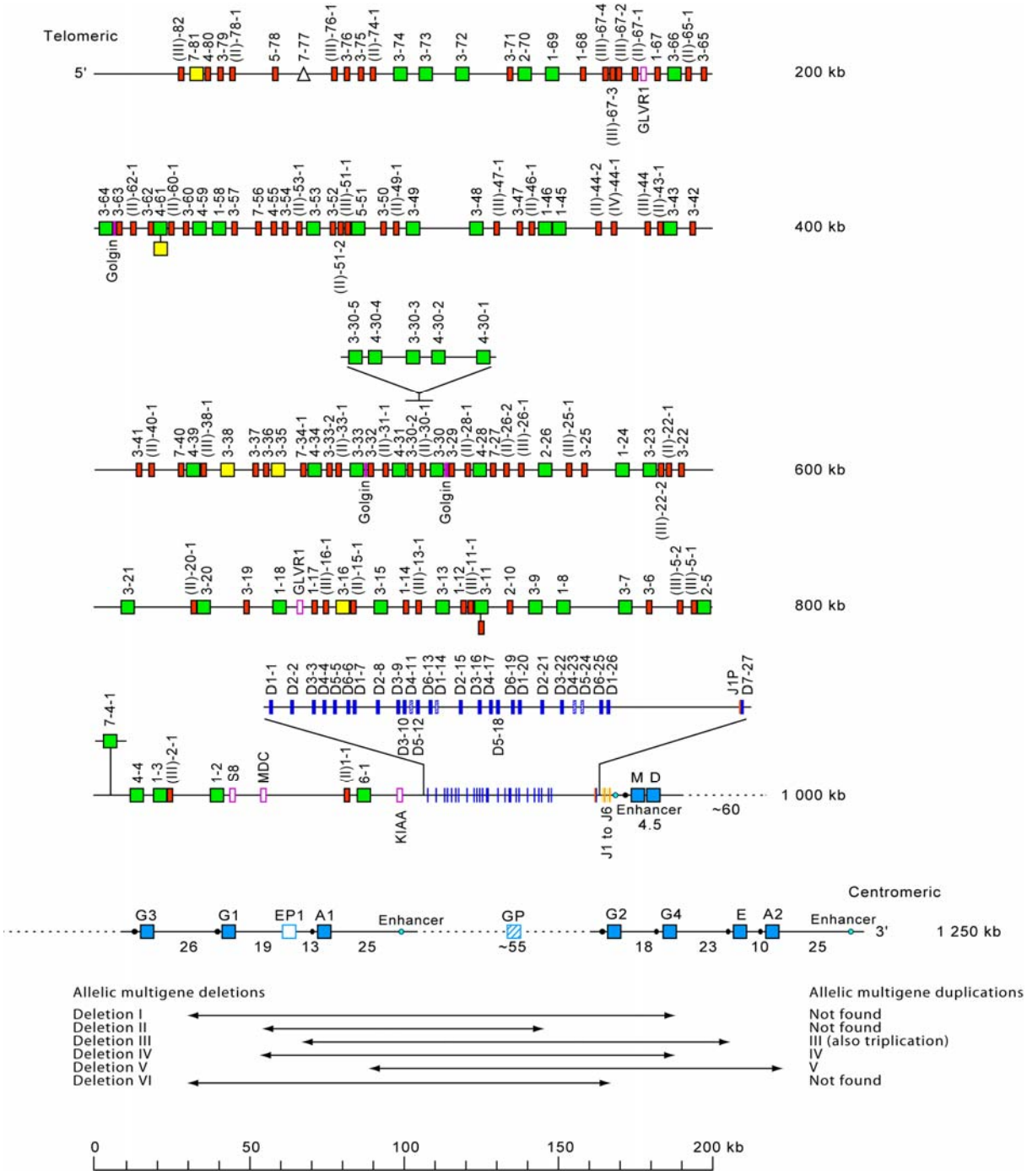


Figure 1.17: Représentation schématique du locus IGH chez l'homme. Le locus IGH comprend de 123 à 129 gènes IGHV dont 38 à 46 sont fonctionnels, 27 gènes IGHD, dont 23 sont fonctionnels et 9 gènes IGHJ, dont 6 sont fonctionnels. Finalement, le locus IGH comprend 11 gènes IGHC dont deux sont des pseudogènes. La plupart des gènes IGHC sont précédés d'une séquence de switch qui joue un rôle important dans la commutation de classe. Sont représentés en vert les gènes V fonctionnels, en jaune clair les gènes V ORF, en rouge les gènes V Pseudogènes, en bleu (carrés) les gènes C fonctionnels, en bleu (traits) les gènes D fonctionnels et en jaune foncé les gènes J fonctionnels [3] (avec la permission de M.-P. et G. Lefranc, IMGT®, http://www.imgt.org).

Tableau 1.1: Nombre total de gènes d'IG par gènome haploïde ([3] avec la permission de M.-P. et G. Lefranc, IMGT®, http://www.imgt.org).

Locus	locus						Nombre d'orphons	Nombre total de gènes (incluant les orphons)
	Localisation chromosomique	V	D	J	C	Nombre total de gènes dans les locus (sans les orphons)		
IGH	14q32.33	123-129	27	9	11[a]	170-176[b]	36[c]	206-212[b,c]
IGK	2p11.2	(40[d] ou) 76	0	5	1	(46[d] ou) 82	25	(71[d] ou)-107
IGL	22q11.2	73-74	0	7-11	7-11	87-96	7[e]	93-103

[a]Des délétions multigéniques, des duplications et triplications alléliques des gènes IGHC ont été décrites chez des individus sains. Le nombre de gènes IGHC peut varier de 5 (délétion I, dans la figure 1.17) à 19 (triplication III, dans la figure 1.17), par génome haploïde.
[b]Comprend les 7 gènes IGHV non localisés.
[c]inclut le gène 'processé' IGHEP2 localisé sur le chromosome 9 (9p24.2-p24.1).
[d]Nombre de gènes dans le rare haplotype IGKV dépourvu du V-CLUSTER distal.
[e] Inclut le gène processé IGLJ-C/OR18
Les locus comprennent le locus IGH (14q32.33), le locus IGK (2p11.2), et le locus IGL (22q11.2). Ces gènes sont impliqués dans la synthèse des chaînes d'immunoglobulines. Les orphons sont localisés en dehors des principaux locus, et ne contribuent pas à la synthèse des chaînes d'IG.
25 IGHV, 10 IGHD, 25 IGKV, 4 IGLV, 2 IGLC orphons ont été identifiés.
Les deux gènes 'processés' d'immunoglobulines décrits à ce jour, IGHEP2 et IGLJ-C/OR18, ont été inclus avec les orphons, dans ce tableau.

Tableau 1.2: Nombre de gènes d'IG fonctionnels par génome haploïde. (avec la permission de M.-P. et G. Lefranc, [3], IMGT®, http://www.imgt.org).

Locus	Localisation chromosomique	Taille des locus (kb)	V	D	J	C	Nombre de gènes fonctionnels	Ordre de grandeur théorique de la diversité combinatoire
IGH	14q32.33	1250	38-46	23	6	9[a]	76-84	38 × 23 ×6 = 5244 (m) 46 × 23 × 6 = 6348 (M)
IGK	2p11.2	1820	34-38	0	5	1	40-44	34 × 5 = 170 (m) 38 × 5 = 190 (M)
		500[b]	17-19[b]	0	5	1	23-25[b]	17 × 5 = 85 (m)[b] 19 × 5 = 95 (M)[b]
IGL	22q11.2	1050	17-19	0	4-5	4-5	37-43	29 × 4 = 116 (m) 33 × 5 = 165 (M)

[a]Il existe des haplotypes avec délétions multigéniques (voir figure 1.17). Le nombre de gènes fonctionnels IGHC est de 5 (délétions I, III, et V), 6 (délétions IV et VI), ou 8 (délétion II), par génome haploïde. Dans le cas des haplotypes avec duplication ou triplication multigénique, le nombre de gènes fonctionnels IGHC par génome haploïde n'est pas connu.
[b]Dans le rare haplotype IGKV dépourvu du V-CLUSTER distal.
L'ordre de grandeur théorique de la diversité combinatoire prend en compte le nombre de gènes V, D et J minimum (m) et maximum (M) dans les principaux locus IGH, IGK et IGL.

1.3.2 Le locus humain IGK

Le locus IGK humain est localisé sur le chromosome 2 [90], sur le bras court, à la bande 2p11.2 [91]. Le locus kappa (IGK) (Figure 1.18) comprend 76 gènes IGKV [92-98] dont 34 à 37 sont fonctionnels qui appartiennent à 5 sous-groupes. Il existe 5 gènes IGKJ [92, 98, 99] situés en 3' des gènes IGKV et à 2,5 kb en 5' de l'unique gène IGKC [100] qui code la région constante (C) des chaînes légères kappa (Tables 1.1, 1.2).

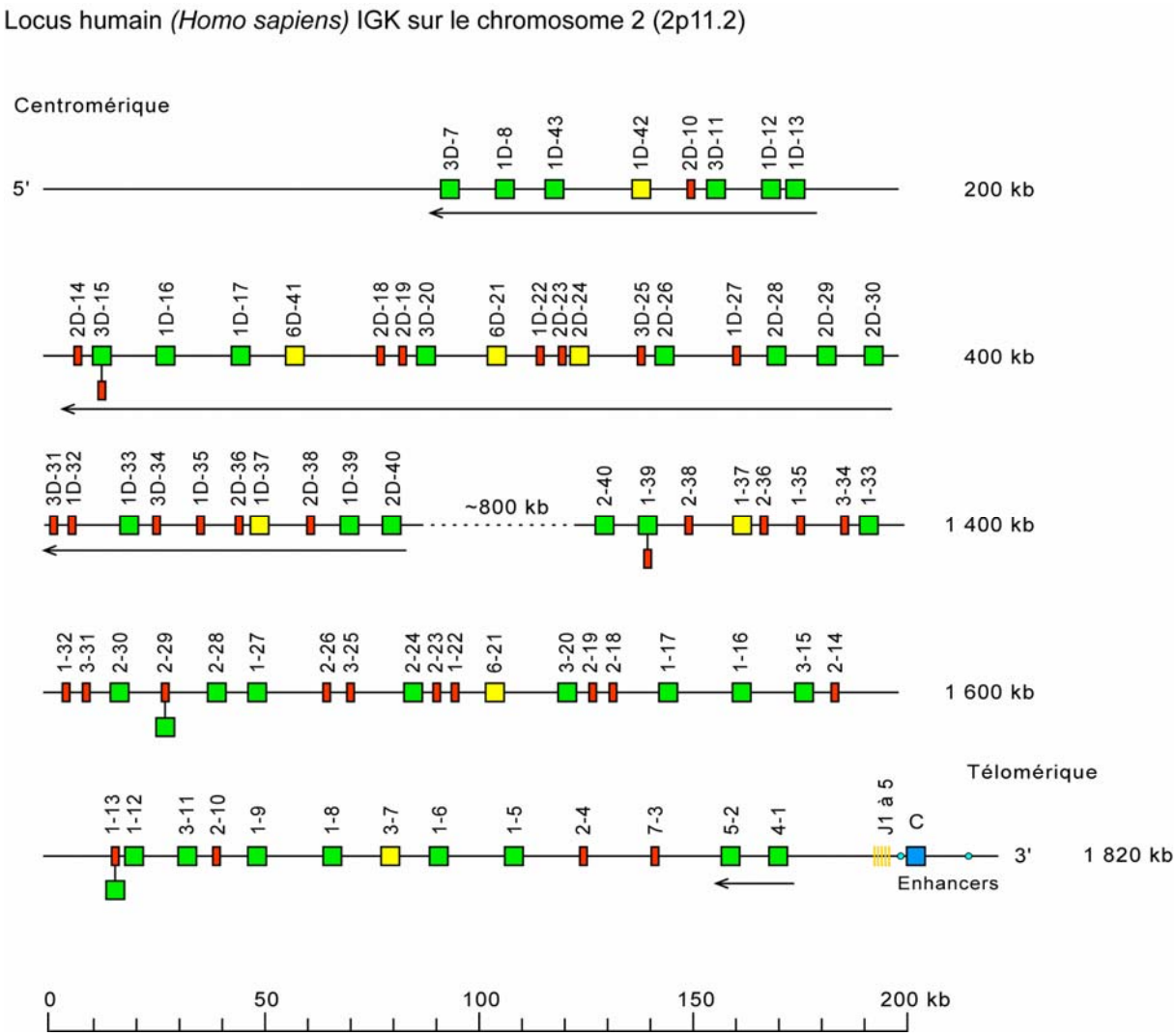


Figure 1.18: Représentation schématique du locus IGK chez l'homme. Le locus IGK comprend 76 gènes IGKV dont 34 à 37 sont fonctionnels, 5 gènes IGKJ et un unique gène IGKC [3] (avec la permission de M.-P. et G. Lefranc, IMGT®, http://www.imgt.org).

1.3.3 Le locus humain IGL

Le locus IGL humain est localisé sur le chromosome 22 [101], sur le bras long, à la bande 22q11.2 [102]. Le locus lambda (IGL) (Figure 1.19) comprend de 70 à 74 gènes IGLV

[92, 103-107] dont 29 à 35 sont fonctionnels qui appartiennent à 10 sous-groupes (Tables 1.1, 1.2). Le nombre de gènes IGLC varie chez l'homme de 7 à 11, dont 4 au moins sont fonctionnels et sont en tandem sur une distance de 50 kb à 70 kb. Chaque gène IGLC fonctionnel est précédé en 5' d'un gène IGLJ [108-111] situé à 1,5 kb.

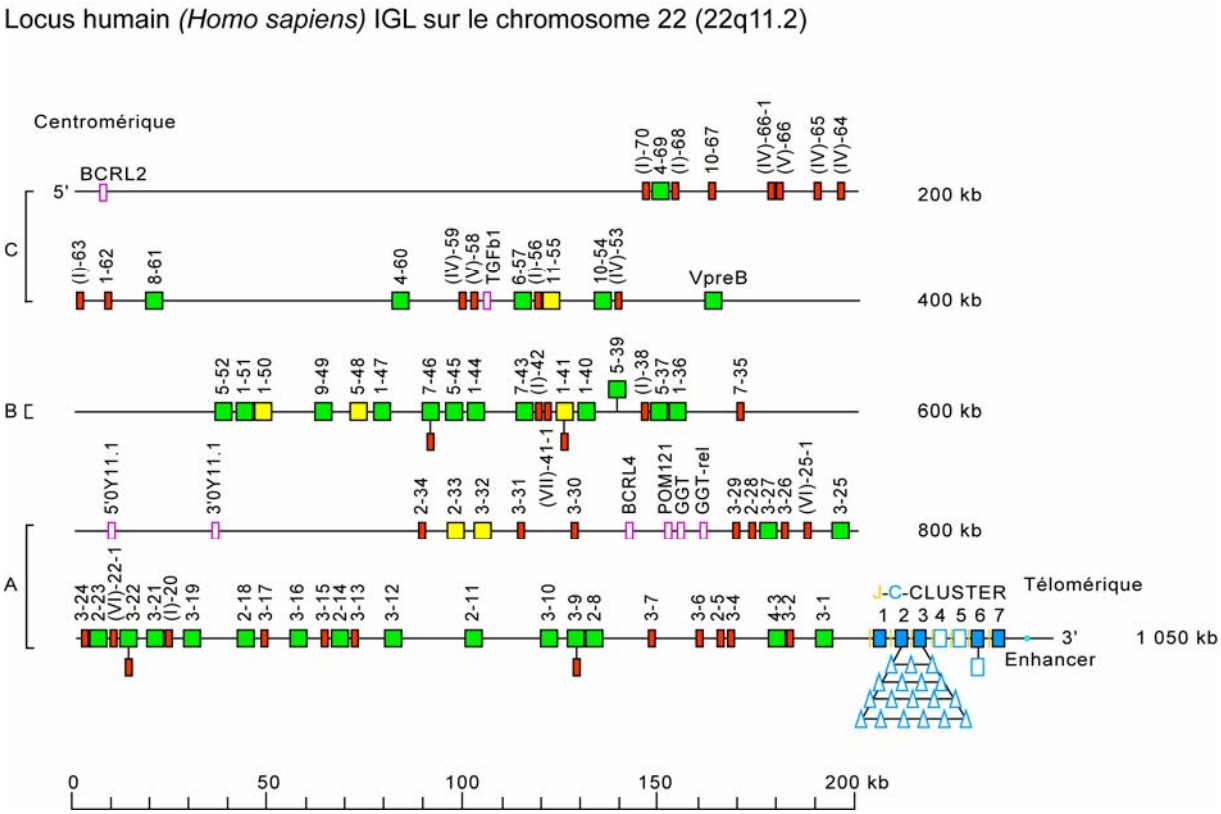


Figure 1.19: Représentation schématique du locus IGL chez l'homme. Le locus IGL comprend 70 à 74 gènes IGLV dont 29 à 35 sont fonctionnels. Le nombre de gènes IGLC varie chez l'homme de 7 à 11, dont 4 au moins sont fonctionnels, et chaque gène IGLC fonctionnel est précédé en 5' d'un gène IGLJ situé à 1,5kb [3] (avec la permission de M.-P. et G. Lefranc, IMGT®, http://www/imgt.org).

1.4. Concepts IMGT de description et de numérotation

1.4.1 Concepts de description: prototypes

Afin de décrire de manière standardisée les récepteurs d'antigènes, IMGT® a établi des règles basées sur les concepts d'IMGT-ONTOLOGY [12, 13]. Dans cette section, nous citons les principaux 'labels' (en majuscules) couramment utilisés pour la description des IG. Les IG sont des hétérodimères constitués de quatre chaînes polypeptidiques, deux chaînes lourdes (H) transmembranaires dont la région constante définit la classe ou isotype d'IG et deux chaînes légères (Figure 1.2 et Figure 1.20).

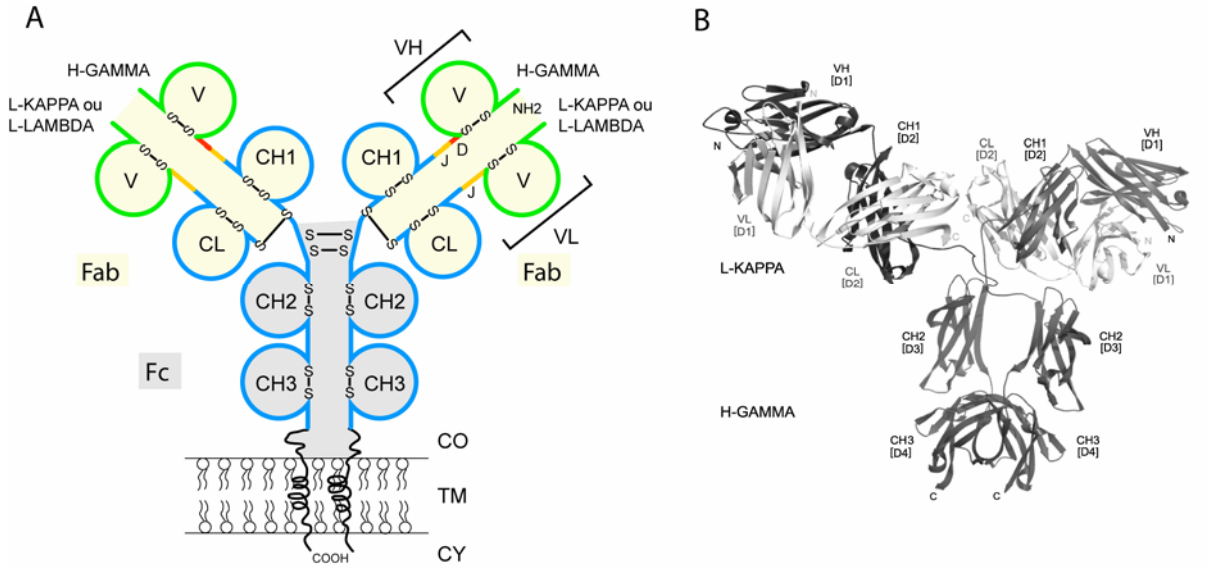


Figure 1.20: présentation schématique d'une IG (IgG1 humaine) (A) Une IgG1 est constituée de deux chaînes lourdes identiques et deux chaînes légères identiques L-KAPPA et H-GAMMA ou L-LAMBDA. Chaque chaîne légère comporte deux domaines: VL (V-KAPPA ou V-LAMBDA) et CL (C-KAPPA ou C-LAMBDA); chaque chaîne lourde comporte les domaines VH, CH1, CH2, CH3, les régions CO, TM et CY (respectivement CONNECTING-REGION, TRANSMEMBRANE-REGION et CYTOPLASMIC-REGION). V, D et J indiquent les séquences codées par chacun des gènes réarrangés pour chaque chaîne (V-D-J pour les chaînes lourdes et V-J pour les chaînes légères), respectivement V-REGION, D-REGION et J-REGION; C indique la C-REGION. Les 2 fragments Fab (liaison avec l'antigène par les domaines variables) et le fragment Fc (propriétés effectrices) obtenus par la digestion de la papaïne sont respectivement indiqués par un fond jaune et gris. (B) Structure 3D d'une IG (IgG1 humaine). Les régions CO, TM et CY non présentes dans la structure 3D ne sont pas montrées (avec la permission de IMGT®, http://www.imgt.org).

Ces chaînes sont caractérisées par leur poids moléculaire: par exemple respectivement 50 kilodaltons (Kd) et 25 Kd pour H-GAMMA et L-KAPPA (ou L-LAMBDA) [112]. La digestion enzymatique d'une IG par la papaïne produit 2 fragments Fab ('antigen binding') et 1 fragment Fc ('cristallisable'). Les chaînes lourdes des IG membranaires sont constituées d'un domaine variable (VH), d'une région constante comprenant trois ou quatre domaines constants (CH), d'une région charnière ou HINGE (excepté pour les IgM et IgE), d'une région de connexion (CONNECTING-REGION, CO) d'une région transmembranaire (TRANMEMBRANE-REGION, TM) et d'une courte région cytoplasmique (CYTOPLASMIC-REGION, CY) (Figure 1.20 Les chaînes lourdes des IG sécrétées (Figure 1.2) sont dépourvues des régions CO, TM et CY.

Le domaine variable (VH) est codé par la V-D-J-REGION issue du réarrangement de 3 gènes, variable (V), diversity (D) et joining (J). Chaque région constante est codée par un gène comprenant 3 ou 4 exons, suivant le nombre de domaines CH, plus un ou plusieurs (pour les IgG3) petits exons pour la région charnière quand celle-ci existe (les IgM et les IgE en sont dépourvues). Les chaînes légères kappa ou lambda comprennent un domaine variable

respectivement V-KAPPA ou V-LAMBDA, codé par la V-J-REGION issue du réarrangement d'un gène V et d'un gène J et d'un domaine constant respectivement C-KAPPA ou C-LAMBDA.

La description standardisée des séquences a permis de définir des prototypes (Figure 1.21) pour l'ADNg (V-D-J-GENE) et l'ADNc (L-V-D-J-C-SEQUENCE).

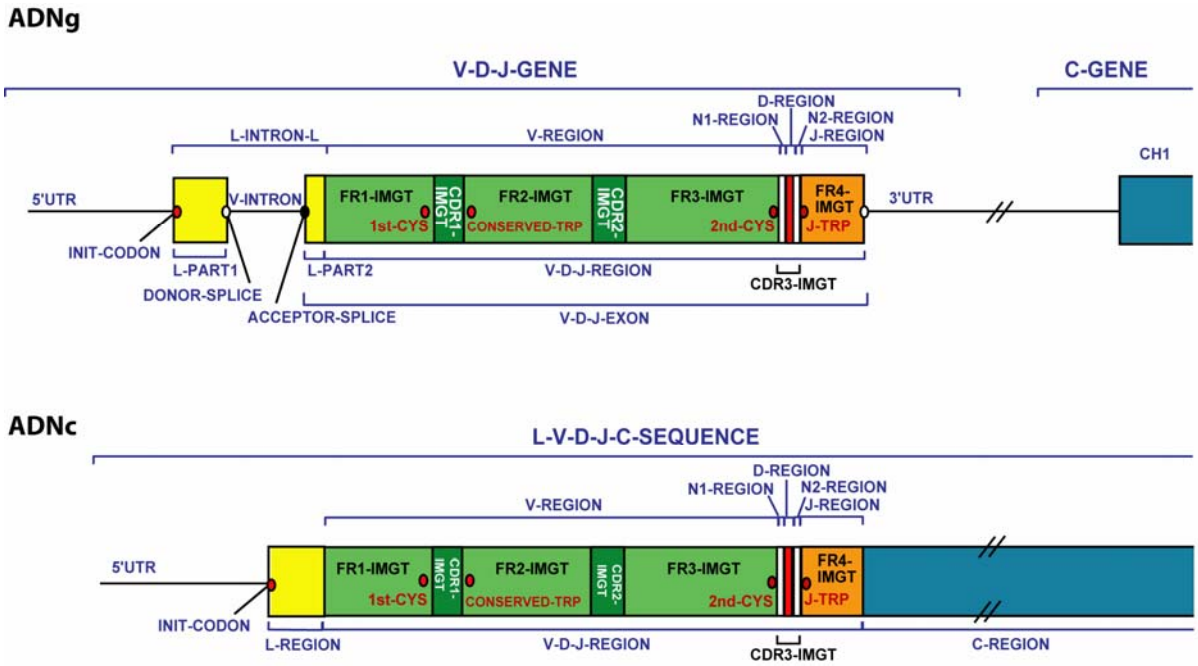


Figure 1.21: Prototypes V-D-J-GENE et L-V-D-J-C-SEQUENCE. Ces prototypes correspondent à l'organisation moléculaire des séquences réarrangées IGH de l'ADN génomique (ADNg) et de l'ADN complémentaire (ADNc). Dans l'ADNg, la séquence réarrangée comprend deux exons: le L-PART1 et le V-D-J EXON. Le V-D-J EXON code le L-PART2 et la V-D-J-REGION. Dans l'ADNc, la L-V-D-J-C-SEQUENCE comprend la région codante complète (L-REGION, V-D-J-REGION et C-REGION). La V-D-J-REGION code le domaine VH. Le V-DOMAIN est constitué de 4 FR-IMGT et de 3 CDR-IMGT. Les acides aminés conservés sont indiqués, 1st-CYS 23, CONSERVED-TRP 41 et 2nd-CYS 104 (avec la permission de IMGT®, http://www.imgt.org).

Le V-DOMAIN est une unité structurale caractérisée par un repliement en sandwich beta, constitué de 9 brins beta antiparallèles (A, B, C, C', C'', D, E, F et G) organisés en deux feuillets [3] (Figure 1.20, Figure 1.22). Les deux feuillets sont maintenus par un pont disulfure entre les brins B et F de deux cystéines très conservées 1st-CYS 23 et 2nd-CYS 104. Le V-DOMAIN est constitué de 4 régions relativement invariantes appelées régions charpentes ou framework (FR-IMGT) qui ont pour rôle de maintenir la structure du domaine variable, et 3 régions hypervariables (complementarity determining region) CDR-IMGT qui forment des boucles et constituent le site de liaison à l'antigène et confèrent la spécificité de l'anticorps. Le CDR3-IMGT est celui qui présente la plus grande variabilité. Il correspond en effet à la

jonction des gènes V, (D) et J et contribue le plus à la spécificité de la reconnaissance et de la liaison à l'antigène. La JUNCTION correspond au CDR3-IMGT avec ses deux ancres 2nd-CYS 104 et J-PHE 118 (VL) ou J-TRP 118 (VH).

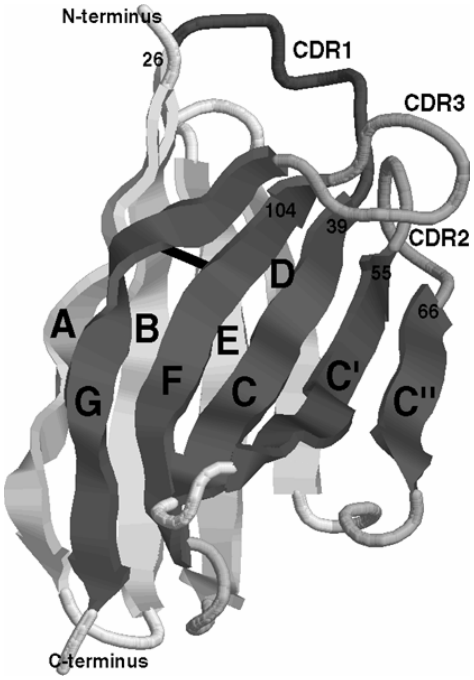


Figure 1.22: Structure 3D des V-DOMAIN des IG. Les V-DOMAIN sont constitués de 2 feuillets; le pont disulfure entre les brins B et F est impliqué dans le maintien de la structure de ce domaine.

Un C-DOMAIN est une unité structurale caractérisée par un repliement en sandwich beta, constitué de 7 brins beta antiparallèles (A, B, C, D, E, F et G) organisés en deux feuillets. Les deux feuillets sont maintenus, comme dans le domaine V, par un pont disulfure, entre les brins B et F, entre 1st-CYS 23 et 2nd-CYS 104. Les brins C' et C'' et le CDR2-IMGT sont absents et remplacés par un brin tranversal CD.

1.4.2 IMGT unique numbering et IMGT Collier de Perles

La numérotation unique des V-DOMAIN [14] (Figure 1.21) établie par Marie-Paule Lefranc en 1997 [14, 113, 114] a permis pour la première fois de standardiser la description des domaines V quels que soient le type de récepteur, le type de chaîne et l'espèce. De plus, ces caractéristiques sont valables aussi bien pour les séquences que pour les structures 3D (Figure 1.20, Figure 1.22). Les acides aminés conservés comprennent les deux cystéines en position 23 (1st-CYS) et 104 (2nd-CYS) (impliquées dans le pont disulfure), un trytophane en position 41 (CONSERVED-TRP) et un acide aminé hydrophobe en position 89. La numérotation unique a permis de définir avec précision les positions de début et de fin des

FR-IMGT et des CDR-IMGT. En particulier la longueur des CDR-IMGT devient en soi une information importante dans la description des V-DOMAIN. Chaque position est donc associée à une localisation structurale et parfois à une propriété physico-chimique. Il est possible de visualiser les V-DOMAIN et leur relation avec la numérotation unique sous forme de IMGT Colliers de Perles (Figure 1.23) [115, 116]. La numérotation unique des V-DOMAIN a de nombreux avantages. Elle permet d'une part de décrire les mutations, le polymorphisme allélique et les hypermutations somatiques de façon standardisée, mais également de visualiser la topologie d'un domaine en l'absence de structure 3D.

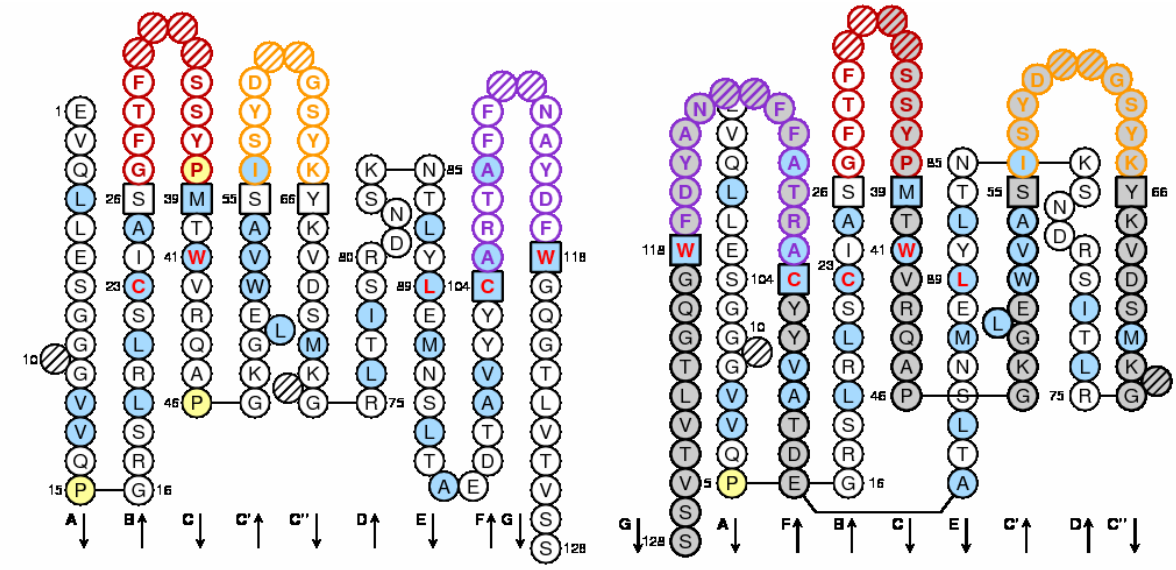


Figure 1.23: IMGT Collier de Perles d'un V-DOMAIN. Domaine V d'une chaîne lourde humaine (AB012909), représenté sur un plan (à gauche) et sur deux plans (à droite); les CDR-IMGT sont coloriés en rouge (CDR1-IMGT), orange (CDR2-IMGT) et violet (CDR3-IMGT). Les acides aminés conservés sont écrits en rouge. Les acides aminés hydrophobes et les résidus tryptophane présents dans plus de 50% des séquences analysées sont représentés en bleu. Les résidus proline sont en jaune. Les positions délimitant les CDR-IMGT sont représentées sous forme de carrés; ces positions appartiennent aux FR-IMGT. Les positions hachurées correspondent à des positions non occupées dans ce domaine, par rapport à la numérotation unique IMGT. Les flèches indiquent l'orientation des brins et leurs labels (avec la permission de IMGT®, http://www.imgt.org).

CHAPITRE 2

Leucémies lymphoïdes chroniques

La leucémie lymphoïde chronique (LLC) est la plus commune des leucémies chez l'adulte en occident. Elle a pour origine une accumulation de lymphocytes B monoclonaux matures dans le sang périphérique ou sang circulant, la moelle osseuse, et les organes lymphoïdes secondaires (ganglions lymphatiques, rate) [117, 118]. L'évolution clinique de la LLC est extrêmement variable: les courbes de survie peuvent aller de quelques mois à plusieurs décades. Certains patients n'ont aucun signe clinique pendant toute l'évolution de la maladie, et ont une survie semblable aux sujets sains du même âge et du même sexe. En revanche, d'autres patients auront une aggravation rapide de leur hémogramme, un envahissement ganglionnaire et une altération de leur état général nécessitant des traitements. Actuellement, aucun traitement ne peut être considéré comme curatif et tous les patients atteints de LLC mourront avec ou à cause de leur maladie. Au cours des dix dernières années, des apports considérables dans la connaissance de l'évolution et l'origine des cellules leucémiques ont permis de redéfinir la maladie. La LLC se différencie en deux sous-types de maladie, les deux ayant pour origine des lymphocytes B matures mais qui différèrent par l'état mutationnel des gènes IGHV réarrangés. Il existe également un défaut constitutif de l'apoptose pour ces cellules leucémiques qui s'accumulent dans l'organisme. Ces dernières années, la recherche sur la LLC a donné lieu à d'importantes découvertes qui permettent de pronostiquer la maladie chez les patients au moment du diagnostic. Ces découvertes correspondent à de nouveaux marqueurs moléculaires et protéiques des lymphocytes leucémiques. Ils sont utilisés pour classifier les patients et évaluer le pronostic, correspondant à des évolutions cliniques différentes (maladie agressive ou indolente), et ainsi adapter les traitements. Ceci est d'autant plus important que la LLC reste, à ce jour, une pathologie incurable.

2.1 Présentation clinique

2.1.1 Incidence et prévalence

La LLC est la plus fréquente des leucémies chez l'adulte en Europe et en Amérique du Nord, alors qu'elle est rare sur le continent asiatique. La LLC représente 25 à 30% de toutes les leucémies [118], et environ 15.000 nouveaux cas sont diagnostiqués chez les adultes aux Etats-Unis chaque année [119, 120]. La LLC est plus commune chez l'homme que chez la femme, avec un sexe ratio d'environ deux hommes pour une femme [118]. Le taux d'incidence tourne autour de 2 à 6 nouveaux cas pour 100.000 individus par an et augmente avec l'âge, pour atteindre 12,8 nouveaux cas pour 100.000 individus par an à 65 ans. Elle atteint principalement les sujets âgés de plus de 60 ans. Néanmoins environ un tiers a moins de 60 ans. La LLC est extrêmement rare au-dessous de 40. La moyenne d'âge se situe autour de 65 ans [118]. Récemment, une augmentation de l'incidence parmi les jeunes individus a été rapportée: 1/3 des nouveaux cas diagnostiqués avant l'âge de 55 ans [121]. En dépit de la découverte de l'augmentation de l'incidence chez les individus 'jeunes', cela n'amène pas de changement substantiel sur l'incidence générale de la maladie [122]. La prévalence de la LLC, c'est-à-dire le nombre de personnes vivantes atteintes actuellement de la LLC, reste autour de 30 à 50 individus pour 100.000 (0,03–0,05%).

2.1.2 Facteurs de risque

Les facteurs environnementaux ne semblent pas jouer un rôle important dans la pathogénie de la maladie. La LLC est la seule leucémie pour laquelle il n'a pas été mis en évidence de corrélation avec l'irradiation ou l'exposition à des composants chimiques. A l'inverse les facteurs génétiques semblent jouer un rôle dans la pathogénie de la maladie. En effet, l'incidence varie selon les pays. Elle représente 3,5% de toutes les leucémies de l'adulte au Japon, alors qu'elle atteint jusqu'à 38% au Danemark. Ce faible taux d'incidence dans les populations orientales est maintenu dans les populations migrantes et chez leur descendance, ce qui permet d'exclure l'existence de phénomènes environnementaux ayant un effet sur les prédispositions génétiques [122, 123]. De plus, des études épidémiologiques récentes démontrent l'existence, dans 5 à 10 % des cas, de prédispositions familiales [124-126], avec au moins deux individus infectés dans une même famille. Le risque de développer la LLC est 2 à 7 fois plus important dans le cas d'une prédisposition familiale, comparé à la population en général [127]. Cette prédisposition familiale est accompagnée par ce que l'on appelle un phénomène d'anticipation [126, 128-130], c'est-à-dire une apparition précoce de la maladie

avec une évolution plus sévère pour les descendants des patients atteints de LLC. En dépit de ces caractères, la LLC familiale est en termes clinique, moléculaire et biologique, très similaire aux cas classiques.

2.2 Diagnostic

2.2.1 Lymphocytose sanguine

Cette hémopathie est le plus souvent découverte au cours d'un examen systématique chez un patient sans symptôme apparent [117, 131]. Elle repose sur la découverte d'une hyperlymphocytose sanguine (augmentation du nombre de lymphocytes) supérieure au seuil de $>5\times10^9$ lymphocytes/L persistante sur plusieurs semaines ou mois. La lymphocytose très variable d'un patient à l'autre, est en moyenne de $30\times10^9L^{-1}$, mais peut atteindre des chiffres beaucoup plus élevés ($> 200\times10^9L^{-1}$).

2.2.2 Aspect cytologique du sang

L'examen morphologique du frottis sanguin constitue l'étape initiale de ce diagnostic et oriente la stratégie des examens ultérieurs. La lymphocytose est constituée de petits lymphocytes matures très proches du petit lymphocyte normal. Toutefois, quelques atypies sont notées: une chromatine moins dense ou très sombre, disposée en grosses mottes et sans nucléole visible. Quelques fois à l'inverse, par un examen attentif du frottis, il est retrouvé des cellules de grandes tailles avec un noyau régulier mais avec une chromatine plus fine et parfois nucléolée et au cytoplasme plus étendu, d'allure prolymphocyte [132]. Ce contingent de cellules anormales ne doit pas dépasser 10% des éléments de la formule sanguine pour porter le diagnostic de LLC (entre 10 et 55%, il s'agit de LLC mixte, au-dessus de 55% de leucémie prolymphocytaire ou LPL). Les anomalies des autres lignées lymphocytaires, entraînant anémie et thrombopénie, ont une valeur pronostique considérable. Elles apparaissent comme signes majeurs de gravité et de mauvais pronostic dans toutes les classifications. Néanmoins les lignées granuleuses, érythroïdes et plaquettaires sont normales dans une forme débutante de la maladie. Le myélogramme par ponction sternale permet de compléter le diagnostic. Il permet de déterminer si la moelle osseuse présente une forte infiltration lymphocytaire (>30%).

2.2.3 Marqueurs de membrane (Immunophénotype)

L'examen immunophénotypique des lymphocytes est actuellement indispensable pour porter le diagnostic de la LLC. En effet, ces cellules portent des marqueurs caractéristiques de la lignée B (Table 2.1), en particulier le CD19 (Cluster of Differentiation). Le caractère monotypique de la prolifération est révélé par l'expression d'une seule chaîne légère d'IG, kappa ou lambda. Les IG membranaires sont faiblement exprimées à la surface de la cellule leucémique et sont plus fréquemment de type IgM ± IgD (dans une plus faible proportion les IgG ou IgA) [133, 134].

Tableau 2.1: Immunophénotypage de la Leucémie Lymphoïde Chronique déterminé par cytométrie en flux.

Marqueurs	Intensité des IG de surface	CD19	CD22	CD23	FMC7
LLC	Faible	+	-	+	-

Marqueurs	CD5	CD10	CD25	CD11	CD103	CD79
LLC	+	-	+/-	-	-	-

+ : exprimé dans la majorité des cas
+/-: exprimé dans la minorité des cas
-: non exprimé dans la majorité des cas

Une autre caractéristique importante est la présence du CD5 [135]. Le CD5 est un marqueur des lymphocytes T et d'une sous population lymphocytaire B rare (< 5%) chez l'adulte mais majoritaire dans le sang du nouveau né. Le CD5 est également observé dans les lymphomes du manteau (catégorie de lymphome non hodgkinien ou LNH, cancer du système lymphatique) et, dans environ, la moitié des leucémies prolymphocytaires. La présence, presque constante, à la surface de ces cellules lymphoïdes de la LLC du marqueur d'activation CD23, plus rarement du CD25 (récepteur de faible affinité pour l'interleukine 2) et du CD71 (récepteur pour la transferine) révèle qu'il s'agit de cellules activées ou préactivées. En revanche, le FMC7 (épitope du CD20), le CD22 et le CD79B sont peu ou pas exprimés. La co-expression de CD5 avec CD23 est pratiquement spécifique de la LLC, ce qui aide à distinguer la LLC des autres leucémies/lymphomes exprimant le CD5, tel que le lymphome du manteau (MCL, mantle cell lymphoma).

Les lymphocytes de la LLC présentent un phénotype spécifique et homogène. Un système simple basé sur l'étude de ces marqueurs et de ces caractéristiques immunologiques a donc été défini pour aider à la discrimination de la LLC: le score de Matutes (Table 2.2) [133, 134, 136]. Il donne une valeur de 1 à chacun des éléments suivants: expression de CD5 et du

CD23, non expression du FMC7 et du CD79B, faible expression d'une seule immunoglobuline membranaire (caractère monotypique) kappa (60% des cas) ou lambda (40% des cas). La LLC se définit par un score de Matutes supérieur ou égal à 4. Un score à 3 correspond généralement à une «LLC atypique». Les scores ≤ 3 permettent de caractériser les autres syndromes lymphoprolifératifs chroniques B (LPL, HCL…). Un score inférieur à 3 exclut formellement le diagnostic de la LLC [134]. D'autres marqueurs peuvent s'exprimer sur les cellules B de la LLC, en particulier le CD38, marqueur de prolifération dont la valeur pronostique péjorative est associée à un mauvais pronostic. Ce fait a été démontré par plusieurs groupes [137-140].

Tableau 2.2: Evaluation du score de Matutes. Système de score du Royal Marsden Hospital (score de Matutes): les LLC classiques ont un score de 5 ou 4. IGm: Immunoglobuline membranaire. +: Expression; -: Aucune expression.

Marqueurs	1 Point	0 Point
IGm	Faible expression	Expression normale
CD5	+	-
CD23	+	-
FMC7	-	+
CD22, CD79b	Faible expression	Expression normale

2.2.4 Conclusion

Deux examens sont donc nécessaires et suffisants pour établir le diagnostic de la LLC: une numération sanguine avec cytologie et un phénotype lymphocytaire sanguin [141]. Selon les critères du National Cancer Institute-Working group (NCI-WG) criteria [136], la LLC peut être diagnostiquée si les conditions suivantes sont respectées:

(1) Une lymphocytose $>5\times10^{9}$ L^{-1} de petits lymphocytes matures dans le sang périphérique depuis plus d'un mois.
(2) L'immunophénotype détecté par cytométrie de flux, doit respecter les caractères suivants:
    a. Expression d'une seule chaîne légère d'immunoglobuline (kappa ou lambda).
    b. Co-expression des marqueurs CD19, CD5 et CD23.
    c. Expression faible des immunoglobulines membranaires et une expression faible ou absence d'expression de CD79B.

Ce phénotype particulier a été proposé dans le passé comme faisant partie d'un système de score de la LLC (score de Matutes) [133, 134]. Dans le cas d'une lymphocytose inférieure à 5×10^9 L^{-1}, une biopsie de la moelle osseuse peut être effectuée pour confirmer le diagnostic. Néanmoins, l'examen de la moelle épinière n'est pas considéré comme nécessaire pour le diagnostic [136].

2.3 Physiopathologie de la LLC

La LLC se présente comme un petit lymphocyte mature au rapport cytoplasmique élevé exprimant CD5, CD23 et les immunoglobulines IgM et IgD. En se basant sur ces caractéristiques cytologiques et phénotypiques, elle a longtemps été considérée comme dérivant d'un lymphocyte B naïf au repos. Les études génétiques et phénotypiques de ces dernières années ont bouleversé ce concept et démontré que la cellule de LLC n'est pas naïve. La rencontre avec un antigène est un évènement majeur dans le développement d'un lymphocyte B. Dans la LLC sont retrouvées des cellules B présentant des mutations somatiques et d'autres sans mutations somatiques. En effet, les gènes V des IG sont non mutés (identiques aux gènes germline) dans un peu moins de la moitié des cas et présentent des mutations somatiques dans l'autre moitié. Le lymphocyte de la LLC ressemble plus à un lymphocyte B mémoire. De plus, d'autres éléments phénotypiques ou moléculaires plaident contre le caractère naïf du lymphocyte de la LLC, avec en particulier, l'expression constante du marqueur B mémoire CD27. Enfin, le lymphocyte de la LLC exprime le plus souvent des IgM et IgD. Cependant, une fraction du clone poursuit, *in vivo*, un processus dynamique de switch conduisant à l'expression d'IgG ou d'IgA. Il est, ainsi aujourd'hui, établi que la cellule de la LLC a fait l'expérience de l'antigène. Cette rencontre avec l'antigène serait une étape déterminante de la leucogenèse. La LLC est définie par la prolifération monoclonale d'une population mature de lymphocytes B (il peut s'agir d'un lymphocyte de type T dans 5% des cas) qui vont envahir progressivement le sang, les organes lymphoïdes et la moelle osseuse. Les mécanismes cellulaires, induisant la transformation des lymphocytes B en cellules leucémiques de LLC et leur prolifération, restent imparfaitement connus.

2.3.1 Diminution de l'apoptose

Il est admis qu'il existe dans les lymphocytes B de LLC une dérégulation des gènes impliqués dans l'apoptose, alors que les gènes impliqués dans la régulation du cycle cellulaire

ne sont pas affectés [142-144]. Une des caractéristiques importantes du lymphocyte de LLC est la présence d'une surexpression de la protéine anti-apoptotique Bcl-2, dont les mécanismes restent à éclaircir. Dans la majorité des cas, la région promotrice du gène est hypométhylée [145] ce qui peut contribuer à une augmentation de la transcription de cette protéine, et par conséquent à une résistance constitutive à l'apoptose. Il n'est pas encore expliqué si ce phénomène est acquis durant la leucogenèse ou si la LLC provient de cellules hyperexprimant cette protéine. La présence de la surexpression de Bcl-2, chez la majorité des patients, est associée à d'autres dérégulations de protéines de la même famille, tel que Mcl-1. De plus, deux microRNAs régulent négativement Bcl-2, miR-15a et miR-16-1 (situés en 13q14.3) et sont, dans la majorité des LLC, délétés ou hypoexprimés [146, 147]. Toutefois, il est important de considérer que d'autres régulateurs classiques de l'apoptose, tels la P53 et ATM sont également altérés dans les cellules leucémiques [148, 149].

2.3.2 Prolifération

Outre un dysfonctionnement des mécanismes de l'apoptose observé chez tous les patients atteints de LLC, plusieurs indices permettent de mettre en évidence une part proliférative dans la population B leucémique. Cette prolifération a été mise en évidence dans la moelle osseuse et dans les ganglions, au sein des centres de prolifération ou pseudofollicules [150]. Les cellules de LLC présentent une part proliférative avec un taux de croissance compris entre 0,1 et 1% de clone par jour [150], dans certains cas il peut être supérieur à 1%, ce qui conduit pour un patient possédant approximativement 10^{12} cellules leucémiques, à la fabrication par jour de 10^9 à 10^{10} nouvelles cellules leucémiques. Ce taux de division cellulaire est suffisant pour permettre l'apparition de nouveaux clones. Il existe une association entre les taux d'apparition de cellules leucémiques et la progression de la maladie. Ainsi le taux d'apparition des cellules peut être un facteur clinique significatif car il reflète la capacité proliférative des cellules leucémiques et leur potentiel à promouvoir les lésions de l'ADN.

2.3.3 Rôle de la stimulation antigénique du récepteur des cellules B (BcR)

Les récepteurs des cellules B (BcR) jouent un rôle très important dans le développement de la LLC. Il semblerait que la stimulation antigénique, serait un facteur important dans le déclenchement de la prolifération et dans l'inhibition de l'apoptose des cellules de LLC (Figure 2.1). Les effets de la stimulation antigénique étant différents selon la LLC, ils aboutiraient à des évolutions cliniques différentes.

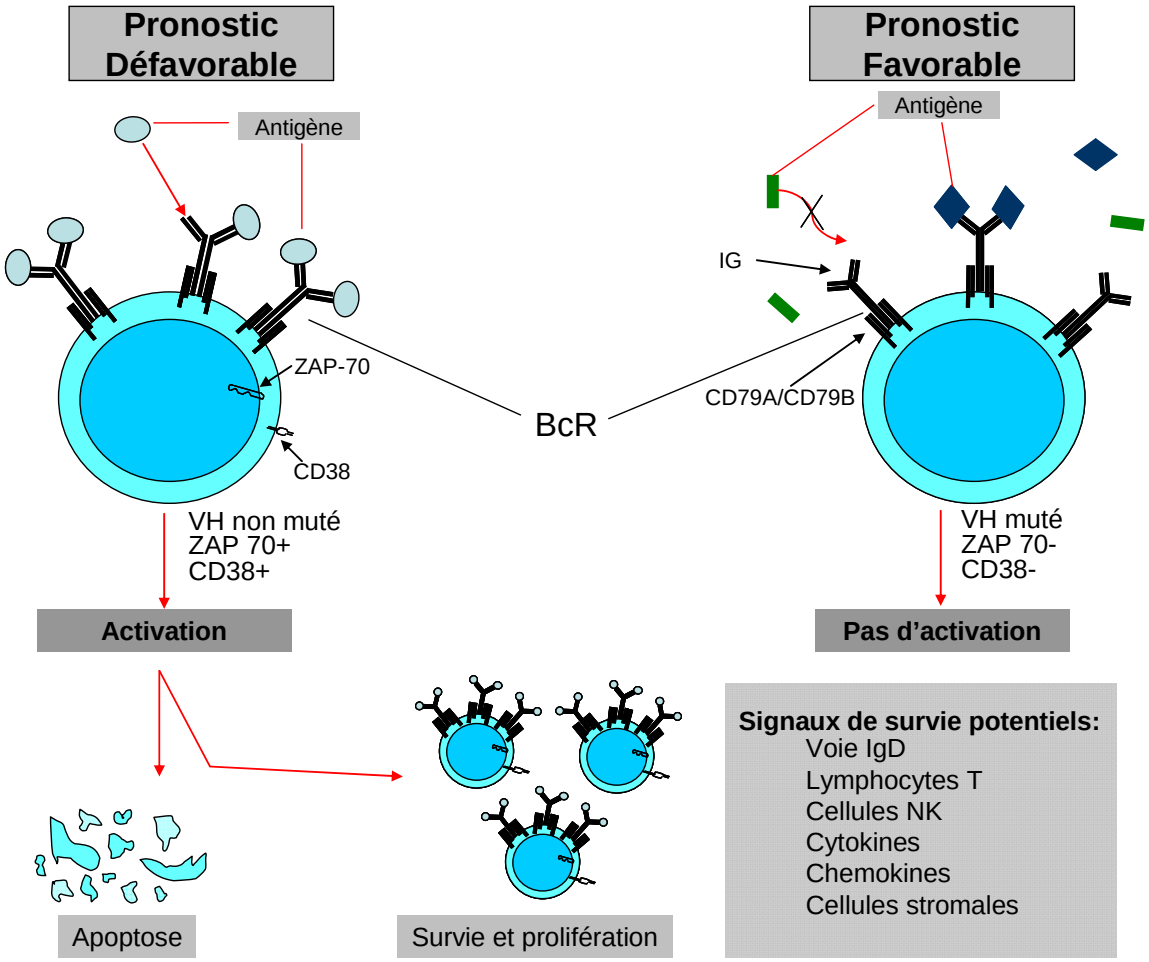


Figure 2.1: Rôle de la simulation d'antigène dans la LLC, d'après Chiorazzi et coll [117]. Les cellules B des patients qui ont des marqueurs pronostiques défavorables (à gauche de la figure) sont stimulées par des antigènes au niveau du BcR. La balance dynamique des signaux positifs et négatifs délivrés par le récepteur B et des signaux de survie transduits par l'IgD et délivrés par d'autres cellules, des cytokines et des chimiokines, va déterminer si la cellule leucémique prolifère ou entre en apoptose. Les cellules B des patients avec LLC qui ont des marqueurs pronostiques favorables (à droite de la figure) sont moins capables de déclencher l'apoptose, la survie et la prolifération, car elles présentent une incapacité soit à lier l'antigène en raison de changements au niveau du BcR, conséquences de mutations du domaine VH de l'IG, soit d'un défaut de transmission du signal par ce récepteur.

Les structures des BcR des cellules de LLC de divers patients sont très proches, ce qui suggère que les antigènes qui se lient à ces récepteurs sont également très proches voire semblables et sont déterminants dans la pathogenèse de la maladie [151, 152]. Ces ressemblances signifiant qu'un nombre restreint d'antigènes est capable d'induire la division des lymphocytes B leucémiques et d'augmenter la fréquence de survenue des mutations de l'ADN. La nature de ces antigènes reste, jusqu'à ce jour, encore inconnue. Il est possible d'émettre l'hypothèse que des virus latents ou des bactéries commensales activent répétitivement des clones de cellules B spécifiques par leur BcR. La LLC serait, alors, la conséquence directe ou indirecte d'infections spécifiques et serait entretenue par celles-ci d'une manière semblable à ce qui a été décrit pour les lymphomes gastriques qui évoluent en

réponse à *Helicobacter pylori* [153]. D'autre part, des antigènes de l'environnement ou des auto-antigènes pourraient entraîner une expansion clonale. Les cellules de LLC ont fréquemment des récepteurs polyréactifs qui lient de multiples antigènes, incluant les auto-antigènes [154-157], ce qui permet leur stimulation simultanée par des auto-antigènes et des antigènes microbiens. Les cellules de LLC ont longtemps été considérées comme anergiques [152]. Or depuis peu, il a été montré que les cellule B de LLC peuvent être activées par le récepteur [158-160] et tout particulièrement dans le cas de LLC non mutées [161-163]. Les cellules qui ne répondent pas à la stimulation du BcR peuvent être figées dans un stade où même les lymphocytes B normaux ne réagiraient pas à l'antigène [151]. Alternativement, ces cellules peuvent être anergiques, certainement en raison d'une expérience antigénique antérieure [152], elles peuvent devenir incapables de répondre à la stimulation d'antigènes dû à un changement dans la structure des BcR par des mutations somatiques où parce qu'elles ont une incapacité à rentrer en contact avec les antigènes *in vivo* [151]. Finalement, cette incapacité à transmettre le signal pourrait être due à un très faible taux de BcR exprimés à la surface des cellules de LLC ou à un défaut d'assemblage entre les différentes protéines du BcR (IGm et ses corécepteurs CD79A/CD79B) [164-166].

Une fois la transduction du signal initiée par le BcR, la cellule va progresser dans le cycle cellulaire ou mourir. La stimulation par une IgM à la surface des cellules de LLC peut entraîner [160] ou inhiber [167] l'apoptose (Figure 2.1), tandis que la stimulation par une IgD de surface inhiberait l'apoptose [160, 168]. Cette différence reste à l'heure actuelle inexpliquée. En conséquence, l'évolution finale de chaque cellule de LLC dépendrait de la balance entre ces signaux [117, 151, 169].

Parallèlement à la stimulation du BcR, il est probable que d'autres facteurs entrent en jeu favorisant probablement la croissance des cellules B, comme par exemple, par des contacts directs de cellule à cellule ou par des facteurs solubles [117, 169]. La résistance spontanée à l'apoptose, *in vivo,* qui définit la maladie, et à l'inverse l'augmentation spontanée de l'apoptose lorsque les cellules sont cultivées *ex vivo,* impliquent que ces cellules ont perdu des facteurs nécessaires à leur survie. En effet, les interactions avec les cellules stromales [170] ou avec les cellules 'nurse-like' [171] à travers le récepteur chemokine, leur permettent d'échapper à l'apoptose par contact direct de cellule à cellule [172]. De la même façon, les interactions entre CD38 et son ligand naturel le CD31 permettent aux cellules leucémiques d'échapper à l'apoptose *in vitro* et probablement *in vivo.* L'interaction CD38/CD31 de surcroît favorise la croissance de la population leucémique, car le contact de ces deux molécules

entraîne l'expression du récepteur CD100, une sémaphorine impliquée dans le maintien de la croissance cellulaire des cellules qui prolifèrent [137, 173]. Les cellules T activées, ou d'autres cellules exprimant le CD40 ligand (CD40L), favorisent également la prolifération des cellules leucémiques. Finalement, les cytokines telles que l'interleukine 4, le facteur de croissance «vascular endothelial growth factor» [174, 175] (VEGF) et les chimiokines telles que SDF1 favorisent l'expansion des clones de LLC. Ces signaux règlent la balance entre les signaux anti-apoptotiques et pro-apoptotiques en faveur de la survie cellulaire. Il y a une augmentation de l'expression des gènes anti-apoptotiques Bcl2, et Mcl-1 dans les cellules leucémiques [176, 177].

Parce que la prolifération d'un clone dépend également, d'une variété d'interactions avec l'environnement, les variations qui résultent de ces interactions avec les cellules leucémiques peuvent être responsables de changements dans l'évolution clinique de la maladie.

2.3.4 Origine cellulaire de la LLC

Une des avancées majeures dans la compréhension de la LLC est survenue quand différents groupes de recherche ont indépendamment démontré que le taux de mutations somatiques dans les gènes réarrangés, codant les domaines variables des chaînes lourdes (VH) des IG, permet de subdiviser la maladie en deux groupes distincts avec des évolutions cliniques différentes [178, 179]. Les IG des cellules de LLC, dont les gènes IGHV réarrangés ont un pourcentage d'identité inférieur à 98% par comparaison avec les séquences germline les plus proches, sont dites «mutées» et sont généralement liées à une maladie plus indolente et un pronostic favorable. Les gènes IGHV réarrangés, avec un pourcentage d'identité supérieur à 98%, sont dits «non mutés» et liés à une maladie agressive et à un pronostic défavorable [178, 179]. Ces éléments permettraient de définir la LLC non plus comme une entité unique mais comme l'existence de deux sous-types de maladies dérivant de deux origines cellulaires différentes, mutées et non mutées, ou de deux stades de différenciation associés à une évolution clinique différente [7]. Les cellules aux IG mutées (pronostic favorable) auraient une origine post-centre germinatif, et dériveraient de cellules des centres germinatifs ou de cellules des zones marginales, et par des mécanismes dépendant ou non des cellules T. Tandis que les cellules non mutées (pronostic défavorable) proviendraient d'une cellule naïve d'origine pré-centre germinatif ou de cellules B des zones marginales, activées par des mécanismes indépendant des cellules T.

D'autre part, l'expression du répertoire des gènes IGHV, IGKV et IGLV des cellules de LLC est biaisée et se distingue du répertoire des cellules B normales [6, 117, 151, 152, 180, 181].

Ce biais est caractérisé, dans la LLC, par une utilisation préférentielle de certains gènes IGHV par rapport aux cellules B normales [180, 181]. Les gènes IGHV les plus fréquents dans la LLC sont IGHV1-69, IGHV3-7, IGHV3-23 et IGHV4-34. Certains gènes: IGHV1-69, IGKV1-33/1D-33 et IGLV3-21, sont préférentiellement utilisés dans les réarrangements non mutés, alors que d'autres gènes: IGHV4-34, IGKV2-30 et IGLV2-8, sont plus fréquents dans les réarrangements mutés. Les mutations somatiques ne semblent pas se produire de façon uniforme entre les gènes IGHV. Par exemple, le gène IGHV1-69 est régulièrement signalé avec très peu de mutations à l'inverse des gènes IGHV3-7, IGHV3-23 et IGHV4-34 qui, sont généralement très mutés.

Récemment, différentes équipes ont démontré (Rosenquist, Stamatopoulos) l'existence dans la LLC de patients présentant une très forte similarité de leurs BcR, dit 'stéréotypés', qui utilisent les mêmes gènes dans leur réarrangement et possèdent, de plus, un CDR3 quasiment identique [7]. Une utilisation préférentielle du gène IGHV1-69 a été observée dans les LLC et plus particulièrement dans les séquences non mutées. Les séquences réarrangées montrent, de plus, un CDR3 conservé avec des acides aminés quasiment identiques [182]. 1,3% des patients dans cette étude (15/1220) avaient des cellules leucémiques qui exprimaient à leur surface des IG pratiquement identiques. Le groupe de Rosenquist [183] a démontré qu'une grande partie des gènes IGHV3-21 réarrangés était caractérisée par un CDR3 court avec une séquence en acides aminés quasiment identique. De plus, le gène IGHV3-21 réarrangé était associé en particulier avec le gène IGLV3-21 de la chaîne légère, montrant ainsi un biais dans l'expression des répertoires des chaînes lourdes et légères. Ces observations suggèrent que les IG exprimées par les cellules B des LLC sont très fortement sélectionnées. En revanche, d'autres équipes [180] ont mis en évidence l'idée que les patients, ayant des BcR stéréotypés, peuvent partager des caractéristiques phénotypiques, cliniques. De ce fait la présence de BcR stéréotypés pourraient avoir une importance dans l'évolution de la maladie (facteur de pronostic) et ceci indépendamment du statut mutationnel des gènes IGHV [183, 184] .

En considérant les événements combinatoires qui se déroulent lors de la synthèse des IG (réarrangement des gènes V, D, J), plus de 1,6 million de combinaisons possibles, ajoutés aux mécanismes de diversité des IG (hypermutation somatique, N-diversité et commutation de classe) [3], les chances d'avoir deux clones de cellules B indépendants qui exprimeraient à leur surface des récepteurs d'antigènes (IG) identiques sont pratiquement nulles. L'existence de BcR stéréotypés est en faveur de l'hypothèse de l'intervention d'un nombre limité d'anticorps dans la leucémogénèse et suggère que la liaison entre un antigène et un récepteur

peut faire une différence en termes de pathogenèse de la maladie, de présentation clinique et par suite, de pronostic [185, 186] .

2.4 Facteurs de Pronostics

La LLC est caractérisée par une évolution clinique hétérogène [187]. Alors que certains patients ont une maladie totalement indolente pendant des dizaines d'années, d'autres meurent plus ou moins rapidement en dépit de tout traitement avec une médiane autour de 7,5 ans [188]. Cette grande hétérogénéité clinique est caractéristique de la LLC et a justifié, depuis des décennies, la recherche de marqueurs de pronostics.

2.4.1 Facteurs de pronostics classiques

Les facteurs de pronostics classiques reflètent vraisemblablement la masse tumorale et/ou sa capacité proliférative. En pratique, l'application clinique des facteurs de pronostic classique reste délicate pour différentes raisons: défaut de standardisation des techniques de mesures, résultats parfois contradictoires des études multivariées, absence d'études prospectives. Actuellement, ce sont les caractéristiques moléculaires, immunophénotypiques et cytogénétiques qui constituent les outils de pronostics les plus performants.

2.4.1.1 Classifications anatomocliniques

L'évaluation clinique standard repose sur les classifications introduites par Rai en 1975, très utilisées aux Etat-Unis [189] (Table 2.3), et Binet en 1981 largement utilisées en Europe [190] (Table 2.4). Elles représentent la première étape indispensable à la décision thérapeutique [189, 190]. Ces systèmes de classification de la maladie, fondés sur des signes cliniques (adénopathies, hépatosplénomégalie) et des paramètres biologiques simples (Numération Formule Sanguin ou NFS ou hémogramme) ont permis de distinguer trois grand groupes de patients ayant des pronostics distincts: des stades précoces (Rai 0 et Binet A), intermédiaires (Rai I/II, Binet B) et avancés (Rai III/IV, Binet C) avec une moyenne de survie estimée respectivement à: 10 ans, 5-7 ans et 1-3 ans. Cependant, il existe une hétérogénéité dans l'évolution clinique au sein d'un même groupe et ces classifications ne permettent pas de prédire précisément à quelle vitesse la maladie va progresser pour un patient se trouvant à un stade précoce (Rai 0, Binet A). Puisque plus de 90% des patients sont actuellement diagnostiqués en stade précoce, il est devenu indispensable d'identifier des marqueurs qui

vont permettre de prédire l'évolution de chacun de ces patients afin d'envisager une thérapeutique adaptée.

Tableau 2.3: Classification anatomoclinique de Rai

Stade		Risque	Survie médiane
0	Lymphocytose sanguine et médullaire isolée	Faible	>15 ans
I	Lymphocytose + adénopathies		9 ans
II	Lymphocytose + splénomégalie et/ou hépatomégalie	Intermédiaire	5 ans
II	Lymphocytose + anémie (Hb < 11 g/dl)	Elevé	2 ans
IV	Lymphocytose + thrombopénie (100.10^9/L)		2 ans

Hb: Hémoglobine

Tableau 2.4: Classification anatomoclinique de Binet

Stade	Pronostic	Syndrome tumoral	Hématopoïèse	% de LLC	Survie médiane
A	Bon	Adénopathies <3 aires palpables	Hb≥10g/dL et plaquettes $\geq 100.10^9$/L	63%	12 ans
B	Intermédiaire	Atteint au moins 3 aires ganglionnaires lymphoïdes	Hb≥10g/dL et plaquettes $\geq 100.10^9$/L	30%	5 ans
C	Mauvais	Quel que soit le nombre d'aires ganglionnaires atteintes	Hb<10g/dL et/ou plaquettes $<100.10^9$/L	7%	2 ans

Aires ganglionnaires: cervicaux, axillaire et inguinaux. Hb: Hémoglobuline.

2.4.1.2 Facteurs de pronostics cliniques

Les principaux facteurs cliniques sont de 2 types:

(a) Le temps de doublement lymphocytaire (LDT), est un marqueur de la prolifération,

mais reste un facteur majeur dans la prise en charge actuelle de la LLC, un LDT inférieur à un an définit un pronostic défavorable et un LDT de moins de 6 mois indique une progression de la maladie [191-193]. La LDT représente même un critère pour débuter un traitement selon le National Cancer Institute sponsored Working Group (NCI-WG) [136].

2.4.1.3 Facteurs de pronostics biologiques

Les principaux facteurs biologiques sont de 5 types:

(a) Le CD23 soluble est un récepteur de faible affinité pour les IgE. La forme soluble se comporte comme une cytokine, induisant la prolifération des lymphocytes B normaux et leucémiques. Dans la LLC, une élévation de son taux sérique a été corrélée à une forte masse tumorale et à un risque accru de progression [194, 195].
(b) L'expression de p53 est corrélée avec une forme aggresive de la maldie et à une résistance au traitement [196].
(c) La β2-microglobulinémie, est une protéine extracellulaire appartenant au complexe HLA de type I. Son taux dans le sérum est inversement corrélé à une diminution de la survie à la fois des patients traités et non traités [197].
(d) La thymidine kinase sérique (TK) est une enzyme impliquée dans le contrôle de la synthèse de l'ADN en particulier des cellules en division. Son taux dans le sérum est inversement corrélé avec le temps de survie, et corrélé avec la prolifération cellulaire [198, 199].

2.4.2 Nouveaux facteurs de pronostics

Récemment, de nouveaux marqueurs biologiques ont été identifiés et permettent une meilleure classification des risques afin de préciser le pronostic pour chaque patient.

2.4.2.1 Statut mutationnel des IGHV

Le marqueur pronostique le plus pertinent, actuellement, repose sur la détermination du statut mutationnel des IGHV [178, 179, 200]. La recherche de mutations somatiques au sein des régions variables des gènes IGHV permet de répartir les patients en deux groupes d'évolution bien distincte, l'absence de mutations des gènes IGHV se définissant par un pourcentage d'identité ≥ 98% à la séquence germline (germinale).
Ainsi, les LLC dites «naïves» ou non mutées, c'est-à-dire développées à partir d'un lymphocyte B naïf n'ayant pas transité par le centre germinatif du follicule lymphoïde

secondaire et n'ayant pas rencontré d'antigènes, sont différenciées des LLC dites «mémoires» ou mutées, c'est-à-dire développées à partir d'un lymphocyte B mémoire qui a transité par le centre germinatif et subi le phénomène d'hypermutation somatique des IGHV. Deux groupes de patients ont été définis:

- Un groupe de patients dont les cellules leucémiques ont des gènes IGHV non mutés (≥98% d'identité avec le gène germline) avec un risque de présenter une pathologie évolutive et une survie raccourcie (pronostic défavorable)
- Un groupe avec des gènes IGHV mutés (<98% d'identité avec le gène germline) [6, 178, 179] avec une évolution favorable et une faible probabilité de développer une maladie agressive (pronostic favorable).

Néanmoins, il existe encore sur cette technologie un débat sur la valeur du seuil permettant de distinguer les séquences mutées des non mutées (98% contre 97%) [201, 202]. Cette recherche implique le séquençage de cette région après amplification par PCR de l'ADN ou de l'ARN extrait des cellules tumorales et sa comparaison avec les séquences germline les plus proches. Actuellement, le séquençage des gènes IGHV n'est pas une technique de routine et le statut mutationnel ne peut donc pas être déterminé pour tous les patients au diagnostic: cet élément majeur du pronostic reste à l'heure actuelle réservé à la recherche clinique.
Il est à noter que certains gènes IGHV semblent échapper à la classification mutée/non mutée et confèrent leur propre valeur pronostique. C'est le cas notamment du gène IGHV3-21 dont la présence est défavorable, indépendamment du pourcentage d'identité [203].

2.4.2.2 Expression de ZAP70

ZAP70 (zeta-chain-associated protein kinase) [204] est une protéine tyrosine kinase qui est très fortement exprimée dans les lymphocytes T et les cellules NK. Elle est impliquée dans la cascade de la signalisation intracellulaire qui transmet l'activation du signal, suite à la reconnaissance d'un antigène par le récepteur des lymphocytes T, et dans les cellules NK par les molécules DAP12 associées aux récepteurs de Fc d'IgG et à d'autres récepteurs caractéristiques. La protéine ZAP70 est rarement présente dans les lymphocytes B normaux matures, mais a été retrouvée dans les cellules B de patients atteints de la LLC. L'étude des profils d'expression génique a montré que les cellules B de LLC avec des gènes IGHV non mutés exprimaient plus de mRNA ZAP70 que des cellules contenant des gènes mutés. Ainsi, la présence de ZAP70 dans les cellules de LLC est corrélée à un pronostic défavorable, et son

absence à un pronostic favorable [162, 205-207]. L'expression d'autres gènes, tels que la Lipoprotein Lipase (LPL est une enzyme ayant un rôle central dans le métabolisme et le transport des lipides) et ADAM29 (gène adisintegrin andmetalloproteinase codant une protéine transmembranaire de la famille des désintégrines et métalloprotéases qui favoriserait l'interaction cellule à cellule et/ou cellule à matrice) permet d'améliorer la corrélation de ZAP70 avec le statut mutationnel [208]. Ainsi, la ZAP70 ou la LPL sont préférentiellement transcrits dans les cellules de LLC présentant des gènes IGHV non mutés, tandis que le gène ADAM29 est majoritairement exprimé par les cellules de LLC présentant des gènes IGHV mutés. Le rapport LPL/ADAM29 a une correspondance de 90% avec le statut mutationnel et représente un marqueur indépendant pour la survie des patients en stade A, B ou C (Binet) [208]. L'analyse des séquences d'ADN pour déterminer le statut mutationnel des gènes IGHV est difficile, longue et coûteuse. Elle ne peut, donc, pas être pratiquée dans les laboratoires d'hématologie de routine. Par opposition, l'expression de ZAP70 est actuellement standardisée et de ce fait, peut plus facilement servir de test clinique, en particulier avec le développement des méthodes de détection en cytométrie de flux [209, 210].

2.4.2.3 Expression du marqueur CD38

Le CD38 est une glycoprotéine membranaire d'environ 45 kDa, qui témoigne de l'activation et de la maturation cellulaire. Elle joue, également, un rôle dans la signalisation cellulaire. Proposé comme marqueur de substitution à l'étude du statut mutationnel des gènes IGHV [179], son intérêt comme facteur de pronostic (du statut mutationnel des IGHV) a été démontré par certaines équipes. En effet, son expression par les cellules de LLC est associée à un mauvais pronostic et à une évolution rapide de la pathologie [137-140, 179]. L'expression de CD38 sur les lymphocytes tumoraux fut le premier marqueur à être corrélé au statut mutationnel des IGHV [178]. Cependant, depuis les travaux de Damle et al. (1999) [179], les résultats obtenus avec l'expression de CD38 et de ZAP70, validés comme marqueurs corrélés à l'état mutationnel des gènes IGHV et comme indicateurs de pronostic, restent incertains et controversés. En effet, il est apparu que le niveau d'expression du CD38 peut varier au cours de l'évolution de la LLC, en particulier après la prise d'un traitement par le patient [187, 211]. Environ 10 à 30% des cas montrent des résultats discordants entre l'expression du CD38 et de ZAP70 et l'état mutationnel des gènes IGHV [187]. D'autre part, il existe également des controverses parmi les différentes études pour définir la valeur seuil de ce marqueur (pourcentage d'expression de CD38). Ce seuil est définit à 30% [139, 179] et voir même à

moins de 20% [212] afin de discriminer au mieux les patients à bon pronostic de ceux à mauvais pronostic.

2.4.2.4 Anomalies cytogénétiques

Les aberrations génomiques et chromosomiques [213] sont d'autres facteurs génétiques connus pour avoir une signification pathogénique et clinique dans la LLC. L'analyse d'hybridation fluorescente in situ (FISH), réalisée sur des noyaux cellulaires en interphase, permet de retrouver des anomalies génomiques dans plus de 80% des cas de LLC. Döhner et al. [213] ont établi un modèle pronostique hiérarchique permettant d'identifier cinq groupes. Ces anomalies améliorent notre compréhension de la pathogenèse de la maladie. Les anomalies chromosomiques les plus fréquentes observées par ordre croissant de pronostic de survie sont:

(a) La délétion du bras court du chromosome 17 (17p) responsable le plus souvent de la délétion de la bande 17p13 et induit une inactivation de p53, protéine impliquée dans la réparation de l'ADN et dans l'induction de l'apoptose (7% des cas) [213]. Elle est corrélée à une médiane de survie courte (32 mois).
(b) La délétion du chromosome 11q. Les délétions de 11q22-23 induisent une mutation du gène suppresseur de tumeur ATM (ataxia telangiectasia) qui peut entraîner un dysfonctionnement de la voie p53 dans la protection du génome en cas de lésions de l'ADN (18%). Elle est corrélée à une médiane de survie de 79 mois [149, 214, 215].
(c) La trisomie 12 n'affecte pas significativement la survie, comparée à des patients ayant un caryotype normal (114 mois contre 111 mois) (18%).
(d) Enfin, la présence d'une délétion 13q, anomalie cytogénétique la plus fréquente, identifie un groupe de pronostic favorable avec une survie médiane de 133 mois (55%) [213].

Ces anomalies définissent des sous-groupes de patients qui diffèrent pour la vitesse d'évolution de la maladie. Ainsi les aberrations chromosomiques défavorables (11q, 17p) surviennent plus fréquemment chez les patients ayant des gènes IGHV non mutés, et les anomalies favorables (13q isolée, ou l'absence d'anomalies) plus souvent dans le groupe des gènes IGHV mutés. Toutefois, environ 2/3 des LLC avec des gènes IGHV non mutés n'ont pas d'anomalies cytogénétiques défavorables, ce qui indique une influence différentielle de ces facteurs pronostiques [213, 216, 217].

2.5 Évolution de la maladie

L'évolution de la LLC est très variable. Un certain nombre de patients est atteint d'une maladie indolente reliée à une courbe de survie très longue pouvant durer plusieurs décades. Tandis que d'autres sont atteints de formes de la maladie très agressives et subissent une aggravation rapide. Outre la progression de la masse tumorale, l'évolution peut être émaillée par plusieurs types de complications. Les complications les plus fréquemment rencontrées sont les cytopénies, les infections, les transformations et les cancers associés.

2.5.1 Cytopénies

Il s'agit avant tout de l'anémie (diminution de globules rouges) et de la thrombopénie (diminution de plaquettes sanguines). La neutropénie (diminution du nombre de neutrophiles dans le sang) est moins fréquente. Les cytopénies sont principalement dues soit à une insuffisance médullaire, à un hypersplénisme (diminution, dans le sang circulant, du nombre de globules rouges, de globules blancs (granulopénie), et de plaquettes) ou encore d'origine auto-immune. L'anémie est soit d'origine centrale, liée à l'infiltration cellulaire de la moelle osseuse, soit d'origine périphérique: anémie hémolytique auto-immune. La thrombopénie peut résulter, comme l'anémie, soit d'une baisse de production par infiltration de la moelle, soit d'une destruction périphérique des plaquettes. Les neutropénies d'origine auto-immune, rarement observées, se voient essentiellement dans les LLC T.

2.5.2 Complications infectieuses

Les infections sont les complications les plus fréquentes et les premières causes de morbidité et de mortalité des patients atteints de LLC. Elles surviennent dans 50% des cas, particulièrement dans les stades avancés de la maladie. La sensibilité aux infections résulte d'une hypogammaglobulinémie, qui est un déficit immunitaire humoral mais aussi cellulaire, ou encore de la neutropénie. Ces infections sont essentiellement d'origine bactériennes, affectant l'appareil respiratoire (pneumococcies, tuberculose), la peau et l'appareil urinaire: les septicémies sont assez fréquentes et d'autres organes peuvent être touchés. Les infections virales les plus fréquentes sont dues aux virus de l'herpès: Herpes simplex ou zona. Les infections mycobactériennes ou fongiques sont moins fréquentes.

2.5.3 Transformation en lymphomes

Une évolution des formes agressives, en particulier le développement d'un lymphome (syndrome de Richter), la transformation en leucémie prolymphocytaire ou plus rarement en leucémie aiguë ou en myélome multiple, ont été décrites. Le syndrome de Richter survient dans 3 à 10 % des LLC. Il doit être évoqué cliniquement avec l'apparition de masses ganglionnaires tumorales (qui peuvent entraîner des douleurs et des signes de compression). Ces aspects tumoraux sont révélés au scanner. Suspecté par la ponction ganglionnaire, le diagnostic sera confirmé par la biopsie. Le type histologique est variable mais il s'agit le plus souvent de lymphomes à grandes cellules parfois de type immunoblastique. La réponse aux différentes chimiothérapies des patients atteints du syndrome de Richter, est généralement faible et la médiane de survie ne dépasse pas quatre mois.

2.5.4 Cancers

Le risque de cancer chez les sujets porteurs de LLC est plus important que chez les sujets témoins du même âge et du même sexe. L'incidence de ces cancers est de 6 à 15%. Le cancer de la peau était considéré comme le plus fréquent mais d'autres types de cancers ont été retrouvés par d'autres équipes, en particulier le cancer du poumon et du côlon. L'incidence de l'apparition de cancers pourrait être augmentée par l'utilisation du Chlorambucil (traitement utilisé dans certaines formes de LLC) chez les patients par rapport au groupe de patients non traités.

2.6 Traitement

La stratégie thérapeutique traditionnellement adoptée pendant des années reposait sur le principe: *'primum non nocere*' ou *'d'abord ne pas nuire*', réflexion éthique du médecin vis à vis du principe de précaution entre la maladie et le risque du traitement. Cette approche était basée sur le fait que les patients appartenaient à une population relativement âgée qui devait à priori mourir d'une autre cause que la LLC. En outre, les traitements généralement utilisés pour traiter la LLC restaient inefficaces pour prolonger la vie du patient et s'accompagnaient très rarement de rémissions complètes [136]. Le développement de nouveaux facteurs de pronostics [6, 178, 179] a permis une bien meilleure classification des patients dans des catégories de risques différents. De plus il est devenu évident au fil des années que les patients qui ont une progression de la LLC mourront des complications de cette maladie, spécialement d'infections. Car, au fur et à mesure que la maladie progresse, le

dysfonctionnement immunitaire et la myélosuppression (diminution de l'activité des cellules de la moelle osseuse) deviennent de plus en plus sévères. La LLC étant reconnue comme une maladie d'évolution très hétérogène, une meilleure caractérisation de l'affection, tant sur le plan diagnostique que pronostique, ainsi que la découverte d'agents thérapeutiques nouveaux, ont permis de diversifier les stratégies de traitement de la LLC. De récentes améliorations des traitements thérapeutiques plus efficaces et moins toxiques ont permis d'atteindre de hauts pourcentages de rémission complète (RC) [136] pour 60 à 70% des patients après traitement. Néanmoins, une rechute est attendue pour la plupart de ces patients. Malgré les avancées dans la prise en charge de la LLC, aucun traitement permettant une guérison totale n'a encore été découvert. La détermination d'un tel traitement reste encore de nos jours un des buts à atteindre pour la médecine moderne.

Etablir le diagnostic de la LLC n'implique pas la mise en place d'un traitement, la décision de traiter ou de ne pas traiter la leucémie lymphoïde chronique dépendant d'un certain nombre de critères, tels que le stade évolutif de la maladie, l'âge du patient, l'existence ou non de facteurs de mauvais pronostics.

En considérant l'hétérogénéité de l'évolution de la maladie, la principale question que se posent les hématologistes est de déterminer le moment pour traiter le patient. Il existe à l'heure actuelle un consensus, une ligne de conduite définie par le NCI-WG [136], pour déterminer le moment approprié de l'initiation de la thérapie durant l'évolution de la LLC:

(1) Les patients en stade Binet A et Rai 0, soit 60% des patients, ne sont pas actuellement traités en première intention. Il faudrait disposer d'un traitement efficace, ayant peu d'effets secondaires et peu coûteux pour changer cette attitude. Néanmoins, la possibilité de mieux caractériser la maladie en accord avec les nouveaux marqueurs de pronostic (tel que le statut mutationnel des IGHV [6, 178, 179]) peut inciter les médecins à tester les avantages d'un traitement pour les patients présentant un mauvais pronostic.
(2) Les patients en stade Binet B ou Rai I et II devront être traités, quand ils présenteront des preuves de la progression de la maladie.
(3) En revanche, les patients en stade Binet C et Rai IV sont traités en première intention.

Le traitement standard a été, pendant des dizaines d'années, basé sur des agents alkylants pris quotidiennement ou par intermittence seuls ou associés avec des corticostéroïdes. La Chlorambucil seule peut fournir jusqu'à 70% de réponse, mais il existe seulement 10% de

rémission complète. De nos jours, la Chlorambucil est seulement indiquée comme traitement palliatif.

Actuellement, et ce depuis le milieu des années 80, les patients sont traités classiquement par des analogues de purines: la Cladribine, la Pentostatin et en particulier la Fludarabine, qui a été l'agent le plus étudié et le plus utilisé. Ils inhibent la ADN polymérase et la ribonucléotide réductase, favorisant l'apoptose, ce qui permet d'obtenir jusqu'à 80% de réponse, avec 38% de RC [218]. La réponse à ce traitement est supérieure, en terme de RC et de durée de rémission, à celle des agents alkylants seuls, tels que la Chlorambucil. Mais aucune différence en terme de survie globale n'a été démontrée [219, 220]. La Fludarabine est souvent associée à un agent alkylant la cyclophosphamide [221], en raison de la synergie dans la LLC entre ces deux molécules [222] et donne de meilleurs résultats en terme de survie [222-224].

D'autres lignes thérapeutiques utilisent les chimiothérapies associées à des anticorps monoclonaux, tel que le Rituximab [225, 226]. Ce dernier est un anticorps monoclonal chimérique anti-CD20 utilisé avec succès pour les patients atteints de lymphomes B. En monothérapie, il ne permet d'obtenir que des réponses partielles, et il est le plus souvent associé à la Fludarabine et au cyclophosphamide (FCR). L'alemtuzumab est lui aussi un anticorps monoclonal dirigé contre l'antigène CD52 qui est exprimé sur les cellules T et B normales et leucémiques, les macrophages et les monocytes. Particulièrement efficace, il a malheureusement une toxicité importante qui limite son utilisation. Sa toxicité est diminuée lorsqu'il est administré par voie sous-cutanée.

De nos jours, la médecine moderne permet également de traiter les patients par des greffes de cellules souches hématopoïétiques (HSC). Cette thérapie donne un taux important de rémission complète. Malheureusement, ce nouveau traitement ne permet toujours pas d'atteindre une guérison définitive [227, 228]. De plus, l'utilisation de ces greffes est limitée par l'âge avancé des patients et le taux de mortalité dû à la transplantation (20 et 41%) [229]. Néanmoins, la greffe de cellules souches hématopoïétiques est la seule stratégie qui permette, à l'heure actuelle, une guérison potentielle de la maladie [230-233]. Le rôle précis des allogreffes n'est pas encore très bien défini quant à sa place dans la stratégie thérapeutique et évoluera probablement avec la définition de nouveaux groupes pronostiques de patients. Néanmoins, un consensus sur les indications de l'allogreffe a été établie et publié par le groupe européen de l'EBMT (European Group for Blood and Marrow Transplantation) [234].

Malgré les avancées thérapeutiques majeures, la LLC est une pathologie incurable. Seule la greffe allogénique a permis d'obtenir des rémissions prolongées, aucun des traitements

actuels n’est curateur, et la maladie évoluant habituellement en phases successives nécessite plusieurs types de traitements.

CHAPITRE 3

IMGT/V-QUEST

Les intérêts croisés de la recherche dans le domaine de la santé, de la recherche fondamentale en immunologie et dans les applications technologiques (ingénierie des anticorps…), ont conduit IMGT® à la création d'un outil web spécialisé, IMGT/V-QUEST ('V-QUEry and STandardization') [235]. Cet outil permet d'analyser des séquences nucléotidiques réarrangées en prenant en compte la structure particulière des domaines V des récepteurs d'antigènes (IG et TR). IMGT/V-QUEST est accessible sur le Web depuis Juillet 1997. IMGT/V-QUEST identifie les gènes et allèles V, D et J dans les séquences réarrangées V-J et V-D-J par alignement avec les gènes et allèles germline d'IG et TR des répertoires de référence d'IMGT. Il délimite les régions charpentes (FR-IMGT) et les régions hypervariables (CDR-IMGT) de la séquence soumise par l'utilisateur en accord avec les règles de la charte scientifique d'IMGT®, basée sur les axiomes et les concepts de description, de classification et de numérotation de l'IMGT-ONTOLOGY [11, 13].

Lors de mon projet de thèse, de mise en œuvre d'un système d'information dédié à l'analyse et la gestion des séquences réarrangées d'IG et TR, il a été nécessaire de réécrire le coeur du programme dans le but d'homogénéiser l'outil et de faciliter ainsi son intégration à de nouvelles applications. Nous avons fait évoluer le logiciel pour répondre aux nouveaux besoins (notamment pour la recherche médicale), avec de nouvelles fonctionnalités telles que la localisation et la caractérisation des mutations, la détermination des insertions et des délétions, une évaluation de la fonctionnalité, le tout associé à une nouvelle interface web que nous avons rendu paramétrable [236] (publication 1).

3.2 Principes de la recherche par IMGT/V-QUEST

Pour identifier les gènes V, D et J impliqués dans le réarrangement, IMGT/V-QUEST cherche les régions constitutives des séquences d'IG et TR en comparant la séquence utilisateur avec les séquences de référence des gènes et allèles d'IG et TR présents dans IMGT/GENE-DB[1] [237]. Les séquences de référence proviennent de données expérimentales, annotées selon les règles standardisées d'IMGT®. Les répertoires des séquences de référence, utilisés par IMGT/V-QUEST, contiennent des séquences correspondant à: la V-REGION pour les gènes et allèles V, la J-REGION pour les gènes et allèles J et la D-REGION pour les gènes et allèles D. IMGT/V-QUEST procède en une analyse séquentielle d'une séquence réarrangée d'IG ou de TR soumise par l'utilisateur, afin de déterminer les gènes et allèles de cette séquence et de délimiter les nucléotides appartenant à chaque région. L'ordre de recherche est le suivant: V-REGION, J-REGION, D-REGION. Lorsqu'une région est délimitée et lorsque les gènes et allèles germline (les plus proches de la séquence utilisateur) sont identifiés, l'algorithme passe à la recherche de la région suivante, en excluant les régions précédemment déterminées. L'analyse précise et complète de la jonction est effectuée par un outil dédié et intégré à IMGT/V-QUEST: IMGT/JunctionAnalysis [238].

La méthode d'alignement employée par l'outil IMGT/V-QUEST doit prendre en compte les règles du IMGT unique numbering [14]. Dans la numérotation IMGT, les acides aminés conservés de la V-REGION (et les codons correspondants) sont toujours localisés aux mêmes positions (cystéine 23, tryptophane 41 et cystéine 104) et des gaps sont introduits dans les CDR pour uniformiser les tailles variables des CDR-IMGT. Pour appliquer et préserver la numérotation, les alignements réalisés par IMGT/V-QUEST ne doivent pas autoriser l'insertion de gaps supplémentaires. Afin de respecter ces contraintes, nous avons mis en place un algorithme d'alignement par paire (alignement de séquences deux à deux) dérivé d'un algorithme d'alignement classique de type global qui n'autorise ni les insertions ni les délétions. Le résultat de ces alignements permet 1) de délimiter chaque région, 2) dans le cas des gènes variables, de déterminer la séquence V germline numérotée selon les règles du IMGT unique numbering [14] qui sert de modèle pour numéroter la V-REGION de la séquence utilisateur. L'outil peut ensuite comparer la région délimitée, V-REGION (numérotée), D-REGION, ou J-REGION, avec l'ensemble des V-REGION, D-REGION ou J-

[1] IMGT/GENE-DB est la base de données de gènes de IMGT® dans laquelle sont répertoriés et classés tous les gènes et allèles des IG et des TR connus de l'homme et de la souris (l'intégration de gènes d'autres espèces est en cours).

REGION respectivement des gènes et allèles germlines afin de déterminer les gènes et allèles impliqués dans le réarrangement de la séquence utilisateur.

3.3 Principes d'alignement global sans insertions ni délétions

Les programmes de comparaison de séquences ont pour but de révéler en quoi ces séquences se ressemblent ou diffèrent en recherchant des régions similaires, afin de discriminer celles qui ont une relation biologique, de celles qui n'en ont pas. En général les algorithmes fonctionnent sur des alignements de segments de séquences (nommés fenêtres, motifs ou mots), qui sont comparés pour déterminer s'il existe ou non une similitude significative. L'évaluation de la similitude entre deux segments se calcule par la somme des scores élémentaires entre deux résidus (nucléotides ou acides aminés). Nous distinguons l'identité entre deux résidus et la ressemblance entre deux résidus non identiques qualifiée de similitude. Il existe bien évidemment plusieurs façons d'évaluer la similitude entre deux résidus, elles sont codées dans les différentes matrices de substitution (par exemple: la matrice NUC4.4 pour les nucléotides, les matrices PAM et BLOSUM pour les acides aminés). Il s'agit de repérer les segments pour lesquels la ressemblance est significative. En fait, un alignement de deux segments est considéré significatif lorsque son score est supérieur ou égal à un score seuil préalablement fixé.

L'algorithme d'alignement par paire employé dans IMGT/V-QUEST (Figure 3.1) calcule le score des alignements possibles entre deux séquences. Dans le but d'optimiser le temps de calcul, l'algorithme n'explore pas l'ensemble des alignements possibles. Par exemple, pour une V-REGION, il considère uniquement les alignements qui comparent deux séquences sur une longueur de plus d'un tiers de la plus petite séquence. Les alignements qui comparent deux séquences sur une longueur de moins d'un tiers de la plus petite séquence ne sont pas considérés, la longueur d'alignement étant jugée trop faible pour déduire une relation entre une séquence réarrangée et la séquence d'un gène germline. Cette longueur minimale est appelée overlap. Les valeurs de l'overlap ont été fixées au cours du développement de la première version d'IMGT/V-QUEST afin de limiter le nombre d'itérations lors de la recherche du meilleur alignement, tout en minimisant les risques de ne pas obtenir le meilleur alignement. Elles sont récapitulées en annexe 4 pour chaque type d'alignement réalisé par IMGT/V-QUEST.

L'algorithme de recherche de la V-REGION dans la séquence utilisateur est détaillé ci-dessous:

Pour deux séquences $X = x_1x_2\ldots x_n$ (séquence utilisateur) et $Y = y_1y_2\ldots y_m$ (séquence de référence) à comparer, les positions de départ de la recherche SP (StartPosition) et de fin de la recherche EP (EndPosition) sont déterminées. La position de départ SP correspond à la position d'un nucléotide (dans le sens 5' -> 3') de la séquence X (utilisateur) qui devra être aligné avec le premier nucléotide de la séquence Y (de référence). La position de fin EP correspond à la position d'un nucléotide (dans le sens 5' -> 3') de la séquence utilisateur qui devra être aligné avec le dernier nucléotide de la séquence Y (de référence). Pour l'alignement de deux V-REGION, nous obtiendrons:

SP = Taille de la séquence utilisateur – partie entière du 1/3 de la taille de la plus petite séquence (des 2 séquences comparées)

EP = partie entière du 1/3 de la taille de la plus petite séquence +1

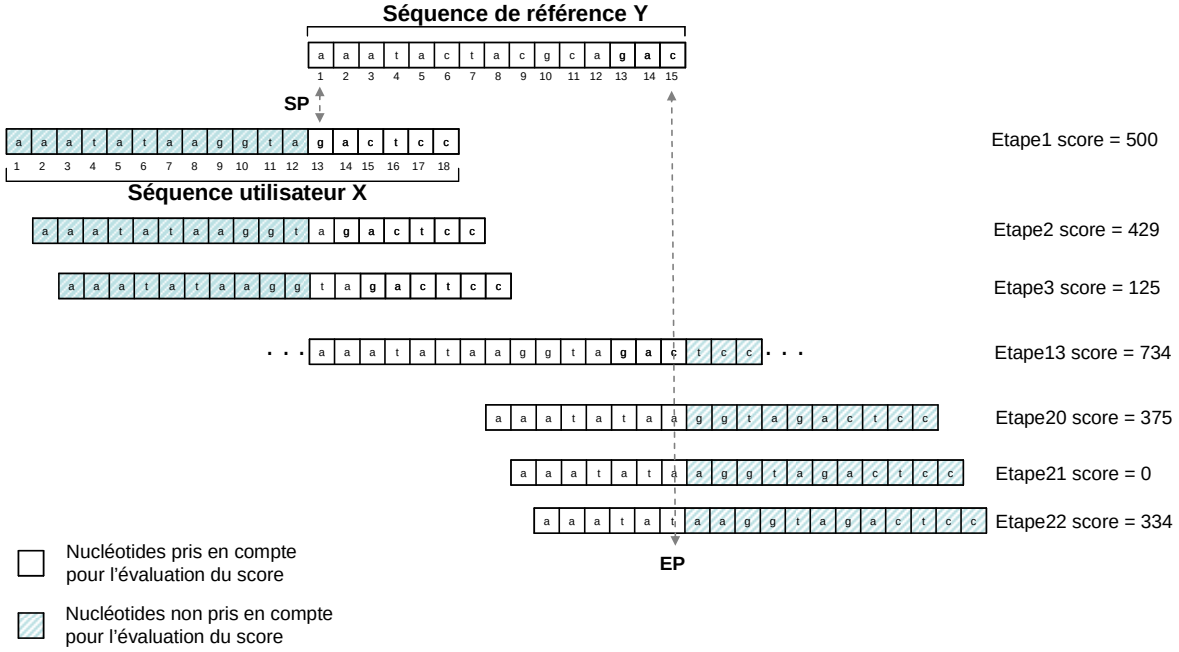


Figure 3.1: Méthode d'alignement entre la séquence utilisateur et une séquence de référence d'IMGT et évaluation du score normalisé. Exemple avec une séquence utilisateur (X) de longueur Tsu = 18 nucléotides, et une séquence de référence de longueur Tsr = 15 nucléotides; SP=18-(15/3)=13, EP=(15/3)+1=6. Le meilleur score est déterminé à l'étape 13.

Le score S d'un alignement est obtenu en soustrayant à 1, la somme des scores élémentaires S(X,Y). Le score S(X,Y) doit être indépendant de la tailles des segments comparés (il est donc divisé par 2 fois la longueur des segments comparés). Nous obtenons alors un score normalisé.

$$S = (1 - (\frac{S(X,Y)}{2 \times LC})) \times 1000$$

LC est la longueur des segments comparés, S(X,Y) est la somme des scores élémentaires en chacune des positions des séquences comparées.

$$S(X,Y) = \sum_{i=1}^{LC} s(x_i, y_j)$$

$s(x_i, y_j)$ est le score élémentaire de comparaison entre deux nucléotides x et y alignés, indiqué dans la matrice de substitution (Annexe 2).

Pour évaluer le score des alignements suivants, la séquence utilisateur est décalée d'une position dans le sens 3'. La recherche du meilleur alignement prend fin quand le nucléotide en position EP de la séquence utilisateur est aligné avec le dernier nucléotide de la séquence de référence. Pour chaque alignement un score est obtenu et seuls les alignements dont le score est jugé significatif seront mémorisés. Un score significatif doit répondre à deux conditions: être supérieur ou égal à un seuil préfixé (le seuil pour la V-REGION étant fixé à 600) et être supérieur au meilleur score précédemment calculé. Le score le plus grand détermine le meilleur alignement. Toutes les valeurs des seuils utilisées pour chaque type alignement réalisé par IMGT/V-QUEST sont indiquées en annexe 4.
Si les séquences ont respectivement une longueur de m et n nucléotides, la complexité de l'alignement sera de l'ordre de n*m. Il s'agit de l'étape limitante du programme, il conviendra donc de choisir judicieusement le nombre et la qualité des séquences de référence qui devront être comparées avec la séquence utilisateur pour chaque type d'alignement pour limiter le temps de calcul.

3.4 Les différentes étapes de l'analyse

L'analyse séquentielle classique effectuée sur une séquence par IMGT/V-QUEST se divise en 8 étapes représentées dans la figure 3.2.

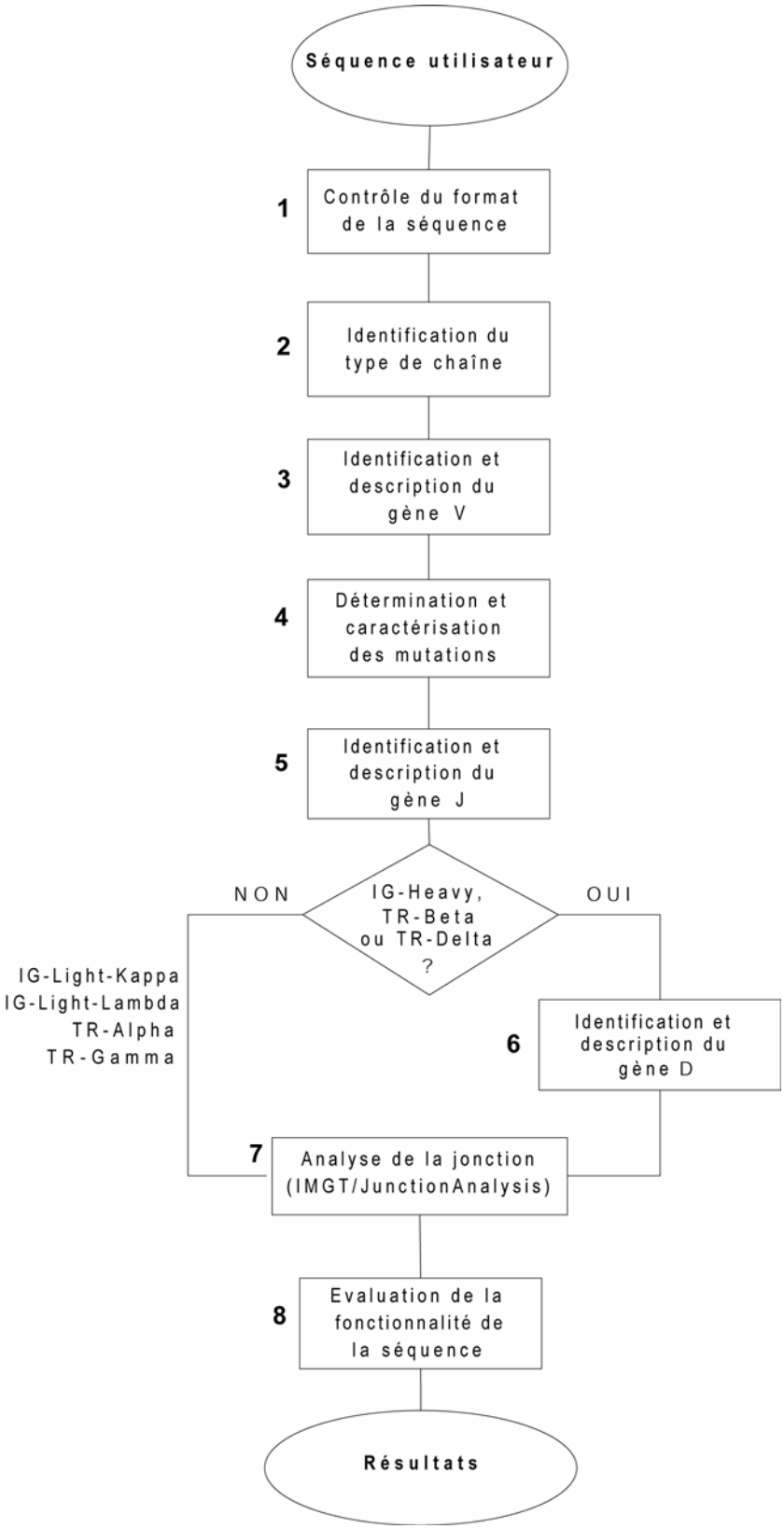


Figure 3.2: Diagramme de la procédure d'analyse d'une séquence réarrangée de récepteur d'antigènes.

3.4.1 Contrôle des séquences utilisateur

IMGT/V-QUEST accepte les séquences nucléotidiques dans le code du International Union of Biochemistry (IUB) [239] et est capable de gérer les séquences en orientation ‘sens’ et ‘anti-sens’ (inverse complémentaire).
Dans une première étape IMGT/V-QUEST vérifie l’intégrité de la séquence utilisateur ce qui inclut:

- La vérification du format de la séquence soumise par l’utilisateur. IMGT/V-QUEST accepte uniquement les séquences en format FASTA (http://www.imgt.org/textes/IMGTindex/Fasta.html).
- L’élimination dans la séquence nucléotidique des caractères qui ont pu être insérés lors du formatage de la séquence, soit: « '. ', '_', '/', '-', '1', '2', '3', '4', '5', '6', '7', '8', '9', '0', '\n', '\t', '\r'.
- L’arrêt de l’analyse si un caractère différent de 'a', 't', 'g', 'c', 'n', 'm', 'd', 'b', 'v', 'w', 'k', 'r', 's', 'h', 'y' est présent dans la séquence nucléotidique. Dans ce cas un message avertit l’utilisateur de la présence de caractères non valides.
- Le remplacement des caractères ‘x’ par le caractère ‘n’.
- L’élimination des nucléotides ‘n’ consécutifs aux extrémités 5’ et 3’ de la séquence (la présence de nucléotides non définis (‘n’) en 5’ et en 3’ peut perturber les alignements). Dans ce cas un message avertit l’utilisateur du nombre de nucléotides ‘n’ éliminés en 5’ et en 3’ dans les résultats.
- La vérification de la taille limite de la séquence utilisateur. Pour toute séquence de plus de 10.000 caractères, l’analyse s’arrête et un message indique que la séquence dépasse les capacités de l’outil.

3.4.2 Identification du type de chaîne

L'espèce, le type de récepteur (IG ou TR) et le type de chaîne de la séquence utilisateur sont indispensables à l'analyse des séquences réarrangées d’IG et TR: ces paramètres conditionnent le choix des répertoires de référence comparés à la séquence utilisateur. Les paramètres sélectionnés par l’utilisateur sont l’espèce et le type de récepteur. Le type de chaîne de la séquence utilisateur sera déterminé par IMGT/V-QUEST par alignement entre la séquence utilisateur et les séquences du répertoire de référence ‘IMGT Group set'. Le ‘IMGT Group set’ contient, pour une espèce et un type de récepteur donnés,

des séquences de référence représentatives de chaque groupe de gène variable (IGHV, IGKV, IGLV pour les IG, TRAV, TRBV, TRGV, TRDV pour les TR). Les séquences sélectionnées pour chaque groupe ont pour caractéristique de présenter toutes les longueurs des CDR-IMGT possibles. Le type de chaîne de la séquence utilisateur est alors celui de la séquence de référence donnant le meilleur score d'alignement.

3.4.3 Identification et description du gène V

Une fois le type de chaîne connu, l'outil peut identifier le gène et l'allèle V de la séquence utilisateur. La recherche du gène et allèle V est l'étape qui représente le plus grand challenge de l'algorithme. Elle suit une procédure spécifique qui a été implémentée pour prendre en compte les règles de description et de numérotation d'IMGT-ONTOLOGY [13]. Elle localise la V-REGION dans la séquence utilisateur par alignement, elle numérote les codons de la V-REGION selon les règles du IMGT unique numbering [14] (insertion des gaps IMGT) et la compare à l'ensemble des V-REGION des gènes et allèles V germline afin de déterminer le gène et allèle V le plus proche de la séquence utilisateur.
Le diagramme de la procédure d'identification et de description des gènes et allèles V est représenté dans la figure 3.3.

3.4.3.1 Délimitation de la V-REGION et détermination de la séquence modèle

Dans une première étape, IMGT/V-QUEST localise la V-REGION dans la séquence utilisateur par alignement, selon la méthode décrite dans le paragraphe 3.2, entre la séquence utilisateur et les séquences du répertoire de référence 'IMGT Subgroup set'.
Pour obtenir une délimitation correcte de la V-REGION, la séquence utilisateur est alignée avec le 'IMGT Subgroup set' qui contient des représentants de chaque sous-groupe de gènes variables (ensemble des gènes d'un même groupe et d'une même espèce dont les V-REGION présentent au moins 75% d'identité au niveau nucléotidique). Les séquences sélectionnées pour chaque sous-groupe doivent être représentatives de toutes les combinaisons de longueur des CDR-IMGT existantes au sein du sous-groupe.

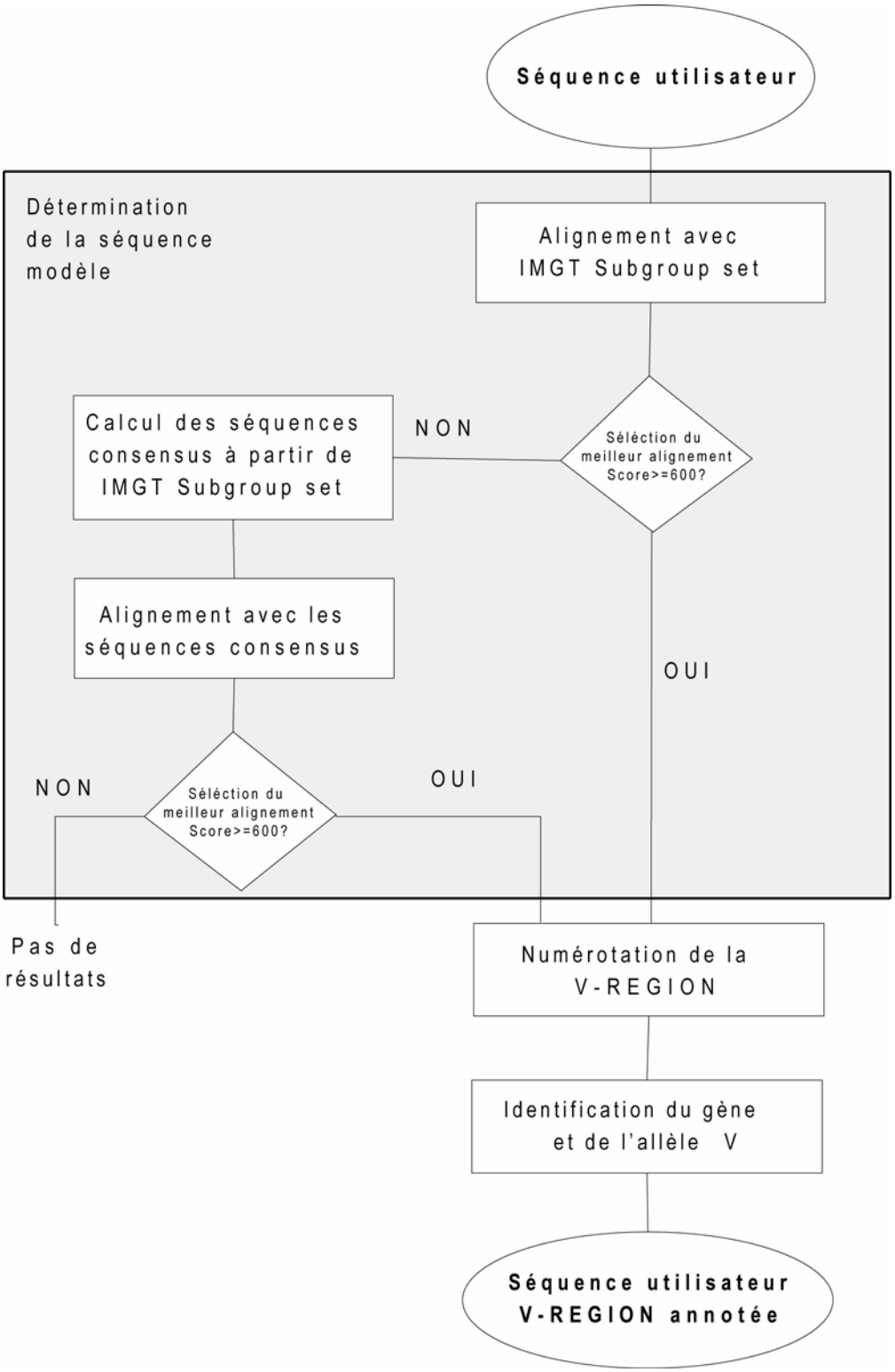


Figure 3.3: Diagramme de la procédure d'identification et de description du gène et allèle V.

Il faut cependant considérer deux cas de figure:

a) Le répertoire des gènes et allèles V germline est bien connu (comme par exemple chez l'homme) et la séquence utilisateur n'est pas excessivement mutée: elle est alignée avec les séquences du 'IMGT Subgroup set' dont les gaps IMGT ont été éliminés (mais mémorisés). Le meilleur alignement permet de délimiter la V-REGION dans la séquence utilisateur et permet de déterminer le nom de la séquence modèle qui servira à numéroter la V-REGION de la séquence utilisateur en accord avec la numérotation IMGT (Figure 3.4).

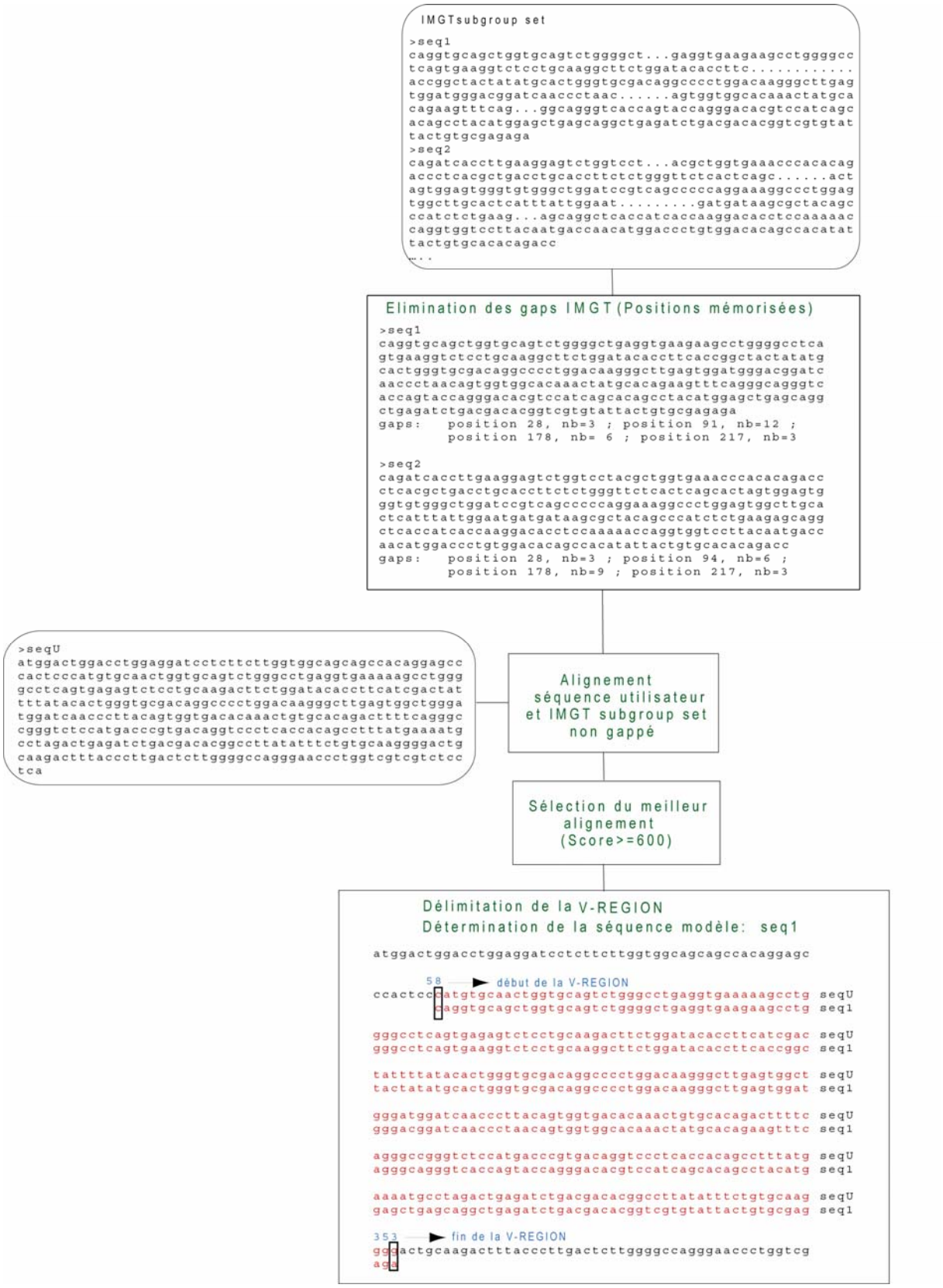


Figure 3.4: Diagramme de la procédure de délimitation de la V-REGION et de la détermination de la séquence modèle.

b) Le répertoire des gènes et allèles V germline n'est pas entièrement connu et/ou la séquence utilisateur est excessivement mutée: comme la suite des nucléotides influence fortement l'alignement, cela peut perturber la détermination de la séquence modèle. La séquence utilisateur sera alors alignée avec un ensemble de séquences consensus établi à partir du 'IMGT Subgroup set'. L'utilisation de séquences consensus permet de focaliser l'alignement sur les particularités de l'organisation structurelle des séquences (positions des nucléotides conservés et longueur des CDR-IMGT) et non pas sur l'enchaînement des nucléotides qui composent la séquence. Par conséquent, dans le cas où il y a peu de séquences de référence pour une espèce donnée, l'outil est capable de déterminer une séquence modèle dont l'organisation structurelle est semblable pour numéroter la séquence utilisateur. La détermination de la séquence modèle correspond à la démarche suivante:

- Dans une première étape IMGT/V-QUEST déduit la séquence consensus de l'ensemble des séquences du 'IMGT Subgroup set' comme le décrit la figure 3.5A. L'alphabet dégénéré de l'ADN (Annexe 1) utilisé est constitué de 15 lettres et d'un point représentant les gaps, soit ('.','g','a','r','c','s','m','v','t','k','w','d','y','b','h','n').
- Dans une seconde étape, chaque séquence de référence est comparée à la séquence consensus: le but est d'obtenir pour chaque comparaison, une nouvelle séquence consensus dont les CDR-IMGT sont de même longueur que ceux de la séquence de référence. Tous les nucléotides ou gaps de la séquence consensus qui sont localisés aux mêmes positions que les gaps IMGT dans la séquence de référence sont éliminés (Figure 3.5B). Le nombre et la position des gaps sont mémorisés par le programme pour être utilisés lors de l'étape de numérotation de la séquence utilisateur. Un lot de séquences contenant autant de séquences consensus modifiées qu'il y a de séquences de référence dans le répertoire 'IMGT Subgroup set' utilisé est alors obtenu.
- Dans une troisième étape IMGT/V-QUEST élimine les séquences consensus modifiées redondantes (Figure 3.5C).
- IMGT/V-QUEST aligne alors chaque séquence du lot de séquences consensus modifiées avec la séquence de l'utilisateur. L'alignement donnant le meilleur score définit la séquence consensus modifiée qui sera utilisée comme modèle

pour numéroter la séquence utilisateur. De plus, le meilleur alignement permet également de délimiter la V-REGION dans la séquence utilisateur.

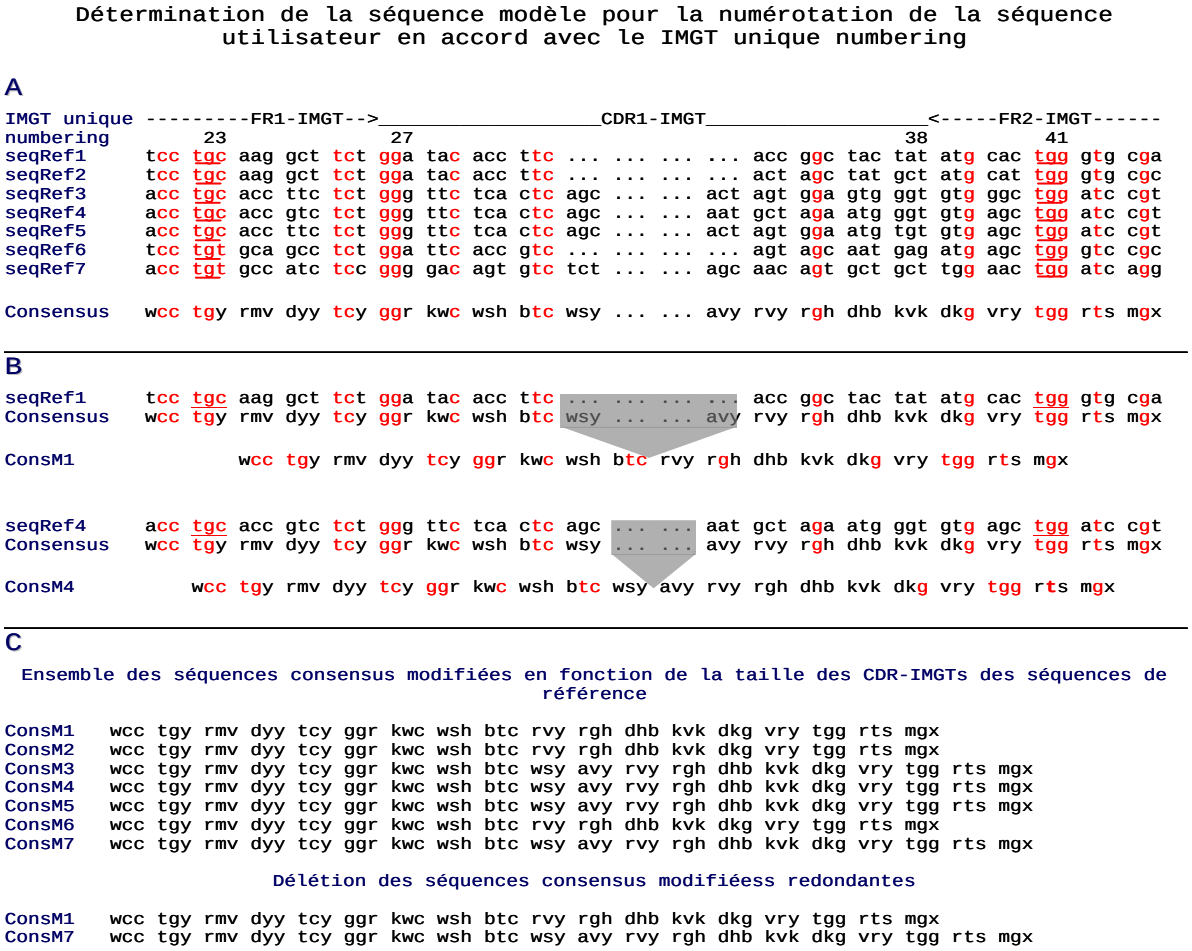


Figure 3.5: Exemple de la création du lot de séquences consensus. Les séquences indiquées sont des segments de V-REGION (FR1-IMGT, CDR1-IMGT, FR2-IMGT) de différents gènes et allèles de la position 22 à 43 selon la numérotation IMGT extrait pour l'illustration. Les nucléotides soulignés correspondent aux AA conservés. **A**. Création de la séquence consensus. Les séquences de référence de cet exemple correspondent à des segments de V-REGION des gènes et allèles suivant: seqRef1 = IGHV1-2*01; seqRef2 = IGHV1-3*01; seqRef3 = IGHV2-5*01; seqRef4 = IGHV2-26*01; seqRef5 = IGHV2-70*01; seqRef6 = IGHV3-d*01; seqRef7 = IGHV6-1*01. **B**. Création de nouvelles séquences consensus ConsM1 et ConsM4 dont la longueur du CDR est identique à celle des séquences de référence seqRef1 et seqRef4 respectivement. Tous les nucléotides ou gaps de la séquence consensus qui sont localisés aux mêmes positions que les gaps IMGT dans la séquence de référence sont éliminés. Un lot de séquence contenant autant de séquence consensus modifiée que de séquences de référence est obtenu. **C**. Elimination des séquences redondantes dans le lot de séquences consensus modifiées précédemment déterminé.

Dans la pratique, IMGT/V-QUEST recherche d'abord la séquence modèle selon la méthode décrite dans a). En absence de résultat, il recherche la séquence modèle en passant par l'intermédiaire des séquences consensus selon b).

A l'issue de cette étape, la V-REGION de la séquence utilisateur est délimitée: la position du début de l'alignement avec la séquence de référence ayant le score le plus élevé permet de déterminer le début de la V-REGION dans la séquence utilisateur. A partir de cette valeur,

nous en déduisons le nombre de nucléotides de la séquence en 5' de la V-REGION (cut-off), qui ne seront pas montrés dans les alignements présentés par IMGT/V-QUEST. Dans le cas de séquences partielles, les nucléotides manquants en 5' par rapport à la séquence modèle (leur nombre est aooelé offset), seront alors remplacés par des '.' dans les résultats affichés par IMGT/V-QUEST. La séquence modèle du meilleur alignement est par là même déduite.

3.4.3.2 Numérotation de la séquence utilisateur selon la numérotation unique IMGT

La séquence modèle déterminée, IMGT/V-QUEST insère dans la séquence utilisateur les gaps IMGT aux mêmes positions que les gaps de la séquence modèle dont les positions ont été mémorisées. La séquence utilisateur répond, désormais aux règles de la numérotation unique IMGT [14], de ce fait les FR-IMGT et CDR-IMGT sont délimités et les acides aminés conservés localisés.

3.4.3.3 Identification du gène et allèle V

Une fois la V-REGION de la séquence utilisateur numérotée selon les règles de la numérotation unique IMGT unique numbering [14], la V-REGION est simplement comparée nucléotide par nucléotide à l'ensemble des V-REGION des séquences du répertoire de référence IMGT. Cet ensemble contient pour un type de chaîne et une espèce donnée, les V-REGION de tous les gènes et allèles V fonctionnels, ainsi que les ORF (open reading frame) et pseudogènes (in-frame). Toutes ces séquences sont numérotées en accord avec les règles de la numérotation unique IMGT (IMGT unique numbering [14]). Le nom des gènes et allèles identifiés par IMGT/V-QUEST est conforme à la nomenclature IMGT, officiellement acceptée par le Human Genome Organisation (HUGO) Nomenclature Committee (HGNC) [240].

Chaque séquence de référence est comparée avec la séquence de l'utilisateur, nucléotide par nucléotide dans le sens 5'->3': de la position 1 à la position 104 pour les V-REGION des chaînes lourdes et de la position 1 à la position 109 et 110 respectivement pour les V-REGION des chaînes légères kappa et lambda. Pour chaque comparaison, un score est évalué (score qui sera affiché dans les résultats de l'outil), la plus grande valeur définissant le gène et allèle germline le plus proche de la séquence utilisateur.

$$Score = (5 \times nbL) - (4 \times nbD)$$

'nbL' étant le nombre de nucléotides identiques et 'nbD' le nombre de nucléotides différents entre les séquences comparées.

Le calcul du score ne prend pas en compte la fin de la V-REGION, étant donné que le CDR3-IMGT fait partie intégrante de la jonction, sujette à d'importantes modifications *via* les mécanismes de diversité des jonctions.
Le pourcentage d'identité est calculé dans les mêmes conditions (de la position 1 à 104 pour les V-REGION des chaînes lourdes, et de la position 1 à 109 et 110 respectivement pour les chaînes légères kappa (IGK) et lambda (IGL)). Si la séquence est partielle, le pourcentage d'identité est calculé uniquement au niveau des nucléotides alignés. Le pourcentage d'identité permet d'évaluer le statut mutationnel de la séquence utilisateur.
IMGT/V-QUEST ajuste la fin de la V-REGION en comparant les nucléotides de la séquence utilisateur avec les nucléotides de la séquence la plus proche dans le sens 3' -> 5'. Par défaut, dans IMGT/V-QUEST, la fin de la V-REGION en 3' est déterminée lorsque dans l'alignement nous obtenons deux nucléotides consécutifs identiques à la séquence germline. Notons ici que si le gène V n'est pas identifié, l'ensemble de cette étape est effectué sur la séquence en inverse complémentaire.

3.4.4 Détermination et caractérisation des mutations dans la V-REGION

La caractérisation des mutations dans la V-REGION de la séquence utilisateur se fait par comparaison avec la V-REGION du gène et allèle V germline identifié comme la séquence la plus proche. Les mutations sont caractérisées selon leur type: soit des transitions (modification d'une purine en purine, ou d'une pyrimidine en pyrimidine), soit des transversions (changement d'une purine en pyrimidine ou d'une pyrimidine en purine), et les mutations silencieuses ou non silencieuses. Dans le cas d'une mutation non silencieuse, (conduisant au remplacement d'un acide aminé par un autre), le remplacement est qualifié selon les classes d'acides aminés définies par IMGT: l'hydropathie, le volume et les caractéristiques physicochimiques [241]. IMGT/V-QUEST localise les positions des motifs où se produisent les mutations somatiques préférentielles (décrites dans la littérature) appelées les positions hot spots. Les motifs des hot spots (a/t)a, (a/g)g(c/t)(a/t), et leur motifs en inverse complémentaire (a/t)(a/g)c(c/t), t(a/t) sont localisés dans le gène et l'allèle V germline le plus proche.

3.4.5 Identification et description du gène J

Une fois la V-REGION délimitée, l'outil recherche la J-REGION.

3.4.5.1 Délimitation de la J-REGION

IMGT/V-QUEST recherche la J-REGION sur un segment d'une longueur de 200 nucléotides, à partir de la fin 3' de la V-REGION, ou jusqu'à la fin de la séquence utilisateur si celle-ci est plus petite. Cet intervalle de recherche permet d'identifier la J-REGION dans le cas de jonctions de très grande taille.

Pour cette recherche, le segment défini est aligné avec l'ensemble des J-REGION des gènes et des allèles J pour un type de chaîne et pour une espèce donnés. L'alignement avec le score le plus élevé permet de délimiter la J-REGION.

3.4.5.2 Identification du gène et allèle J

IMGT/V-QUEST compare ensuite chaque J-REGION des gènes et allèles J germline avec la J-REGION de la séquence utilisateur, nucléotide par nucléotide dans le sens 5' -> 3'. A chaque gène et allèle J comparé à la séquence utilisateur, un score est évalué comme dans le paragraphe 3.3.3.3 pour déterminer le gène et l'allèle J germline le plus proche de la séquence utilisateur. Le pourcentage d'identité est calculé entre la J-REGION délimitée et le gène et allèle J germline le plus proche. Pour éviter des résultats non significatifs, le nom des gènes et allèles J les plus proches ne sont pas fournis si moins de 6 nucléotides ont été alignés. Si le gène et allèle J n'a pas pu être identifié, la procédure de recherche de la J-REGION est de nouveau appliquée, en utilisant un nouveau intervalle de recherche, localisé entre quelques nucléotides en amont de la fin de la V-REGION et 200 nucléotides en aval. Cet intervalle de recherche a été défini empiriquement et est différent en fonction du type de chaîne de la séquence utilisateur et du type de récepteur IG ou TR. Pour les chaînes lourdes des IG, l'intervalle de recherche sera compris entre la position 103 selon la numérotation unique IMGT et 200 nucléotides en aval de cette position ou jusqu'à la fin de la séquence utilisateur. Pour les chaînes légères kappa (IGK) et lambda (IGL), cet intervalle est compris entre la position 107 selon la numérotation unique IMGT et 200 nucléotides en aval de cette position ou jusqu'à la fin de la séquence utilisateur. Pour les chaînes alpha des TR, l'intervalle est compris entre la position 103 selon le IMGT unique numbering et 200 nucléotides en 3' de cette position ou jusqu'à la fin de la séquence utilisateur. Cette procédure permet en particulier de localiser les J-REGION correspondant à des gènes J avec de nombreux nucléotides en 5' éliminés par l'activité exonucléase au cours du réarrangement.

Si l'identification du gène et allèle J est obtenue en utilisant un intervalle de recherche empiétant sur la fin de la V-REGION (intervalle précédemment décrit), la délimitation de la fin de la V-REGION de la séquence utilisateur reste par défaut la même. Et le début de la J-REGION est alors consécutif de la fin de la V-REGION.
La description (annotation) de la séquence est mise à jour: la V-J ou V-D-J-REGION est délimitée, ainsi que les acides aminés conservés de la J-REGION, soit le tryptophane 118 pour les IGH ou la phénylalanine 118 pour les autres types de chaînes. Ce qui permet de délimiter la JUNCTION de la cystéine 104 (2nd-CYS) au tryptophane ou phénylalanine (J-TRP ou J-PHE) 118 (C 104 -> F/W 118) et d'ajuster l'extrémité 3' du CDR3.

3.4.6 Identification et description du gène D

IMGT/V-QUEST recherche la D-REGION et le gène et allèle D le plus proche de la séquence utilisateur, uniquement pour les chaînes lourdes des IG (IGH) et les chaînes beta (TRB) et delta (TRD) des TR. IMGT/V-QUEST aligne (cf. paragraphe 3.2) la région de la séquence utilisateur située entre la fin de la V-REGION en 3' et le début de la J-REGION en 5' avec l'ensemble des D-REGION des gènes et allèles D du répertoire de référence IMGT (pour un type de chaîne et une espèce donnés).
Notons cependant que l'identification et la description d'un gène D réalisé par IMGT/JunctionAnalysis s'avèrent plus précises et ce sont les informations de IMGT/JunctionAnalysis qui sont affichées dans la page de résultats de IMGT/V-QUEST.

3.4.7 Analyse détaillée de la JUNCTION

Une fois la jonction délimitée et les gènes et allèles V et J déterminés, l'outil IMGT/JunctionAnalysis [238] intégré à IMGT/V-QUEST identifie précisément le gène et allèle D germline le plus proche de la séquence utilisateur. Cette recherche est faite pour les chaînes lourdes des IG (IGH), et les chaînes beta (TRB) et delta (TRD) des TR. Il délimite les régions palindromiques ou P-REGION et les N-REGION résultant de la N-diversité. De plus, IMGT/JunctionAnalysis [238] ajuste la fin de la V-REGION en 3' et le début de la J-REGION en 5'. L'annotation de la séquence est mise à jour en prenant en compte les résultats de l'analyse de IMGT/JunctionAnalysis.

3.4.8 Evaluation de la fonctionnalité

La dernière étape de l'analyse consiste à évaluer la fonctionnalité de la séquence utilisateur en accord avec les règles d'IMGT-ONTOLOGY [13]. Une séquence réarrangée des récepteurs d'antigènes est soit productive ou unproductive. Elle est définie comme productive si aucun codon stop et aucun changement de cadre de lecture n'ont été détectés dans la V-D-J-REGION ou V-J-REGION et si la jonction est 'in-frame' c'est-à-dire si les codons correspondant aux AA conservés J-TRP ou J-PHE 118 sont dans le même cadre de lecture que la 2nd-CYS 104. Elle est définie comme 'unproductive' si un ou plusieurs codons stops et un changement de cadre de lecture sont détectés dans la V-D-J-REGION ou V-J-REGION, et/ou si la jonction est 'out-of-frame', c'est-à-dire que les codons correspondant aux AA conservés J-TRP ou J-PHE ne sont pas dans le même cadre de lecture que la 2nd-CYS 104.

3.5 Recherche des insertions et des délétions

Il a été montré dans la littérature que des insertions et des délétions peuvent se produire respectivement dans 1,11% et 1,85% de cas dans la V-REGION de séquences réarrangées d'IG de cellules B normales ou malignes [242]. Dans l'analyse classique de IMGT/V-QUEST, les insertions et les délétions ne sont pas recherchées. Cependant, des indices permettent de suspecter ce phénomène, notamment un faible pourcentage d'identité, l'absence de la 2nd-CYS en position 104 ou bien encore des différences de longueur des CDR1-IMGT et/ou CDR2-IMGT entre la séquence utilisateur et le gène et allèle V germline le plus proche identifié.

Récemment, nous avons intégré une nouvelle fonctionnalité dans l'outil IMGT/V-QUEST qui permet de rechercher les insertions et délétions potentielles lors de l'analyse.

Dans le principe, la recherche de ces insertions et/ou des délétions dans la séquence utilisateur est basée sur un alignement de type local. Les alignements locaux ont pour but de trouver seulement les segments les plus similaires entre deux séquences en excluant les fragments plus distants. Pour évaluer les scores d'alignement avec des insertions et des délétions, il est nécessaire de les prendre en compte dans le calcul du score. L'approche la plus simple consiste à étendre la définition de nos matrices et de rajouter des coefficients pour l'alignement d'un nucléotide avec une 'brèche' ou gap. Nous utilisons généralement une valeur Δ de pénalité d'apparition de gap déterminée empiriquement pour obtenir les meilleurs résultats. Il est également possible d'appliquer une pénalité fixe pour toutes insertions, plus une pénalité pour étendre l'insertion. Cette dernière pénalité est moins lourde mais permet de

prendre en compte la longueur. L'expression de cette pénalité, à deux paramètres, peut être décrite par l'équation suivante:

$$P = \Delta + yl$$

P étant la pénalité pour une insertion de longueur l, Δ la pénalité fixe d'apparition d'une insertion indépendante de la longueur et y la pénalité d'extension pour un élément. En conséquence, une longue insertion est légèrement plus pénalisante qu'une courte, ce qui revient en fait à minimiser le poids de la longueur des insertions par rapport à l'introduction même d'une insertion.

Le temps de comparaison de deux séquences de longueur équivalente n est proportionnel à n^2. L'exploration de chaque position de chaque séquence pour la détermination d'une insertion ou d'une délétion augmente considérablement le nombre total de façons différentes d'aligner entre elles deux séquences. Ce nombre total de combinaisons est approximativement de 2^{2n}. Par conséquent, pour deux séquences de 20 nucléotides cela représente $2^{2*20}=2^{40}$ ce qui représente déjà plus de 137 milliards de combinaisons. Il est donc impensable d'employer une approche de type 'force brute' (qui consiste à tester tous les alignements), contrairement à l'alignement global sans insertions/délétions utilisé par IMGT/V-QUEST (cf. 3.2).

Nous avons utilisé, pour la détection des insertions/délétions, une méthode algorithmique très générale appelée programmation dynamique, qui permet de trouver l'alignement optimal en temps $O(n^2)$. Elle est basée sur le fait que tous les événements sont possibles et calculables mais que la plupart sont rejetés en considérant certains critères. La recherche des insertions et des délétions est donc basée sur l'algorithme dynamique d'alignement local appelé Smith et Waterman [243].

Lors de l'étape d'identification des insertions et délétions, la V-REGION de la séquence utilisateur, délimitée dans l'étape d'identification et description du gène et allèle V, est alignée aux V-REGION des 10 gènes et allèles V germline les plus proches. Les paramètres employés lors de l'alignement utilisé par IMGT/V-QUEST pour la détection des insertions et des délétions sont décrits dans la table 3.1.

Tableau 3.1: Paramètres d'alignement Smith&Waterman.

Paramètres	Valeurs
Match	-1
Remplacement	0
Insertion	30
Délétion	30
Extension de gap	1

Les différentes étapes sont représentées dans la figure 3.6.

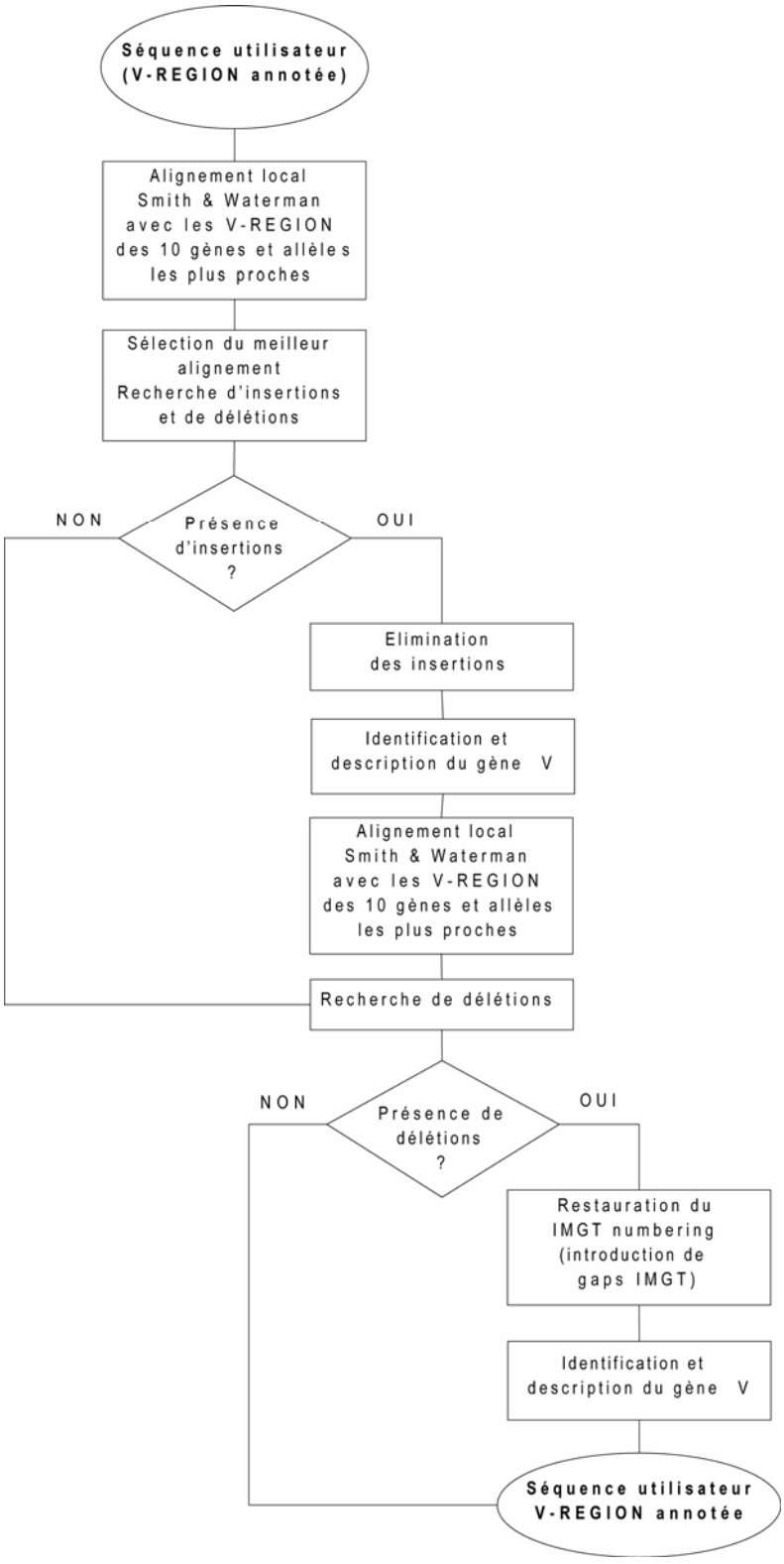


Figure 3.6: Diagramme de procédure de la détermination des insertions délétions dans la séquence utilisateur.

1. Si des insertions sont détectées dans l'alignement donnant le meilleur score, elles sont alors exclues de la séquence utilisateur pour restaurer la numérotation unique IMGT perturbé par les insertions. La position des insertions est mémorisée par le programme afin de les afficher au niveau des résultats et pour informer l'utilisateur. Après l'exclusion des insertions, l'outil applique de nouveau l'ensemble de la procédure de détermination du gène et allèle V le plus proche de la séquence utilisateur (cf 3.3.3). Nous réalisons une seconde fois l'alignement Smith et Waterman [243] pour rechercher les délétions potentielles.
2. Si des délétions sont détectées, des gaps sont insérés dans la séquence utilisateur pour restaurer la numérotation unique IMGT [14]. Et nous appliquons de nouveau l'ensemble de la procédure de détermination du gène et allèle V le plus proche de la séquence utilisateur sans la recherche des insertions et délétions.

En cas d'insertions ou de délétions, les alignements de Smith et Waterman [243] ne tiennent pas compte du cadre de lecture de la V-REGION. Voici une illustration d'une insertion de 3 nucléotides:

```
Séquence analysée (cadre de lecture 1)      ATCGTCGTC
Séquence de référence                       A---TCGTC
```

Lorsque les insertions et les délétions concernent un nombre de nucléotides multiple de 3, et si un ou les deux premiers nucléotides de l'insertion de la séquence analysée (dans notre exemple TC en vert) sont identiques aux un ou deux nucléotides de la séquence de référence suivant l'insertion (dans notre exemple TC en vert), IMGT/V-QUEST les localise de préférence au niveau de codons pleins. Ainsi, dans l'illustration choisie, suite à la modification effectuée par IMGT/V-QUEST, la localisation de l'insertion correspond au codon GTC:

```
Séquence analysée                           ATCGTCGTC
Séquence de référence                       ATC---GTC
```

Dans nos calculs de score, nous avons utilisé une pénalité fixe pour l'apparition d'une insertion ou d'une délétion, ajoutée à une pénalité d'extension de gaps, tandis que pour les substitutions, nous avons utilisé une matrice de substitution décrite en annexe 3. Les

paramètres utilisés dans les alignements Smith et Waterman [243] (Table 3.1) ont été obtenus de façon empirique dans le but de détecter au mieux les insertions et délétions potentielles.
Le réglage de ces paramètres a été obtenu par une série de tests effectués sur un lot de 26 séquences de patients atteints de LLC répertoriés par le European Research initiative on CLL (ERIC). Les séquences sont considérées comme problématiques car elles présentent des insertions et des délétions identifiées (Annexe 5). Lors de la détermination de ces paramètres, nous avons cherché à minimiser la possibilité d'obtenir des délétions en 3' de la V-REGION.

3.6 L'interface utilisateur IMGT/V-QUEST

IMGT/V-QUEST est accessible sur le Web (http://www.imgt.org), via une interface utilisateur, qui permet d'analyser les séquences réarrangées d'IG et TR et d'en visualiser les résultats.

3.6.1 IMGT/V-QUEST Search

IMGT/V-QUEST peut analyser jusqu'à 50 séquences simultanément [236]. Les séquences doivent être soumises en format FASTA, et peuvent être des séquences réarrangées des 3 locus (IGH, IGK et IGL) pour les IG et des 4 locus (TRA, TRB, TRG, TRD) pour les TR. Les séquences génomiques d'ADN (ADNg) ou d'ADN complémentaire (ADNc) sont analysées selon la même procédure (l'organisation d'une V-D-J-REGION et d'une V-J-REGION étant la même pour les 2 types de molécules). IMGT/V-QUEST analyse les séquences dans l'orientation 'sens', puis si aucun résultat n'est obtenu, l'analyse est appliqué dans l'orientation 'anti-sens'. Les résultats de l'analyse sont présentés dans l'orientation 'sens' des séquences.
L'utilisation de IMGT/V-QUEST se divise en deux étapes:

(i) L'utilisateur sélectionne l'espèce et le type de récepteur d'antigènes des séquences à analyser.
(ii) Il peut ensuite soumettre jusqu'à 50 séquences en format FASTA.

Avant chaque analyse, les utilisateurs ont la possibilité de personnaliser l'affichage des résultats (Figure 3.7). Ils peuvent obtenir les résultats sous un format texte et paramétrer le nombre de nucléotides pour chaque ligne de l'alignement. IMGT/V-QUEST offre deux formes de présentations différentes des résultats:

- Le «Detailed view» affiche les résultats des séquences individuellement.
- Le «Synthesis view» affiche les alignements des séquences assignées au même gène et allèle V.

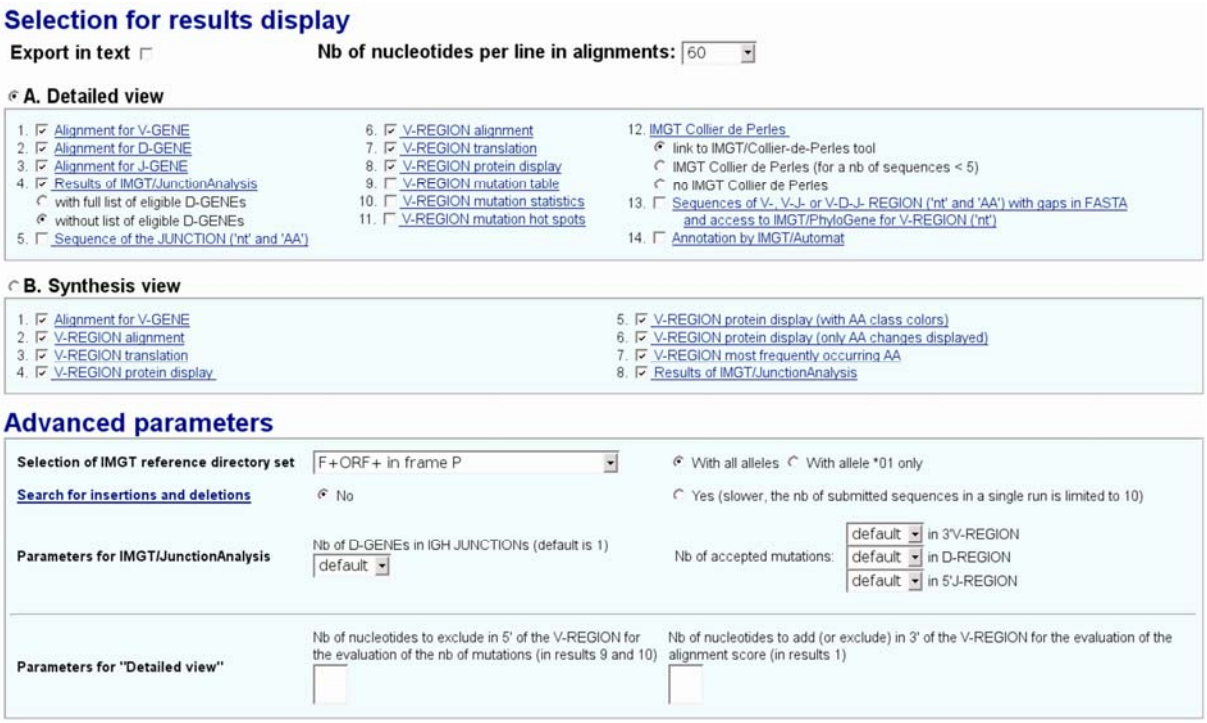


Figure 3.7: Paramètres de l'interface IMGT/V-QUEST. 'Selection for results display' permet à l'utilisateur de sélectionner le type d'affichage des résultats et les options d'analyse pour chacune d'elle (14 pour 'Detailed view' et 8 pour 'Synthesis view'). Les paramètres avancés permettent aux utilisateurs de modifier les valeurs par défaut.

IMGT/V-QUEST offre la possibilité de paramétrer les analyses des séquences d'IG et TR selon les besoins de chaque utilisateur dans 'Advanced parameters'.

i. 'Selection of IMGT reference directory set' permet de sélectionner les répertoires de références utilisés pour l'identification des gènes et allèles V, D et J et les alignements ('F + ORF', 'F + ORF + in-frame P', 'F + ORF including orphons' and 'F + ORF + in-frame P including orphons', où F signifie fonctionnel, ORF signifie cadre de lecture ouvert et P pseudogène). Le choix du répertoire de référence permet à l'utilisateur de travailler uniquement avec les séquences des gènes appropriées (par exemple les séquences orphons sont appropriées pour l'analyse de séquences génomiques mais ne le sont pas pour l'étude de l'expression des répertoires). Les répertoires de référence peuvent être également utilisés soit avec tous les allèles ('With all alleles'), soit uniquement avec l'allèle *01 ('With allele *01 only').

ii. ‘Search for insertions and deletions’ permet de rechercher les insertions et délétions dans la séquence utilisateur. Dans ce cas, le nombre de séquences soumises est limité à 10, le temps de l’analyse étant sensiblement plus long.

iii. ‘Parameters for IMGT/JunctionAnalysis’ permet de déterminer le nombre de D-GENE à rechercher dans une jonction IGH, TRB et TRD (par défaut 1 pour le locus IGH, 1 pour le locus TRB et 3 pour le locus TRD) et le nombre de mutations acceptées dans la partie 3’ de la V-REGION (3’V-REGION), la D-REGION et dans la partie 5’ de la J-REGION (5’J-REGION) (par défaut le nombre de mutations accepté pour le locus IGH est de 2 pour la 3’V-REGION et la 5’J-REGION et de 4 pour la D-REGION, pour les locus IGK et IGL, il est de 7 pour la 3’V-REGION et 5’J-REGION). Les valeurs par défaut sont égales à 0 pour tous les locus TR car il n’y a pas d’hypermutations somatiques.

iv. ‘Parameters for Detailed View’: ces options sont uniquement disponibles pour le ‘Detailed view’. ‘Nb of nucleotides to exclude in 5′ of the V-REGION for the evaluation of the nb of mutations’ permet d’exclure un certain nombre nucléotides dans l’extrémité 5’ de la V-REGION pour limiter l’impact des amorces (primers utilisés pour les amplifications) lors de l’analyse des mutations. ‘Nb of nucleotides to add (or exclude) in 3' of the V-REGION for the evaluation of the alignment score’ permet de définir le nombre de nucléotides à ajouter ou à exclure en 3’ de la V-REGION pour limiter l’impact d’une forte activité exonucléase lors de l’évaluation du score d’alignement.

3.6.2 IMGT/V-QUEST Results

3.6.2.1 Detailed view

Le «Detailed view», affiche les résultats de l’analyse des séquences individuellement. C’est la sortie par défaut de l’outil IMGT/V-QUEST.

Pour chaque séquence un tableau récapitulatif résume les principaux résultats de l’analyse (Figure 3.8). Chaque ligne du tableau est décrite ci-dessous:

A

Result summary:	**Productive IGH rearranged sequence** (no stop codon and in-frame junction)		
V-GENE and allele	IGHV1-18*01	score = 1426	identity = **99,65%** (287/288 nt)
J-GENE and allele	IGHJ3*02 (a)	score = 164	identity = 81,63% (40/49 nt)
D-GENE and allele by IMGT/JunctionAnalysis	IGHD2-2*02	D-REGION is in reading frame 3	
[CDR1-IMGT.CDR2-IMGT.CDR3-IMGT] lengths and AA JUNCTION	[8.8.9]	CARGIRAFDIW	

(a) Other possibilities: IGHJ6*02 (highest number of consecutive identical nucleotides)

B

Result summary:	**Unproductive IGH rearranged sequence** (stop codons, out-of-frame junction)		
V-GENE and allele	IGHV3-h*01(P)	score = 1321	identity = **96,14%** (274/285 nt)
J-GENE and allele	IGHJ5*02	score = 237	identity = 96,08% (49/51 nt)
D-GENE and allele by IMGT/JunctionAnalysis	IGHD3-3*01	D-REGION is in reading frame 3	
[CDR1-IMGT.CDR2-IMGT.CDR3-IMGT] lengths and AA JUNCTION	[8.7.X]	CSRDKVE*VLRFLEWLFYE#NWFDPW	

C

Result summary:	**Productive IGH rearranged sequence** (no stop codon and in-frame junction) (a)		
V-GENE and allele	IGHV5-a*03	score = 607	identity = **67,71%** (195/288 nt)
J-GENE and allele	IGHJ4*02	score = 163	identity = 82,98% (39/47 nt)
D-GENE and allele by IMGT/JunctionAnalysis	IGHD2-8*01	D-REGION is in reading frame 2	
[CDR1-IMGT.CDR2-IMGT.CDR3-IMGT] lengths and AA JUNCTION	[8.8.17]	MYYCARHGTTSAWSRFDYW (2nd-CYS 104 not identified)	

(a) Low V-REGION identity (67,71%), and non-identification of 2nd-CYS 104. This may indicate potential nucleotide insertion(s) and/or deletion(s): try 'Search for insertions and deletions' in 'Advanced parameters' at the bottom of the Search page

D

>X94075
gaggtgcagctggtgcagtccggagcagaggtgaaaaagcctggggagtctctgaggatc
tcctgtaaggcttctggatacatctttaccagctactggatcaactgggtgcgccagatg
cccgggaaaggcctggagtggatggggaggattgatcctagt**GATCCTAAT**gactcttat
aacaactacaacccgtccttccaaggccacgtcaccatctcagccgacaagtccatcaga
gccgcctatctgcagtggagcagcctggaggcctcagacaccgccatgtattactgtgcg
agacatgggactactagtgcctggtcaagatttgactactggggcgagggaagcctggtc
atcgtctcctca

Result summary:	Nucleotide insertions have been detected and automatically removed for this analysis: they are displayed as capital letters in the user submitted sequence above.					
	localization in V-REGION	nb of inserted nt	inserted nt	causing frameshift	from V-REGION codon	from nt position in user submitted sequence
	CDR2-IMGT	9	GATCCTAAT	no	62	163

IMGT/V-QUEST results after removal of the insertion(s):
Potentially productive IGH rearranged sequence: no stop codon and in-frame junction
(Check also your sequence with BLAST against IMGT/GENE-DB reference sequences to eventually identify out-of-frame pseudogenes)

V-GENE and allele	IGHV5-a*03	score = 1309	identity = **95,14%** (274/288 nt)
J-GENE and allele	IGHJ4*02	score = 163	identity = 82,98% (39/47 nt)
D-GENE and allele by IMGT/JunctionAnalysis	IGHD2-8*01	D-REGION is in reading frame 2	
[CDR1-IMGT.CDR2-IMGT.CDR3-IMGT] lengths and AA JUNCTION	[8.8.14]	CARHGTTSAWSRFDYW	

Figure 3.8: Exemples de tableaux récapitulatifs fournis par IMGT/V-QUEST. **A**. Exemple d'une séquence réarrangée, productive, sans codon stop avec une jonction 'in-frame'. **B.** Exemple d'une séquence réarrangée 'unproductive' avec un codon stop dans la V-D-J-REGION et une jonction est 'out-of-frame'. **C.** Exemple d'une séquence montrant un faible pourcentage d'identité avec le gène germline (67,7%). Une note indique en bas du tableau l'existence d'insertion(s) et/ou délétion(s) potentielles dans la séquence utilisateur. **D.** Exemple d'une séquence analysée avec l'option de recherche des insertions et délétions. Les séquences utilisées pour ces exemples correspondent aux numéros d'accès de IMGT/LIGM-DB [244] AB063682 (**A**), AJ889829 (**B**) et X94075 (**C** et **D**).

1. ***Result summary:*** fournit les informations les plus importantes, c'est-à-dire l'évaluation de la fonctionnalité de la séquence utilisateur. Par exemple la Figure 3.8A montre une séquence d'IG productive, sans aucun codon stop détecté et une jonction 'in frame'.
2. ***V-GENE and allele:*** indique le nom (nomenclature IMGT) des gènes et allèles germline V les plus proches de la séquence utilisateur [3, 237]. Si le gène et allèle déterminé comme étant le plus proche est un pseudogène ou un ORF, la lettre (P) ou les lettres (ORF) sont ajoutées respectivement après le nom du gène et allèle V, par exemple IGHV3-h*01 (P) (Figure 3.8B) indique que l'allèle le plus proche de la séquence utilisateur est un pseudogène. Il est intéressant de noter que des séquences réarrangées productives peuvent résulter d'un réarrangement impliquant un pseudogène, par exemple par l'élimination d'un codon stop durant les réarrangements ou lors d'hypermutations somatiques (HMS). Dans ce cas précis, un message est affiché pour alerter l'utilisateur de la présence d'un pseudogène. Avec le nom du gène et allèle identifié sont indiqués le score d'alignement et le pourcentage d'identité (ainsi que le nombre de nucléotides pris en compte pour l'évaluation). Si plusieurs gènes et allèles ont le même score, ils sont tous indiqués. Un faible pourcentage d'identité peut résulter de caractéristiques particulières dans la séquence utilisateur, telles que les hypermutations somatiques (SHM) par insertions ou par délétions. En clair, quand le pourcentage d'identité est inférieur à 85%, une note est affichée dans le 'Results summary' pour attirer l'attention de l'utilisateur.
3. ***J-GENE and allele***: indique le nom des gènes et allèles J germline les plus proches de la séquence utilisateur [3, 237] ainsi que le score d'alignement et le pourcentage d'identité. L'identification des gènes et allèles J est déterminée par le plus haut score d'alignement. Cependant, une note est ajoutée quand il existe d'autres alternatives comme par exemple des gènes et allèles J ayant un plus grand nombre de nucléotides consécutifs identiques (voir note (a) dans la figure 3.8A). Si la J-REGION délimitée dans la séquence utilisateur fait moins de 6 nucléotides, elle n'est pas considérée comme significative et dans ce cas, le message «Moins de 6 nucléotides sont alignés" est affiché.
4. ***D-GENE and allele by IMGT/JunctionAnalysis:*** indique le nom du gène et allèle D germline le plus proche [3, 237] de la séquence utilisateur, identifié par le programme IMGT/JunctionAnalysis [238], avec son cadre de lecture dans la séquence utilisateur. Cependant, les utilisateurs ont la possibilité de changer le nombre de mutations

tolérées dans la jonction (3' V-REGION, D-REGION et 5'J-REGION) lors de la recherche de la D-REGION (paramètres avancés dans IMGT/V-QUEST Search page). IMGT/JunctionAnalysis gère toutes les difficultés intrinsèques de l'analyse des jonctions et de la détermination du gène et allèle D (une taille réduite de la D-REGION, l'usage du 2^{ième} ou 3^{ième} cadre de lecture, l'élimination des nucléotides dans la jonction par les exonucléases et la présence de mutations).

5. ***CDR1-IMGT, CDR2-IMGT, CDR3-IMGT lengths and AA JUNCTION:*** définit la taille des 3 CDR-IMGT (CDR1-IMGT, CDR2-IMGT et le CDR3-IMGT) et affiche la JUNCTION en AA. Ces informations sont des caractéristiques importantes de la séquence. De plus, la différence de longueur des CDR1-IMGT et CDR2-IMGT lorsqu'elles sont comparées aux CDR1-IMGT et CDR2-IMGT de la séquence germline peut indiquer la présence d'insertions et de délétions. Dans ce cas une note est ajoutée au tableau pour attirer l'attention de l'utilisateur. La taille du CDR3-IMGT dépend du réarrangement des gènes V-D-J ou V-J. Un X définit une séquence dont le CDR3-IMGT n'est pas déterminé, dans le cas par exemple d'une jonction 'out-of-frame' [8,7,X] (Figure 3.8B). La séquence de la jonction en AA permet de visualiser si:

    - la jonction est 'in-frame' ou 'out-of-frame' (un changement de cadre de lecture est montré par le signe dièse '#'). La présence ou l'absence de codons stop est indiqué (les codons stop sont identifiés par des astérisques '*').
    - la présence ou l'absence des acides aminés conservés la 2nd-CYS en position 104, W J-TRP en position118 dans la chaîne lourde ou F J-PHE en position 118 dans les autres types de chaîne.
    - des motifs spécifiques éventuels peuvent être caractéristiques des réarrangements des LLC.

Alignements, jonction, mutations, IMGT Collier de Perles

Les détails des résultats de l'analyse des séquences utilisateurs sont affichés à la suite du tableau récapitulatif: ils comprennent les alignements obtenus pour l'identification des gènes et allèles V, D et J, les résultats de l'analyse de la jonction par IMGT/JunctionAnalysis, les différents alignements de la V-REGION ('V-REGION alignment' et 'V-REGION translation'), les tableaux regroupant les résultats de l'analyse des mutations dans la séquence

utilisateur ‘V-REGION mutation table’, et la ‘V-REGION mutation statistics’, le ‘IMGT Collier de Perles’.

Alignments for V, D and J GENE and allele identification

IMGT/V-QUEST affiche les alignements entre la séquence utilisateur et les cinq gènes et allèles germline V, D et J les plus proches, ainsi que leur score d’alignement et leur pourcentage d’identité.

Results of IMGT/JunctionAnalysis

Les résultats de IMGT/JunctionAnalysis (Figure 3.9) comprennent l’analyse de la jonction avec:

1. les détails de la jonction au niveau nucléotidique (en précisant les nucléotides éliminés à la fin des gènes V, D et J et les N nucléotides ajoutés par la TdT) avec la délimitation précise des extrémités de la V-REGION en 3’, de la D-REGION et de la J-REGION en 5’. Les points représentent les nucléotides qui ont été éliminés par l’activité des exonucléases. Le nombre de mutations dans la 3’V-REGION, D-REGION et 5’J-REGION est indiqué, par exemple Vmut 0, Dmut 1 et Jmut 0 (Figure 3.9), et les nucléotides mutés correspondants sont soulignés. Les nucléotides de la N-REGION sont affichés dans N1 et N2. Le ratio du nombre de nucléotide g/c sur le total de nucléotides est indiqué sous le titre ‘Ngc’, par exemple caa (N1) et gc (N2) donne un Ngc de 3/5 (Figure 3.9). Le nombre de mutations acceptées dans les jonctions IGH sont de 2 dans la 3’V-REGION, de 4 dans la D-REGION et de 2 dans la 5’J-REGION. Cependant dans le cas de gènes IGHV non mutés (pas de mutations du FR1-IMGT jusqu’au FR3-IMGT), ces nombres sont modifiés: 0 pour la 3’V-REGION et 5’J-REGION, et 2 pour la D-REGION dans le but de refléter la faible probabilité de HMS.
2. Les gènes éligibles ou ‘Eligible D gènes’: cette option permet aux utilisateurs de comparer le gène D identifié par IMGT/JunctionAnalysis avec tous les D gènes candidats qui ont un score d’alignement au minimum de quatre. Le gène et allèle D avec le meilleur score, sélectionné par IMGT/JunctionAnalysis, est encadré dans le rectangle rouge de la figure 3.9. La localisation d[8-28]s[13-33] signifie que les nucléotides aux positions 8 à 28 du gène germline IGHD3-22*01 identifié comme étant le plus proche de la séquence utilisateur correspondent aux nucléotides des positions 13 à 33 de la séquence.

3. ‘Translation of the JUNCTION’ affiche les séquences en acides aminés de la jonction, et le cadre de lecture (le symbole ‘+’ indique que la jonction est in-frame alors que le ‘-‘ signifie que la jonction est out-of-frame, la taille du CDR3-IMGT, la masse moléculaire et le point isoélectrique (pI). Les acides aminés sont coloriés selon les 11 classes d’acides aminés définies par IMGT [241]. Dans le cas d’un frameshift (changement de cadre de lecture), les gaps (représentés par un ou des points) sont insérés pour maintenir le cadre de lecture de la J-REGION et faciliter la comparaison des séquences. Le codon correspondant qui ne peut pas être traduit est représenté par ‘#’. La séquence des jonctions est disponible en texte, pour simplifier l’exportation, par sélection des options dans les paramètres avancés de IMGT/V-QUEST.

Analysis of the JUNCTION

D-REGION is in reading frame 2.

Input	V name	3'V-REGION	N1	D-REGION	N2	5'J-REGION	J name	D name	Vmut	Dmut	Jmut	Ngc
seq1	IGHV1-69*01	tgtgcgaga..	caa	.....tatggtagtagtggttattac...	gc	.....ctactgg	IGHJ4*01	IGHD3-22*01	0	1	0	3/5

Eligible D genes

D name	D length	tgtgcgagacaatatggtagtagtggttattacgcctactgg	Score#	Mutation#	Location
IGHD1-1*01	17	----ca-c---	8	3	d[1-11],s[16-26]
IGHD2-2*01	31	----t--------	12	1	d[4-16],s[12-24]
IGHD2-2*02	31	----t--------	12	1	d[4-16],s[12-24]
IGHD2-2*03	31	----t--------	12	1	d[4-16],s[12-24]
IGHD3-3*01	31	----------	11	0	d[17-27],s[22-32]
IGHD3-3*02	31	----------	11	0	d[17-27],s[22-32]
IGHD4-4*01	16	--a-c---	6	2	d[9-16],s[26-33]
IGHD5-5*01	20	------	6	0	d[14-19],s[24-29]
IGHD6-6*01	18	---a-c--	6	2	d[4-11],s[13-20]
IGHD1-7*01	17	----ta-c---	8	3	d[1-11],s[16-26]
IGHD2-8*01	31	----t---c--a----	13	3	d[4-19],s[12-27]
IGHD2-8*02	31	----t---c-g-----	13	3	d[4-19],s[12-27]
IGHD3-9*01	31	---------	9	0	d[19-27],s[24-32]
IGHD3-10*01	31	---ta----a--	9	3	d[20-31],s[22-33]
IGHD3-10*02	30	---ta----a--	9	3	d[19-30],s[22-33]
IGHD4-11*01	16	--a-c---	6	2	d[9-16],s[26-33]
IGHD5-12*01	23	---ata------c--	11	4	d[2-16],s[15-29]
IGHD6-13*01	21	--c--ct-g--	7	4	d[10-20],s[19-29]
IGHD1-14*01	17	---a--	5	1	d[3-8],s[28-33]
IGHD2-15*01	31	----t-----g-----	14	2	d[4-19],s[12-27]
IGHD3-16*01	37	--g---ta--c----	11	4	d[20-34],s[16-30]
IGHD3-16*02	37	--g---ta-c------	11	4	d[20-34],s[16-30]
IGHD4-17*01	16	---g-c---	7	2	d[8-16],s[25-33]
IGHD5-18*01	20	------	6	0	d[14-19],s[24-29]
IGHD6-19*01	21	---c-----c-gg---	12	4	d[6-21],s[18-33]
IGHD1-20*01	17	----ta-c---	8	3	d[1-11],s[16-26]
IGHD2-21*01	28	----t--g--g---a--	13	4	d[4-20],s[12-28]
IGHD2-21*02	28	----t--g--g---	11	3	d[4-17],s[12-25]
IGHD3-22*01	31	----a----------	20	1	d[8-28],s[13-33]
IGHD4-23*01	19	--c---g---	8	2	d[5-14],s[13-22]
IGHD5-24*01	20	----aga---c--	9	4	d[1-13],s[17-29]

Translation of the JUNCTION

	104	105	106	107	108	109	110	112	113	114	115	116	117	118	Frame	CDR3-IMGT length	Molecular mass	pI
	C	A	R	Q	Y	G	S	S	G	Y	Y	A	Y	W				
seq1	tgt	gcg	aga	caa	tat	ggt	agt	agt	ggt	tat	tac	gcc	tac	tgg	+	12	1,674.81	8.89

Figure 3.9: Résultats de IMGT/JunctionAnalysis. Les résultats de l'analyse par IMGT/JunctionAnalysis de la seq1 avec le numéro d'accès DQ100777 dans IMGT/LIGM-DB [244], comprend 'Analysis of the JUNCTION', 'Eligible D genes' et 'Translation of the JUNCTION' avec les acides aminés coloriés en accord avec les classes physicochimiques de IMGT [241].

'V-REGION alignment' and 'V-REGION translation'

Ces deux alignements sont particulièrement intéressants pour visualiser les mutations. Le 'V-REGION alignment' affiche l'alignement entre la séquence utilisateur avec les cinq gènes et allèles V germline les plus proches. Dans cet alignement les FR-IMGT et les CDR-IMGT sont délimités, les nucléotides mutés sont affichés tandis que les tirets correspondent à des nucléotides identiques. La 'V-REGION translation' (Figure 3.10) montre l'alignement entre la séquence utilisateur avec le gène et allèle V le plus proche et sa traduction en acides aminés. Les FR-IMGT et CDR-IMGT sont délimités dans l'alignement, et dans le cas de mutations non silencieuses les changements d'AA sont indiqués au-dessus du codon muté. Cet alignement permet de visualiser dans la séquences les mutations décrites dans les tableaux 'V-REGION mutation table' et 'V-REGION mutation statistics'.

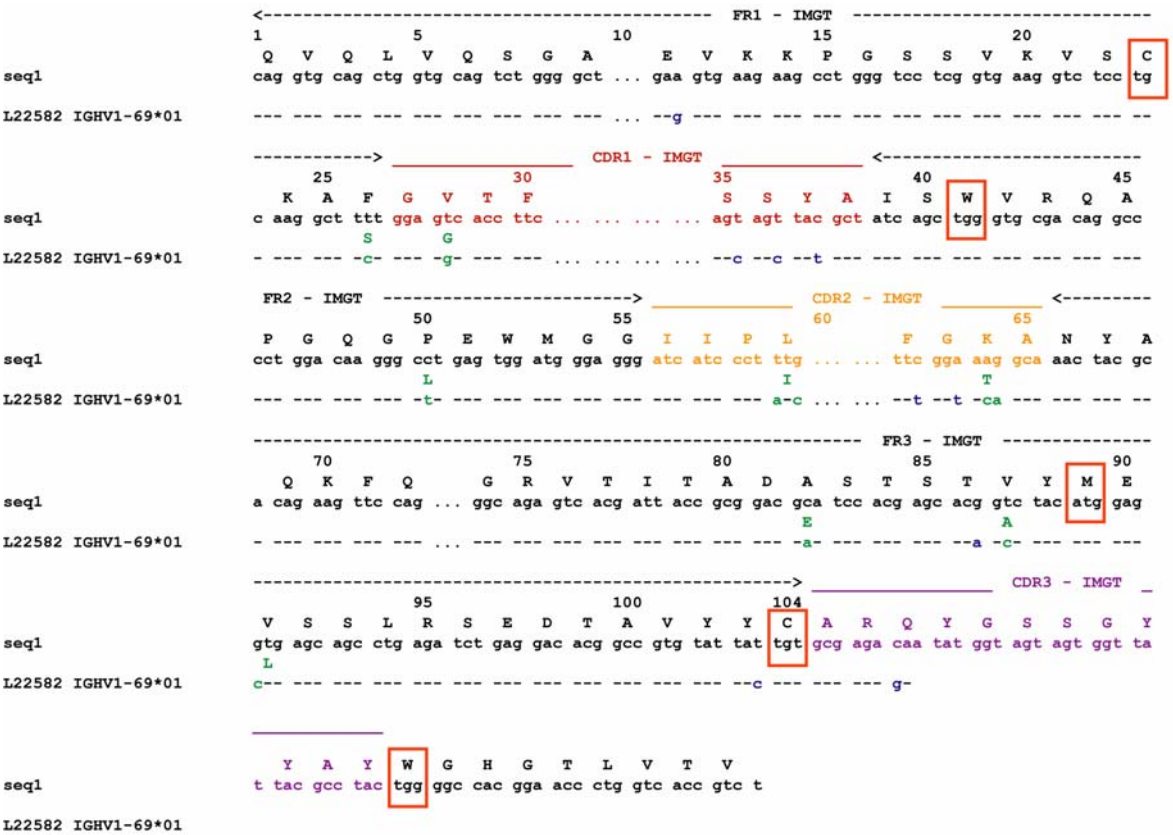


Figure 3.10: 'V-REGION translation'. La séq1 correspond au numéro d'accès DQ100777 dans la base de données IMGT/LIGM-DB [244]. Le CDR1-IMGT, CDR2-IMGT, CDR3-IMGT sont coloriés en rouge, orange et violet respectivement. Les mutations silencieuses, par comparaison avec le gène et allèle V germline le plus proche (IGHV1-69*01) sont en bleu. Les mutations non silencieuses et les changements d'acides aminés qui en résultent sont en vert.

'V-REGION mutation table' and 'V-REGION mutation statistics'

Le tableau 'V-REGION mutation table' (Figure 3.11A) liste les mutations en nucléotides et en acides aminés dans la séquence utilisateur comparée avec la séquence germline du gène et allèle V le plus proche. Les mutations sont décrites pour la V-REGION et pour chaque FR-IMGT et CDR-IMGT avec pour chacune d'elle, sa position en accord avec les règles de la numérotation unique IMGT [14]. Pour les mutations non silencieuses, on indique entre parenthèses si les propriétés des acides aminés originaux sont conservées après mutation (hydropathie, volume et caractéristiques physicochimiques). Par exemple c260>t, A87>V (+ - +) signifie que la mutation du nucléotide c en t conduit au codon 87 à un changement d'acide aminé qui, par rapport à l'AA original, est dans la même classe pour l'hydropathie (+), dans une classe différente pour le volume (-) et dans une même classe pour les propriétés physicochimiques (+).

Le tableau 'V-REGION mutation statistics' (Figure 3.11B) évalue le nombre de mutations silencieuses et non silencieuses, le nombre de transitions et de transversions dans la séquence utilisateur. Pour éviter toute erreur d'interprétation lors de l'étude des mutations dans la séquence utilisateur qui contient une amorce dans la partie 5' de la V-REGION, il est possible d'exclure un nombre donné de nucléotides, défini par l'utilisateur (les paramètres peuvent être modifiés pour le 'Detailed view' sur la page de IMGT / V-QUEST Search page).

Le tableau 'V-REGION mutation hot spot' montre la localisation des motifs hot spot dans le gène et allèle V germline le plus proche de la séquence utilisateur. La séquence utilisateur en nucléotides et en acides aminés peut être exportée en texte en utilisant l'option «Sequence of V-, V-J ou V-D-J-REGION ('nt and 'AA) with gaps in FASTA and access to IMGT/PhyloGene for V-REGION ('nt')». Cette option propose un lien vers l'outil de phylogénie d'IMGT, IMGT/PhyloGene [245].

A - V-REGION mutation table

FR1-IMGT	CDR1-IMGT	FR2-IMGT	CDR2-IMGT	FR3-IMGT	CDR3-IMGT
g33>a	g83>t, G28>V (- - -)	t149>c, L50>P (- - -)	a175>t, I59>L (+ + +)	a245>c, E82>A (- - -)	g319>c
c77>t, S26>F (- - -)	c105>t		c177>g, I59>L (+ + +)	a258>g	
	c108>t		t186>c	c260>t, A87>V (+ - +)	
	t111>c		t189>a	c271>g, L91>V (+ - +)	
			c191>a, T64>K (- - -)	c309>t	
			a192>g, T64>K (- - -)		

B - V-REGION mutation statistics

Nucleotides

IMGT labels		V-REGION	FR1-IMGT	CDR1-IMGT	FR2-IMGT	CDR2-IMGT	FR3-IMGT	CDR3-IMGT
Nb of positions including IMGT gaps (nt)		**318** (320)	78	36	51	30	117	6 (8)
Nb of nucleotides		**294** (296)	75	24	51	24	114	6 (8)
Nb of identical nucleotides		**276** (277)	73	20	50	18	109	6 (7)
Mutations	Total	**18** (19)	2	4	1	6	5	0 (1)
	Silent	**8** (9)	1	3	0	2	2	0 (1)
	Nonsilent	**10**	1	1	1	4	3	0
Transitions	a>g	**2**	0	0	0	1	1	0
	g>a	**1**	1	0	0	0	0	0
	c>t	**5**	1	2	0	0	2	0
	t>c	**3**	0	1	1	1	0	0
Transversions	a>c	**1**	0	0	0	0	1	0
	c>a	**1**	0	0	0	1	0	0
	a>t	**1**	0	0	0	1	0	0
	t>a	**1**	0	0	0	1	0	0
	g>c	**0** (1)	0	0	0	0	0	0 (1)
	c>g	**2**	0	0	0	1	1	0
	g>t	**1**	0	1	0	0	0	0
	t>g	**0**	0	0	0	0	0	0

Amino acids

IMGT labels			V-REGION	FR1-IMGT	CDR1-IMGT	FR2-IMGT	CDR2-IMGT	FR3-IMGT	CDR3-IMGT
Nb of positions including IMGT gaps (AA)			**106**	26	12	17	10	39	2
Nb of AA			**98**	25	8	17	8	38	2
Nb of identical AA			**90**	24	7	16	6	35	2
AA changes	Total		**8**	1	1	1	2	3	0
	Conserved IMGT AA classes (hydropathy, volume, chemical)	(- - -)	**5**	1	1	1	1	1	0
		(+ + +)	**1**	0	0	0	1	0	0
		(+ - -)	**0**	0	0	0	0	0	0
		(- + -)	**0**	0	0	0	0	0	0
		(- - +)	**0**	0	0	0	0	0	0
		(+ + -)	**0**	0	0	0	0	0	0
		(+ - +)	**2**	0	0	0	0	2	0

Figure 3.11: Caractérisation des mutations. Lors de l'analyse de la séquence DQ100777 de la base de données IMGT/LIGM-DB [244]. **A**. Le 'V-REGION mutation table' indique les mutations des nucléotides et des AA (pour les mutations non silencieuses) avec leur position en accord avec la numérotation unique IMGT [14]. **B**. Le 'V-REGION mutation statistics' caractérise les mutations des nucléotides (transitions et transversions) et des AA, dans la V-REGION, et par FR-IMGT et CDR-IMGT. Les statistiques sont calculées jusqu'à la fin de la V-REGION en 3' déterminée par les 2 nucléotides identiques consécutifs en 3' avec le gène et allèle V germline le plus proche. Les nombres entre parenthèses dans les colonnes intitulées V-REGION et CDR3-IMGT correspondent aux statistiques calculées jusqu'à la fin de la V-REGION du gène et allèle V germline le plus proche.

IMGT Collier de Perles

IMGT/V-QUEST fournit un lien vers l'outil IMGT/Collier-de-Perles. Il permet d'afficher automatiquement l'IMGT Collier de Perles [115, 246, 247] du domaine V, à partir de la traduction de la séquence utilisateur (Figure 3.12). L'outil IMGT/Collier-de-Perles permet de relier la séquence à la structure des récepteurs d'antigènes [116], il est très largement utilisé dans les protocoles d'humanisation des anticorps ou pour évaluer les anticorps monoclonaux thérapeutiques [5]. Ces informations sont utiles pour localiser les acides aminés des CDR-IMGT qui peuvent être impliqués dans les contacts avec l'antigène, mais également pour visualiser, par exemple la répartition des motifs stéréotypés dans les CDR3 des patients atteints de LLC (voir chapitre 2). Il offre la possibilité à l'utilisateur d'étudier les propriétés physicochimiques des acides aminés à une position donnée dans un jeu de séquences, et de comparer les acides aminés mutés avec le IMGT Collier de Perles du gène et allèle V germline le plus proche de la séquence utilisateur. Dans l'IMGT Collier de Perles, le CDR1-IMGT, CDR2-IMGT et CDR3-IMGT sont respectivement coloriés en rouge, orange et violet pour le domaine V de la chaîne lourde et en bleu, vert et vert-bleu pour les CDR-IMGT du domaine V de la chaîne légère. Les ancres (acides aminés situés dans les FR-IMGT qui encadrent les CDR-IMGT) sont affichées dans des carrés. Dans les FR-IMGT, les acides aminés hydrophobes et le tryptophane (W) qui sont observés dans plus de 50% des séquences à une position donnée sont affichés avec une couleur de fond bleu. L'outil IMGT/Collier-de - Perles peut être personnalisé pour afficher les acides aminés, coloriés en fonction de leur hydropathie, leur volume ou encore leurs propriétés physicochimiques en accord avec les classes AA d'IMGT [241]. Par défaut, les IMGT Colliers de Perles sont affichés sur un seul plan. Ils peuvent également être affichés en deux plans afin d'obtenir une représentation graphique plus proche de la structure 3D. Il est à noter que dans le cas de jonction out-of-frame, le cadre de lecture du J n'est pas restauré (la séquence ne comprend pas le signe '#').

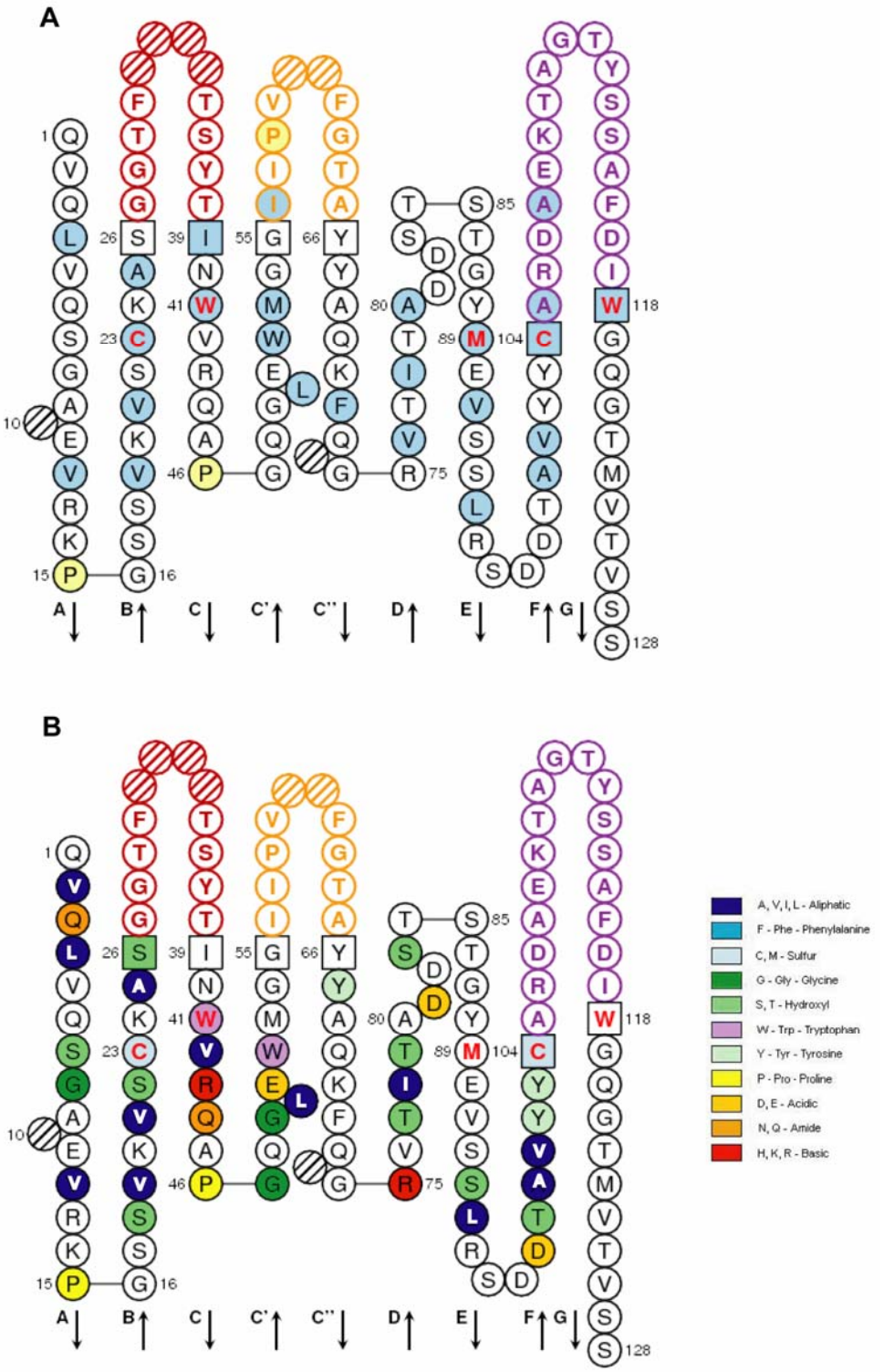


Figure 3.12: IMGT Colliers de Perles. A. Les acides aminés hydrophobes et le tryptophane (W) conservés à une position donnée dans plus de 50% des séquences sont coloriés en bleu. B. Les acides aminés d'une classe physicochimique donnée conservés à une position donnée dans plus de 80% des séquences sont coloriés en accord avec les classes physicochimiques d'IMGT [241]. Dans la séquence dont le numéro d'accès d'IMGT/LIGM-DB [244] est AF021992, la C 23 (1st-CYS), W 41 (CONSERVED-TRP), l'acide aminé hydrophobe (M) 89, et le C 104 (2nd-CYS) sont en rouge. Les cercles rayés correspondent aux positions manquantes selon la numérotation unique IMGT [14]. Les flèches indiquent la direction des feuillets beta antiparallèles et la longueur des CDR-IMGT est [8.8.17].

3.6.2.2 Synthesis view

La sortie ‘Synthesis view’ est une alternative intéressante au ‘Detailed view’. Cette vue facilite la comparaison des séquences qui expriment le même gène et allèle V via différents alignements, et fournit les résultats de IMGT/JunctionAnalysis des séquences par locus (Figure 3.13).

Summary table

Le tableau ‘Summary table’ qui se trouve en haut de la page de résultat regroupe les principales informations déduites de l’analyse des séquences utilisateur. Le tableau montre les résultats de l’analyse de chaque séquence utilisateur, avec le gène et allèle V germline le plus proche, l’évaluation de la fonctionnalité de chaque séquence, le score d’alignement de la V-REGION et le pourcentage d’identité entre les séquences utilisateur et la V-REGION du gène et allèle germline le plus proche. Le tableau regroupe également le nom des gènes et allèles J et D, le cadre de lecture de la D-REGION, la taille des trois CDR-IMGT, la jonction en AA, et le cadre de lecture de la JUNCTION (Figure 3.13A). Le tableau ‘Summary table’ peut également inclure des notes pour alerter l’utilisateur sur des insertions ou des délétions potentielles dans la V-REGION ou encore sur la possibilité d’autres gènes et allèles J germline. Dans le cas où cette configuration se produit, il est alors fortement recommandé de vérifier les résultats des analyses individuelles des séquences utilisateur concernées avec le ‘Detailed view’.

Synthesis view: alignments, junction, mutations

Alignments

Le ‘Synthesis view’ fournit différents alignements entre les séquences utilisateur qui sont réarrangées avec le même gène et allèle V germline. Trois types d’alignements sont fournis aux utilisateurs et sont du même type que ceux présentés dans le ‘Detailed view’, l’alignement du gène V (‘Alignment for V-GENE’), l’alignement de la V-REGION (‘V-REGION alignment’) et l’alignement de la traduction de la V-REGION (‘V-REGION translation'). Ces différents alignements permettent une comparaison aisée des séquences avec une localisation évidente des mutations. Dans tous les alignements, les motifs correspondant à des hot spots dans la V-REGION de la séquence germline sont soulignés (Figure 3.13B), et le nom du gène et allèle J le plus proche est indiqué à la fin de la séquence en 3’.

'V-REGION protein display'

'V-REGION protein display' affiche les alignements entre les séquences utilisateur en acide aminés avec la V-REGION du gène et allèle V germline le plus proche. Trois types d'alignements sont proposés à l'utilisateur, un alignement simple 'V-REGION protein display', un alignement avec les AA coloriés selon les classes AA de IMGT [241] 'V-REGION protein display with colored AA according to the AA IMGT Classes' et finalement un alignement avec uniquement les AA mutés par rapport à la séquence germline.

'V-REGION most frequently occurring AA per position and per FR-IMGT and CDR-IMGT'

Cette sortie affiche un tableau pour les FR1-IMGT, CDR1-IMGT, FR2-IMGT, CDR2-IMGT, FR3-IMGT avec pour chaque position l'acide aminé le plus représenté.

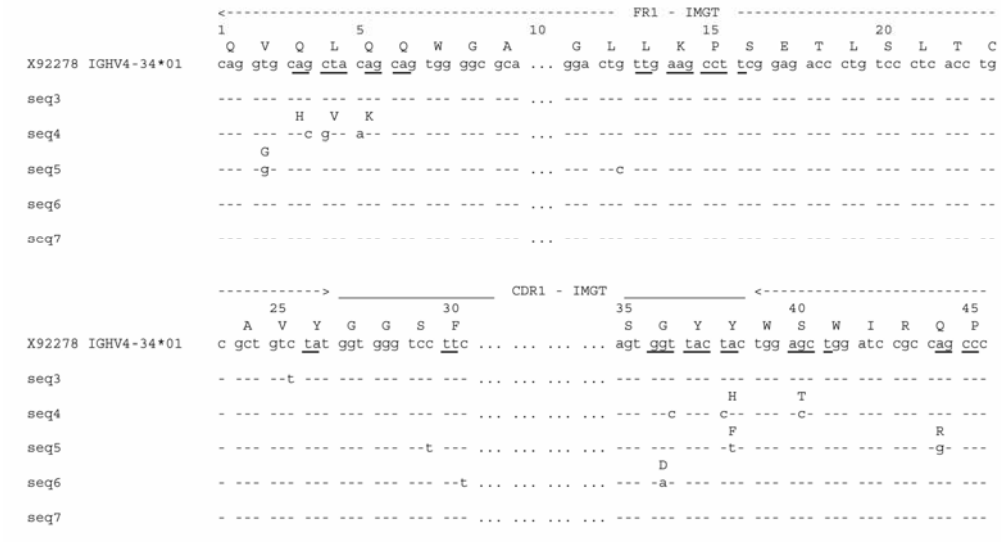
'Results of IMGT/JunctionAnalysis'

'Results of IMGT/JunctionAnalysis' affiche les résultats de l'analyse des jonctions des séquences appartenant à un même locus par IMGT/JunctionAnalysis (IGH, IGK et IGL pour les chaînes lourdes, kappa et lambda pour les chaînes légères, TRA, TRB, TRG et TRD pour les chaînes alpha, bêta, gamma et delta des TR). (Figure 3.13C).

A Summary table

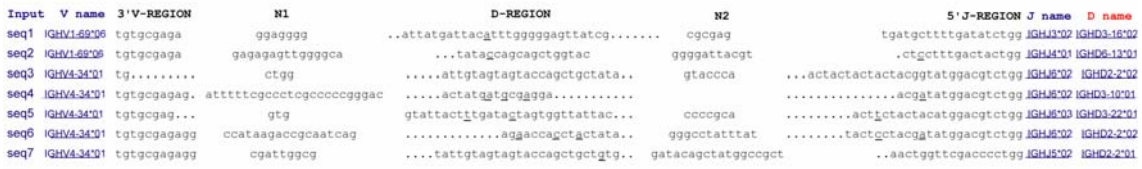
Sequence ID	V-GENE and allele	Functionality	V-REGION score	V-REGION identity % (nt)	J-GENE and allele	D-GENE and allele	D-REGION reading frame	CDR-IMGT lengths	AA JUNCTION	JUNCTION frame
seq1	IGHV1-69*06	Productive	1050	95,63% (219/229 nt)	IGHJ3*02	IGHD3-16*02	2	[8.8.21]	CARGGDYDYIWGSYRASDAFDIW	in-frame
seq2	IGHV1-69*06	Productive	1300	94,79% (273/288 nt)	IGHJ4*01	IGHD6-13*01	1	[8.8.21]	CARERVGAYTSSWYGDYVSFDYW	in-frame
seq3	IGHV4-34*01	Productive	1411	99,30% (283/285 nt)	IGHJ6*02	IGHD2-2*02	3	[8.7.20]	CWIVVVPAAIVPNYYYYGMDVW	in-frame
seq4	IGHV4-34*01	Productive	1294	95,09% (271/285 nt)	IGHJ6*02	IGHD3-10*01	2	[8.7.20]	CARDFSPSPPGHYDARNDMDVW	in-frame
seq5	IGHV4-34*01	Productive	1285	94,74% (270/285 nt)	IGHJ6*03	IGHD3-22*01	2	[8.7.21]	CARWYYFDTSGYYPRNFYYMDVW	in-frame
seq6	IGHV4-34*01	Productive	1258	93,68% (267/285 nt)	IGHJ6*02	IGHD2-2*02	3	[8.7.25]	CARGHKTAIREPPTIGPIYYSYDMDVW	in-frame
seq7	IGHV4-34*01	Productive	1420	100,00% (285/285 nt)	IGHJ5*02	IGHD2-2*01	3	[8.7.25]	CARGDWRIVVVPAAVDTAMAANWFDPW	in-frame

B Sequences aligned with IGHV4-34*01: V-REGION translation



C Results of IMGT/JunctionAnalysis:

Analysis of the JUNCTIONs



Translation of the JUNCTIONs

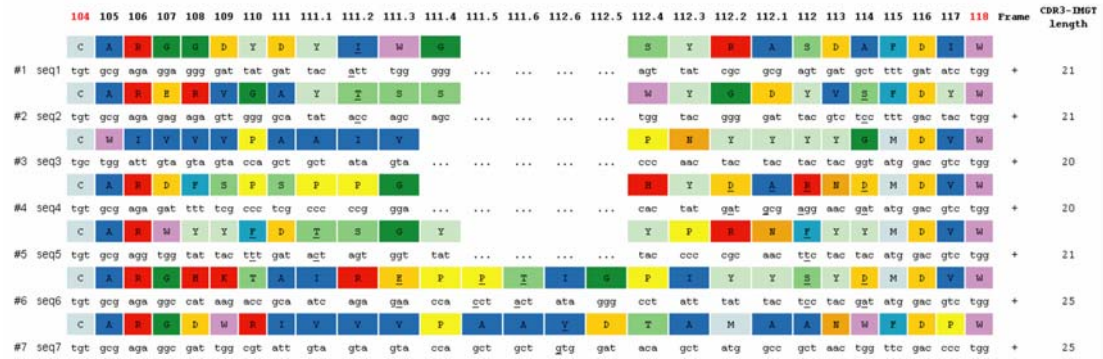


Figure 3.13: IMGT/V-QUEST 'Synthesis view'. A. 'Summary table'. Tableau des principaux résultats de l'analyse. Les séquences, seq 1 (AY998726), seq 2 (AY998742), seq3 (DQ100725), seq4 (AF022004), seq5 (X84333), seq6 (AF022002) et seq7 (AF022005) proviennent de la base de données IMGT/LIGM-DB [244]. B. Extrait de 'V-REGION translation' pour les séquences alignées avec l'allèle IGHV4-34*01. C. 'Results of IMGTJunctionAnalysis'. Le résultat comprend l'analyse de la jonction 'Analysis of the JUNCTIONs', et la traduction de la jonction 'Translation of the JUNCTIONs' avec les acides aminés coloriés selon la classification AA d'IMGT [241].

3.7 Etat de l'art des logiciels d'analyse des séquences réarrangées des IG et TR

IMGT/V-QUEST a été le premier logiciel dédié à l'analyse des séquences réarrangées des IG et des TR. Il est accessible sur le web depuis 1997. Entre 1997 et 2008, d'autres outils ont été développés, tous accessible sur le Web, selon des approches similaires ou différentes. Nous établissons ici un état de l'art de l'existant. Depuis 1997, 6 logiciels ont été développés par différentes équipes, JOINSOLVER [248], IgBLAST (http://www.ncbi.nlm.nih.gov/igblast/), VDJsolver [249], SoDA [250], DNAPLOT [251], iHMMune-align [252]. A l'exception de VDJsolver qui est un outil dédié à l'analyse de la jonction des séquences réarrangées des IG et qui par conséquent se rapproche d'avantage d'IMGT/JunctionAnalysis [238], tous les autres programmes sont dédiés à l'analyse des réarrangements VDJ. Contrairement à IMGT/V-QUEST qui est capable d'analyser les séquences réarrangées des IG et des TR, les autres outils, à l'exception de SoDA, sont limités aux IG. Cependant pour SoDA, il est nécessaire de sélectionner chaque locus indépendamment. Pour la plupart, ces outils utilisent comme séquences de référence les séquences provenant de IMGT®. Cependant, les sites à l'exception d'IgBLAST ne renseignent pas les utilisateurs sur la date de mise à jour ou le suivi des données IMGT®. En ce qui concerne SoDA la publication indique l'utilisation d'IMGT® mais aucune information n'est fournie sur le site, tandis que DNAPLOT utilise les séquences de VBASE [253]. Progressivement les outils se convertissent à la numérotation unique IMGT (IMGT unique numbering) [14] qui est devenu le standard international. A l'heure actuelle VDJsolver, SoDA et iHMMune-align continuent d'utiliser la numérotation de Kabat, ce qui maintient un état de confusion.

L'ensemble de ces outils fournit le nom des gènes et allèles V, D et J les plus proches et la plupart les alignements entre la séquence utilisateur et les gènes et les allèles les plus proches.

JOINSOLVER

JOINSOLVER [248] analyse uniquement les IG d'une seule espèce (humaine). Il est basé sur un algorithme spécialisé dans la recherche de motifs spécifiques conservé pour délimiter les régions à aligner (sans gaps), associé à un algorithme d'alignement qui ne permet d'analyser qu'une seule séquence à la fois. JOINSOLVER ne fournit pas le pourcentage d'identité entre la séquence utilisateur et les gènes et les allèles les plus proches ce qui ne permet pas de visualiser rapidement les séquences les plus proches, et ne peux pas

être utilisé pour l'aide au pronostic de la LLC via le taux de mutations des IGHV. JOINSOLVER ne détecte pas les insertions ou les délétions potentielles dans les séquences utilisateurs et ne fournit aucune annotation de la séquence. Enfin JOINSOLVER ne propose pas de vue synthétique des alignements par gène, ni de représentation 2D de la séquence utilisateur.

IgBLAST

IgBLAST (http://www.ncbi.nlm.nih.gov/igblast/) analyse les IG pour deux espèces (homme et souris). Il utilise un algorithme d'alignement de séquences de type Blast. L'intérêt de ce programme est de pouvoir analyser des séquences très inhabituelles (translocation). C'est le seul outil qui présente une démarche complémentaire à IMGT/V-QUEST. Cependant pour les séquences classiques ou avec des insertions/délétions modérées, IgBLAST fournit beaucoup moins d'informations qu'IMGT/V-QUEST et celles-ci risquent d'être moins précises. IgBLAST n'offre pas d'analyse de la jonction, ni d'analyse des mutations, pas d'évaluation de la fonctionnalité, pas d'annotation de la séquence, pas de résumé des principaux résultats, pas de vue synthétique des alignements par gène, ni de représentation 2D de la séquence utilisateur.

VDJsolver

VDJsolver [249] analyse la jonction des séquences réarrangées des IG pour une seule espèce (humaine), et par conséquent se rapproche d'avantage d'IMGT/JunctionAnalysis [238]. VDJsolver utilise deux programmes JointHMM qui est un algorithme probabiliste sur un modèle de Markov caché et JointML qui est un algorithme basé sur une méthode de maximum-likelihood qui prend en compte la longueur de la jonction et le statut mutationnel des gènes IGHV. Par rapport à IMGT/JunctionAnalysis, VDJsolver ne fournit pas la traduction de la séquence utilisateur en acides aminés et d'alignements, ce qui ne permet pas de vérifier et de comparer la pertinence des gènes et allèles germline définis comme les plus proche. Il n'identifie pas les nucléotides P.

SoDA

SoDA [250] analyse les séquences réarrangées IG et TR pour trois espèces (homme, souris, opposum) et également les IGH du Rhesus monkey (Macaca mulatta). Il utilise une variation de l'algorithme d'alignement de séquences de type programmation dynamique. Il ne

peut pas analyser les séquences de plus de 645 nucléotides. Le nombre de fonctionnalités est limité :
SoDA ne fournit pas de pourcentage d'identité entre la séquence utilisateur et les gènes et les allèles les plus proches, pas d'analyse de la jonction, pas d'annotation de la séquence, pas de vue synthétique des alignements par gène, ni de représentation 2D de la séquence utilisateur.

DNAPLOT

DNAPLOT [251] analyse les séquences réarrangées des IG de deux espèces (homme et souris). Il utilise un algorithme d'alignement de séquences. DNAPLOT ne fournit pas le pourcentage d'identité entre la séquence utilisateur et les gènes et allèles les plus proches identifiés. Lors de l'analyse des mutations, DNAPLOT ne fournit pas de détails sur le nombre de mutations dans la région V par FR et CDR, ni le type des mutations identifiées. Il ne permet pas d'identifier la présence des insertions ou des délétions dans la séquence utilisateur. D'autre part le nombre de fonctionnalités est limité: DNAPLOT ne fournit pas d'évaluation de la fonctionnalité, d'annotation de la séquence, pas de résumé des principaux résultats, pas de vue synthétique des alignements par gène, ni de représentation 2D de la séquence utilisateur.

iHMMune-align

iHMMune-align [252] analyse les séquences réarrangées des IG pour une seule espèce (humaine). Il utilise un modèle probabiliste basé sur un algorithme utilisant un modèle de Markov caché (MMC) (Hidden Markov Models, HMM). iHMMune-align ne fournit pas le pourcentage d'identité entre la séquence utilisateur et les gènes et les allèles germline les plus proches. Lors de l'analyse de la jonction IHMMune-align n'identifie pas les nucléotides P. Il ne permet pas d'identifier la présence des insertions ou des délétions dans la séquence utilisateur, ne fournit aucune annotation de la séquence. Enfin iHMMune-align ne propose pas de vue synthétique des alignements par gène, ni de représentation 2D de la séquence utilisateur.

Conclusion

IMGT/V-QUEST est un outil spécialisé dans l'analyse standardisée des séquences réarrangées d'IG et TR. L'ajout de nouvelles fonctionnalités (évaluation de la fonctionnalité, localisation et caractérisation des mutations, détermination des insertions/délétions, le

‘Synthesis view') et l’utilisation des paramètres avancés fournit un large éventail de nouveaux types d'analyses.
IMGT/V-QUEST était en 1997 le premier outil accessible sur le Web. Par l’intégration de nouvelles fonctionnalités, par son large choix d’options et de paramètres personnalisables, et par son niveau élevé de standardisation, IMGT/V-QUEST reste unique parmi les autres logiciels, JOINSOLVER, VDJSolver, SoDA, iHMMune [248-250, 252], IgBLAST (http://www.ncbi.nlm.nih.gov/igblast/) spécialisés dans l’analyse des récepteurs d’antigènes. Les informations fournies par IMGT/V-QUEST sont en particulier indispensables pour l'analyse comparative de séquences d'IG et TR et des répertoires en situation normale et pathologique, pour les analyses statistiques de la jonction [254], pour l'ingénierie des anticorps et des anticorps thérapeutiques [5].

Dans le but de concevoir un système d’information dédié à l’analyse et à la gestion des récepteurs d’antigènes appliqué à une pathologie du système immunitaire la LLC, le premier objectif était d’adapter l’outil spécialisé IMGT/V-QUEST pour le rendre:

1) plus simple et plus souple dans son développement et dans sa maintenance.
2) portable afin de l’intégrer au sein d’un ensemble plus vaste en interaction directe avec un système de gestion de données (base de données, outil d’administration, interface utilisateur).
3) plus performant en ajoutant des fonctionnalités pour enrichir l’analyse des séquences réarrangées correspondant aux attentes des utilisateurs et particulièrement dans le domaine de la recherche clinique.

Ma contribution à IMGT/V-QUEST s’est traduite par une réécriture du coeur de l’outil. La première version d’IMGT/V-QUEST était constituée de deux modules de fonctions distincts. Le premier module en langage C, était responsable de l’algorithme d’alignement et de la détermination des gènes et allèle V, D et J de la séquence utilisateur. Le second module codé en langage JAVA était responsable de la gestion de l’interface utilisateur et de la mise en forme des résultats.
Dans un premier temps, j’ai analysé et adapté la méthode de détermination des gènes et allèles V, D et J, pour réécrire le programme dans le but d’intégrer les 2 modules en une seule et même structure (codée en JAVA). Ce travail de réécriture a permis de simplifier et de réorganiser le programme mais également d’améliorer la cohérence, les performances, et la portabilité de l’outil.

Dans un deuxième temps, j'ai contribué aux perfectionnements et à l'enrichissement significatif de l'analyse des séquences réarrangées d'IG et TR, par l'ajout de nouvelles fonctionnalités, avec:

- l'évaluation systématique du pourcentage d'identité entre chaque région V, D et J de la séquence utilisateur et les cinq gènes et allèles germline les plus proches (le marqueur pronostic le plus pertinent actuellement pour la LLC est le statut mutationnel des gènes IGHV).
- la localisation précise et la description complète des mutations présentes dans la V-REGION de la séquence utilisateur.
- l'identification de la position des nucléotides préférentiellement mutés appelés hot spots dans le gène et allèle V germline le plus proche de la séquence utilisateur.
- la détection des insertions et des délétions de nucléotides pouvant survenir dans la V-REGION des séquences réarrangées d'IG lors des mécanismes d'hypermutations somatiques.
- l'intégration du programme IMGT/Automat [255] qui permet d'obtenir une annotation complète des V-J et V-D-J-REGION.
- l'amélioration des performances d'IMGT/V-QUEST qui nous a permis proposer des analyses par lots pouvant contenir jusqu'à 50 séquences.

Parallèlement, l'interface utilisateur a été entièrement refondue. L'interface utilisateur de IMGT/V-QUEST est maintenant paramétrable en fonction des besoins des utilisateurs. Les paramètres avancés offrent le choix d'une analyse personnalisée selon la problématique rencontrée: choix des répertoires des séquences de références, recherche des insertions et des délétions, paramètres de l'outil IMGT/JunctionAnalysis, possibilité d'exclure un nombre de nucléotides de la partie 5' de la V-REGION pour l'évaluation des mutations, possibilité d'exclure en 3' de la V-REGION pour l'évaluation du score. IMGT/V-QUEST affiche maintenant deux types de résultats le 'Detailed view' qui correspond à l'analyse de séquences individuelles et le 'Synthesis view' qui présente les alignements des séquences qui expriment le même gène et allèle V. Enfin, nous avons défini un tableau récapitulatif présentant les principaux résultats de l'analyse, avec notamment une évaluation de la fonctionnalité de la séquence utilisateur (productive ou unproductive) et des alertes pour avertir l'utilisateur de caractéristiques particulières telles que la possibilité d'insertions ou de délétions dans la séquence, ou la possibilité d'avoir un pseudogène etc…

IMGT/V-QUEST est largement utilisé par la communauté scientifique, avec une moyenne de 30.000 requêtes depuis le début 2008. Nous avons enregistré au mois de septembre 2008 des pics de 50.000 requêtes sur un seul week-end. Dans ce contexte, nous avons le projet d'adapter IMGT/V-QUEST à une analyse à haut débit. Ceci nécessite le développement de méthodologies spécifiques reliées au calcul intensif, pour la soumission des séquences, l'analyse des données et la restitution des résultats. La mise en place de IMGT/V-QUEST haut débit va de pair avec l'intégration de nouvelles fonctionnalités inhérentes à l'analyse d'un grand nombre de séquences (telles que la réalisation de statistiques sur les résultats obtenus pour chaque pool de séquences).

Un second projet consiste à adapter IMGT/V-QUEST à l'analyse des séquences protéiques. Une modification de l'algorithme d'alignement pourrait nous permettre de déterminer les insertions et les délétions en routine. L'utilisation d'un algorithme d'alignement de type semi-global, dérivé de l'alignement de Needleman et Wunsch permettrait lors de l'identification des gènes V et J de 1) délimiter le début de la région en cours d'analyse, 2) de numéroter la V-REGION de la séquence utilisateur en fonction de la séquence la plus proche et 3) de détecter les insertions et les délétions en une seule et même étape. Cette approche pourra être envisagée car les séquences protéiques contiennent un nombre de résidus (ici acides aminés) à comparer plus limité que les séquence nucléotidiques. Il sera également nécessaire de définir une matrice de substitution qui prendra en compte la classification IMGT des acides aminés.

CHAPITRE 4

IMGT/CLL-DB

Les différents perfectionnements apportés à l'outil d'analyse IMGT/V-QUEST, décrits dans le chapitre précédent, nous ont permis de concevoir et d'intégrer au sein d'IMGT® un système d'information dédié à l'analyse et au stockage des séquences réarrangées des récepteurs d'antigène, associées à des informations relatives à des patients atteints d'une pathologie du système immunitaire. La conception et l'implémentation du système d'information ont été élaborées pour être appliquées à la leucémie lymphoïde chronique, dans le cadre d'une collaboration internationale avec des équipes cliniques, spécialistes de la LLC (Annexe 8): celle-ci nous ont apporté d'une part l'expertise médicale nécessaire à la mise en place du système et d'autre part les séquences et les données relatives aux patients. Ce système d'information a été conçu comme un outil en appui à la recherche médicale. Il présente l'originalité de corréler les caractéristiques génétiques des IG exprimées dans les cellules leucémiques de patients atteints de la LLC à certaines données cliniques afin de mieux cerner les mécanismes impliqués. Notons ici que ce système n'a pas pour vocation de gérer la totalité des données cliniques des patients.

Le système d'information est constitué de l'outil IMGT/V-QUEST, couplé à la base de données IMGT/CLL-DB accessible par une interface Web. En raison de la nature du projet, qui inclut l'analyse, la gestion et la restitution de données confidentielles médicales de patients atteints de la LLC, il a été nécessaire de modifier l'approche d'IMGT en matière de gestion des bases données (toutes les autres bases d'IMGT étant publiques) pour n'autoriser que les partenaires du projet à accéder au système. D'autre part, nous avons réactualisé les technologies mises en œuvre, en nous appuyant sur les standards actuels des applications Web.

IMGT/CLL-DB fournit (1) les séquences et les résultats de leur analyse par IMGT/V-QUEST (identification des gènes et allèles V-D-J, le pourcentage d'identité, une évaluation de la fonctionnalité, la JUNCTION en acides aminés, la taille des 3 CDR-IMGT, le cadre de lecture de la jonction, la fiche complète des résultats de l'analyse 'Detailed view' de IMGT/V-QUEST), (2) les données relatives aux patients et à leur pathologie (état civil préservant l'anonymat, tests cliniques, évolution de la maladie, etc …), et (3) la description de l'échantillon biologique (date de prélèvement, origine tissulaire , etc…).

4.1 Organisation de IMGT/CLL-DB

La base de données IMGT/CLL-DB est gérée par le gestionnaire de base de données de type relationnel Sybase (http://www.sybase.com), rapide, largement répandu et utilisé pour la gestion de toutes les bases de données IMGT®. Les données sont gérées dans des tables organisées en 5 systèmes ou groupes (Figure 4.1). (1) Les données relatives aux patients et aux pathologies (Disease System) décrivent et définissent les patients et leur pathologie. (2) Les données relatives aux échantillons (Sample System) définissent les caractéristiques des échantillons telles que le tissu, le statut, ainsi que la date des prélèvements. (3) Les données relatives aux séquences (Sequence System) décrivent et caractérisent les séquences d'immunoglobulines. (4) Le système automatique d'analyse (Automatic Analysis System) regroupe les annotations des séquences d'IG analysées par le programme IMGT/V-QUEST. Enfin (5) les données relatives à la gestion des utilisateurs (User System) permettent de gérer les droits d'accès de connexion à la base de donnée IMGT/CLL-DB. Dans le but de gérer l'historique et le suivi des modifications, nous avons créé deux tables 'rawHistoricPat' et 'rawHistoricSeq' respectivement intégrées dans les 'Disease System' et 'Sequence System', chargées de répertorier l'intégralité des modifications qui concernent les données relatives aux séquences et celles relatives aux patients.

Les trois premiers systèmes répertorient les données fournies par les laboratoires. L'organisation des tables, des champs et le vocabulaire contrôlé, résulte d'un consensus établi par tous les laboratoires participant au projet. Les données relatives aux patients et les données relatives aux échantillons biologiques qui doivent être gérées dans la base ont été sélectionnées lors de réunions préliminaires avec les partenaires du projet. Seules les informations médicales les plus pertinentes pour le projet ont été sélectionnées, et pour chaque type de données, dans la mesure du possible, un vocabulaire a été défini.

Les données relatives aux séquence fournies par IMGT/CLL-DB sont conformes aux axiomes IDENTIFICATION, DESCRIPTION et NUMEROTATION d'IMGT-ONTOLOGY [13] et aux concepts correspondants d'identification, de description, de numérotation ainsi qu'aux règles de la charte scientifique d'IMGT générées à partir de ces concepts.

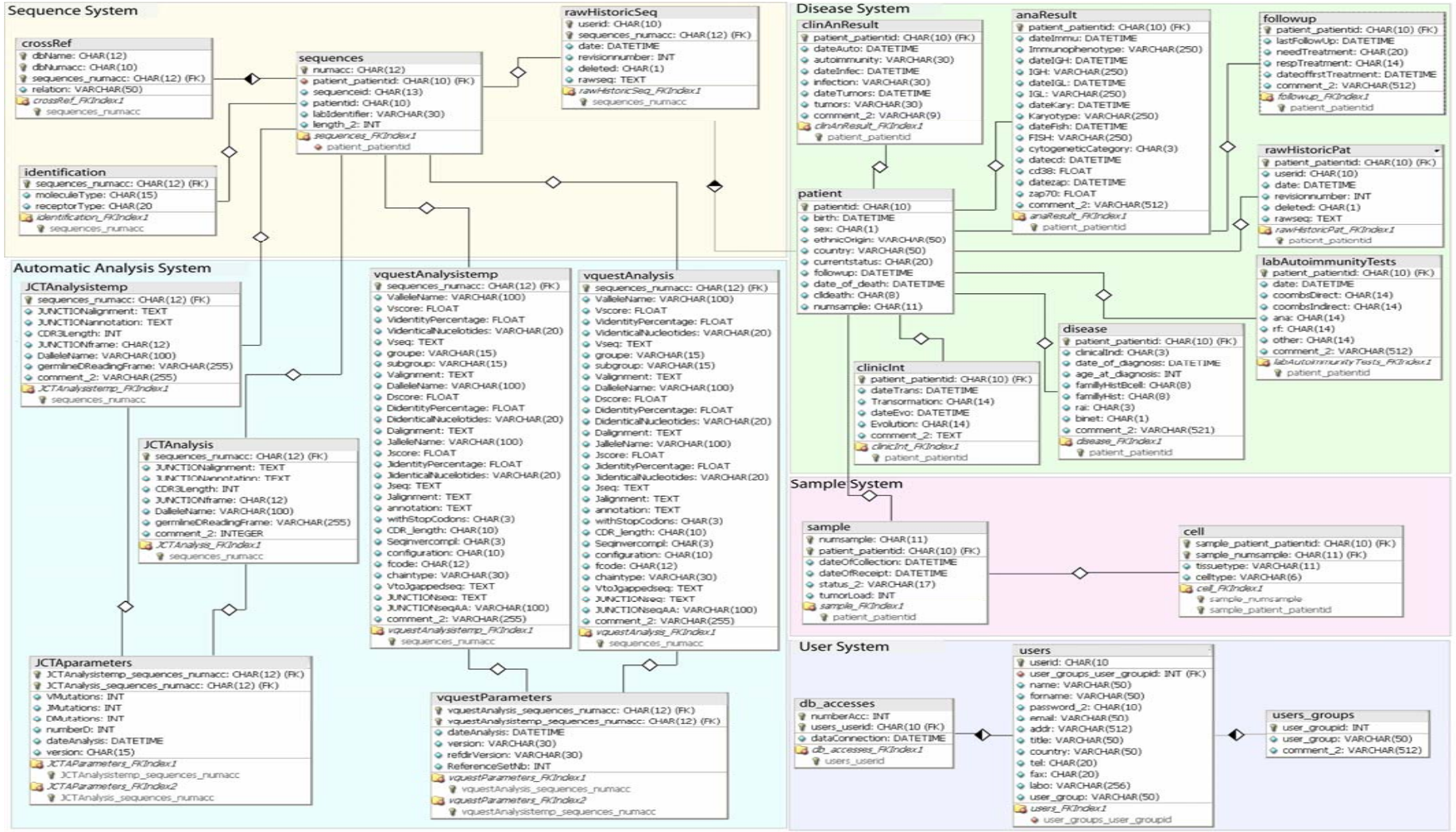


Figure 4.1: Organisation de la base de données IMGT/CLL-DB. Les tables sont organisées en 5 groupes: 'Sequence System' gère les séquences d'immunoglobulines, 'Disease System' décrit les patients et leur état clinique, 'Sample System' définit les données relatives aux échantillons (tissu, statut…), 'Automatic Analysis System' regroupe les annotations des séquences d'IG analysées par le programme IMGT/V-QUEST et 'User System' gère les utilisateurs et les connexions à la base de donnée IMGT/CLL-DB. Les relations entre ces tables sont représentées et la cardinalité de chaque relation est indiquée par: ◇ pour une relation de type 1:1, où une entrée de la table A est reliée à une entrée de la table B et par ◆ pour une relation de type 1:n, où une entrée de la table A peut être reliée à n entrées de la table B.

4.1.1 Données relatives aux séquences

Ce sont les informations centrales de la base de données gérées dans le 'Sequence System', toutes les autres tables y sont reliées. Chaque séquence de la base est associée à deux identifiants uniques différents:

1) le numéro d'accès interne à IMGT/CLL-DB (IMGT/CLL-DB accession number), défini automatiquement lors du chargement des données dans la base; il sert à identifier de façon unique chaque séquence de la base.
2) l'identifiant de la séquence (Sequence ID) dont le format a été défini en accord avec les équipes cliniques partenaires. Il est constitué d'une série de lettres et de chiffres (par exemple: GR-02-0001-H1), les deux premières lettres définissent le pays d'origine du laboratoire propriétaire de la séquence, les deux chiffres suivants définissent le laboratoire d'origine, enfin la suite de l'identifiant (une série de 4 chiffres suivie d'une lettre (H, L ou K) et d'un chiffre) identifie de façon unique chaque séquence du laboratoire ainsi que leur type de chaînes.

Les laboratoires propriétaires des séquences peuvent également indiquer leur propre identifiant, généralement celui utilisé dans leur système local, sur lequel n'est effectué aucun contrôle. Cet identifiant est présent dans la base dans le seul but de faciliter le travail des utilisateurs qui ont l'habitude de manipuler leurs données. Toutes les séquences sont caractérisées par leur taille (table Sequences), leur type moléculaire (gDNA ou cDNA), leur type de récepteur (immunoglobuline ou récepteur T) (table Identification), la date de dépôt de la séquence dans la base données, un historique des modifications (table rawHistoricSeq). Certaines des séquences de la base IMGT/CLL-DB peuvent également avoir été publiées dans les bases de données publiques, elles sont alors reliées par leur numéro d'accès à la base de données IMGT/LIGM-DB [244] (table crossRef).

4.1.2 Données relatives aux patients

'Disease System' est constitué de 7 tables qui regroupent les données relatives aux patients et à leur pathologie. Chaque patient est défini par un identifiant et caractérisé (1) par une fiche d'identité (préservant l'anonymat) (table Patient), (2) par les données décrivant et caractérisant leur pathologie (table Disease), (3, 4) par les données concernant leur suivi médical (tables followUp, clinicInt), (5, 6, 7) ainsi que les résultats des principaux tests

cliniques effectués dans le cadre de la prise en charge de leur pathologie en l'occurrence la CLL (tables anaResult, clinResult, LabAutoimmunityTests).

(1) Table 'patient'

La table 'patient' correspond à la fiche d'identité de chaque patient et préserve leur anonymat. Chaque patient est caractérisé par un identifiant, son sexe, sa date de naissance, son origine ethnique, son pays de domicile, son statut actuel (vivant, décédé ou inconnu), la dernière date du suivi médical par l'équipe actuelle, la date du décès potentiel, et enfin indique le cas échéant, si le patient est mort des suites de la pathologie (LLC).

(2) Table 'disease'

La table 'disease' identifie et définit la pathologie par la date du diagnostic, l'âge du patient au diagnostic ainsi que les antécédents familiaux du patient en rapport avec la LLC et avec d'autres pathologies en rapport avec un syndrome lymphoprolifératif des cellules B, et le stade de la maladie selon les deux systèmes anatomo-cliniques actuels de classification de la LLC. La classification de Binet [190] subdivise l'évolution de la LLC en trois stades distincts A, B et C correspondant respectivement à un stade précoce, intermédiaire et avancé de la LLC. La classification de Rai [189] subdivise l'évolution de la maladie en 5 classes distinctes 0, I, II, III, IV correspondant pour les classes 0 et I à un risque faible, II à un risque intermédiaire et enfin III et IV à un risque élevé.

(3 - 4) Tables 'followup' et 'clinicInt'

La table 'followup' relate les informations concernant le suivi médical, avec la date de la dernière consultation, si un traitement est nécessaire, la date du début du premier traitement et la réponse au traitement.

La table 'clinicInt' décrit l'évolution de la maladie en trois phases ('stable', 'en progression' ou en 'régression') et définit les éventuelles complications vers une forme agressive de la maladie, en particulier pour la LLC, le développement d'un lymphome (syndrome de Richter) et la transformation en leucémie prolymphocytaire, en leucémie aiguë ou en myélome multiple. Chaque donnée est associée à une date.

(5) Table 'anaResult'

La table 'anaResult' caractérise pour chaque patient l'immunophénotypage des lymphocytes B selon qu'ils sont typiques de la LLC ou atypiques, l'isotype des chaînes

lourdes et légères exprimé à la surface des cellules B, le caryotype du patient, les résultats de l'analyse par la technologie FISH et la classification de la catégorie cytogénétique du patient. Elle contient également les valeurs de deux facteurs de pronostic le CD38 et le ZAP 70. Chaque résultat est associé à une date, excepté pour la catégorie cytogénétique.

(6 – 7) Tables 'LabAutoimmunityTests' et 'clinAnResult'

La table 'LabAutoimmunityTests' contient les résultats des tests classiques de recherche des manifestations autoimmunes. Les tests répertoriés comprennent le Coombs direct (positif ou négatif), le Coombs indirect (positif ou négatif), la recherche d'auto anticorps avec les dépistages des facteurs rhumatoïdes (RF) et d'anticorps antinucléaires (ANA) (positif ou négatif). A chaque résultat est associé une date.
La table 'clinAnResult' décrit les éventuelles complications qui peuvent survenir à la suite de la LLC et leur date de diagnostic. Il existe trois types de complications possibles: des maladies autoimmunes, des infections et des cancers.

4.1.3 Données relatives aux échantillons

Chaque patient possède son propre identifiant auquel sont reliées les informations relatives aux échantillons ('Sample System). Chaque échantillon est défini par un identifiant et décrit par la date de son prélèvement, la date de réception par les cliniciens responsables du patient, l'état physique de l'échantillon lors de sa réception, à savoir s'il est frais ou congelé, le nombre de cellules B cancéreuses mesurées dans l'échantillon (table sample) et le tissu dans lequel a été effectué le prélèvement et le type de cellules prélevées (cellules B dans le cas des LLC) (table cell).

4.1.4 Système automatique d'analyse

Toutes les séquences présentes dans la base de données IMGT/CLL-DB sont analysées et les résultats enregistrés dans les tables du système 'Automatic Analysis System' 'vquestAnalysis' et 'JCTAnalysis' avec respectivement 30 types de résultats d'IMGT/V-QUEST et 7 résultats d'IMGT/JunctionAnalysis [238] potentielement accessible à la requête. La table 'vquestAnalysis' répertorie, le nom des gènes et allèles V, D et J associé à leur score de comparaison et leur pourcentage d'identité, le nombre de nucléotides identiques entre la V-, D- et J-REGION des séquences utilisateur, et les V-, D- et J-REGION respectivement des gènes et allèles les plus proches, les V-, D- et J-REGION en nucléotides, l'orientation de la séquence (sens ou antisens), la séquence numérotée selon la numérotation unique IMGT, la

séquence de la jonction en nucléotides et en acides aminés, l'alignement des V-, D- et J-REGION, la position des codons stop, la longueur des CDR-IMGT, le groupe et le sous-groupe de la V-REGION selon la nomenclature IMGT, les données d'annotation de la séquence, la configuration de la séquence (rearranged), la fonctionnalité de la séquence (productive ou unproductive) et le type de chaîne.
La table 'JCTAnalysis' répertorie, le nom du gène et allèle D, l'alignement de la jonction, les annotations de la jonction (avec délimitation des P-REGION et des N-REGION, évaluation du nombre de hypermutations somatiques pour chaque gène dans la jonction), la longueur du CDR3-IMGT, le cadre de lecture de la jonction, et finalement le cadre de lecture de la D-REGION.
Les tables 'vquestAnalysis' et 'JCTAnalysis' sont dupliquées pour former des tables 'VquestAnalysistemp' et 'JCTAnalysistemp', qui stockent temporairement les résultats de l'analyse par IMGT/V-QUEST des séquences en cours de chargement dans la base.
Les tables 'JCTAparameters' et 'vquestParameters' répertorient les paramètres définis pour analyser les séquences contenues dans la base de donnée IMGT/CLL-DB.

4.1.5 Système de gestion des utilisateurs

Le système utilisateur 'User System' est constitué de trois tables dédiées à la gestion des comptes utilisateurs, la table 'Users' sert à la gestion des droits d'accès et à l'exploitation des données de la base IMGT/CLL-DB, la table 'User_groups' permet de différencier les utilisateurs de la base en différents groupes selon leur rôle dans le projet: soit simple utilisateur, soit utilisateur et fournisseur de données. Enfin la table 'db_accesses' est utilisée uniquement en interne et sert au suivi des connexions vers la base de données IMGT/CLL-DB.

4.2 Administration de IMGT/CLL-DB

IMGT/CLL-DB est maintenue par un système semi-automatique. Les données sont fournies directement par les cliniciens via des fichiers Excel dont le format répond à des règles strictes déterminées en accord avec l'ensemble des partenaires (ordre des colonnes, vocabulaire contrôlé).
L'insertion des données dans la base est effectuée par un programme Java dédié responsable de l'ensemble des modalités de chargement: vérification du respect du fichier Excel,

lancement des analyses des séquences par le programme IMGT/V-QUEST, vérification du bon déroulement des insertions des données dans la base.

La base de donnée IMGT/CLL-DB comprend un système de validation de l'intégrité des données lors de chargement ou de mise à jour. Ce système d'intégrité est constitué de 3 niveaux de sécurité complémentaires; (1) un vocabulaire contrôlé (Annexe 7), (2) des contraintes d'intégrité (règles définissant dans la base de données le type et la taille des informations pouvant être stockées pour un champ donné) et (3) des 'Triggers' (programmes informatiques sur le serveur qui déclenchent des actions de contrôle lors de l'insertion ou lors de la mise à jour de données). Le vocabulaire contrôlé est une liste de termes qui n'a pas l'ambiguïté du langage naturel. Dans une base de données, le vocabulaire contrôlé garantit qu'un sujet sera décrit avec les mêmes termes préférentiels. Durant une recherche, nous pouvons ainsi trouver plus facilement tous les renseignements relatifs à un sujet précis. Cette liste a été définie en accord avec l'ensemble des partenaires du projet. Les 'Triggers' et contraintes d'intégrité permettent de minimiser le risque d'erreur lors de chargement et de mise à jour en vérifiant automatiquement l'intégrité des données, avec par exemple: le respect des formats, des dates, des valeurs numériques, du vocabulaire contrôlé, de la syntaxe des identifiants ou bien encore l'unicité des données.

La base de données IMGT/CLL-DB implémente un système de suivi des données dans le temps et un système de suivi de l'historique des opérations en sauvegardant automatiquement toutes les entrées dans les tables dédiées à l'historique.

Finalement, nous avons instauré un protocole de correction des données en accord avec tous les partenaires du projet. L'administrateur de la base est habilité à faire des mises à jour ou des corrections seulement avec l'accord formel des propriétaires des données concernées.

4.2.1 Sélection des nouvelles données

Les données stockées dans la base IMGT/CLL-DB sont de deux types. Les séquences et les données associées aux patients atteints de LLC fournies par les cliniciens, et les annotations automatiques obtenues par analyse bioinformatique, avec le programme expert IMGT/V-QUEST. Toute nouvelle entrée est fournie par les utilisateurs de la base, sous forme de fichier Excel. Deux types de fichiers Excel ont été définis pour faciliter la préparation des données et limiter au maximum les erreurs de saisie.

- Un fichier Excel dédié aux nouvelles séquences avec les données relatives aux nouveaux patients.

- Un fichier Excel dédié aux nouvelles séquences dont les patients correspondants ont été préalablement enregistrés dans la base de données.

Le format des fichiers Excel a été défini en accord avec l'ensemble des partenaires du projet pour faciliter son utilisation. Ils sont constitués de 58 (informations complètes) ou 5 (si information partielles) colonnes correspondant aux différentes informations fournies par les cliniciens et sont subdivisés en 6 grandes thématiques: 'Sequences', 'Sample', 'Patient', 'Analytic tests', 'Clinical analysis', et 'Clinical interpretation'. Dans les fichiers Excel chaque ligne correspond à une entrée c'est-à-dire à une séquence. Cinq colonnes doivent être obligatoirement être complétées afin qu'une nouvelle entrée puisse être enregistrée dans la base de données IMGT/CLL-DB: (1) La séquence en nucléotides, (2) l'identifiant de la séquence, (3) l'identifiant du laboratoire, (4) le type de molécule, (5) le numéro d'accès de la séquence dans la base de données IMGT/LIGM-DB [244] s'il existe. Les autres colonnes sont libres d'être remplies ou non par les cliniciens. Le respect de l'ordre des colonnes dans les fichiers Excel est primordial pour la procédure de chargement dans la base de données. Les règles de vocabulaire contrôlé mises en place pour la base de données IMGT/CLL-DB ont été intégrées au fichier Excel (par un système de macros) créant ainsi un fichier de saisie contenant pour chaque champ une liste unique de valeurs possibles.

4.2.2 Chargement des nouvelles données

Le chargement des données est effectué uniquement par le manager de la base par l'intermédiaire d'un programme en langage JAVA qui gère l'ensemble des étapes de chargement. L'outil de chargement prend en entrée le fichier Excel envoyé par les laboratoires et préalablement transformé en fichier CSV. Il vérifie d'abord l'intégrité du fichier, détermine automatiquement un numéro d'accès pour les nouvelles séquences, vérifie le format des identifiants de la séquence et du patient. Si aucune séquence dans la base ne correspond à la séquence prise en charge par le programme de chargement, celui-ci vérifie alors si les données respectent le vocabulaire contrôlé mis en place pour cette application, vérifie la validité des données fournies par le cliniciens (validité des dates et des valeurs numériques), il détermine l'âge du patient au diagnostic de la LLC, détermine la date du chargement des données, et lance le programme expert IMGT/V-QUEST pour l'analyse des séquences dont les résultats sont chargés dans les tables temporaires ('vquestAnalysistemp' et 'JCTAnalysistemp'). Finalement, il vérifie que le chargement des données dans la base ne rencontre aucune erreur. Dans le cas contraire le programme annule le chargement du lot en cours de traitement et le programme envoie au gestionnaire un fichier explicatif contenant

pour chaque entrée ayant provoqué une erreur les causes du problème. Si aucune erreur n'est générée lors du chargement des données ou de l'analyse par IMGT/V-QUEST, les données sont insérées définitivement dans la base, les données des tables temporaires 'vquestAnalysistemp' et 'JCTAnalysistemp' sont copiées dans les tables 'vquestAnalysis' et 'JCTAnalysis', et les tables temporaires sont effacées.
Si des erreurs sont générées au cours du chargement, le protocole instauré dans le cadre de la gestion de la base de données IMGT/CLL-DB interdit au gestionnaire de modifier par lui-même les données fournies par les cliniciens. Le gestionnaire doit alors soumettre les erreurs rencontrées au responsable du lot incriminé pour vérification et/ou correction.

4.2.3 Mises à jour

La procédure de mise à jour est identique à la procédure de chargement des nouvelles données. Les fichiers Excel utilisés sont les mêmes, mais il est également possible d'utiliser un fichier dédié uniquement aux données de patients. L'utilisateur doit simplement remplir les colonnes appropriées avec les données de séquences ou de patients à mettre à jour. Le même programme implémenté en JAVA pour le chargement des données est utilisé pour la mise à jour. Il vérifie l'intégrité des données, les dates, les valeurs numériques, le vocabulaire contrôlé, et l'existence des données qui doivent être mises à jour.
Cette procédure permet de faire des mises à jour à grande échelle sur de grands lots de séquences. Pour des mises à jour ponctuelles, une interface web a été implémentée.

4.2.4 Interface de gestion des données

Nous avons récemment mis en place une interface Web dédiée à la gestion des données. Cette interface permet une interaction simple et performante entre le gestionnaire et la base de données. Cette interface permet à ce jour d'effectuer uniquement des mises à jour ponctuelles (Figure 4.2). Le gestionnaire a la possibilité d'éditer, pour chaque entrée de la base, les données relatives aux patients, aux échantillons et à la séquence (excepté pour les annotations obtenues par l'analyse par IMGT/V-QUEST, le numéro d'accès de la séquence, l'identifiant de la séquence, l'identifiant du patient, l'identifiant de l'échantillon et enfin l'âge du patient au moment du diagnostic). Il peut modifier et sauvegarder les valeurs correspondantes dans la base. L'interface de gestion de la base IMGT/CLL-DB permet de sélectionner une entrée unique, soit par le numéro d'accès de la séquence (IMGT/CLL-DB accession number), ou bien encore par l'identifiant de la séquence (Sequence ID).

Les informations sont affichées et triées en 6 catégories: (Figure 4.2) ‘Sequences’, ‘Disease’, ‘Sample’, ‘Analytical tests’, ‘Autoimmunity tests’, et ‘Other tests’. Pour chaque catégorie, l'interface permet d’éditer les données et d’obtenir un formulaire de saisie pré-rempli. Les champs sont renseignés par les données stockées dans la base (Figure 4.3). Chaque mise à jour est enregistrée dans les tables 'rawHistoricPat' et 'rawHistoricSeq' suivant les données mises à jour. Elle est identifiée par une combinaison de valeurs, le numéro d’accès ou l'identifiant du patient et un numéro définissant le nombre de modifications effectuées à ce jour. Dans le cadre du protocole instauré pour la gestion des données, le gestionnaire est autorisé à modifier les données uniquement sous la recommandation des propriétaires.

Cette interface de gestion de la base IMGT/CLL-DB est destinée dans le futur aux utilisateurs qui auront la possibilité de modifier leurs données. Ils pourront directement, sans passer par le gestionnaire de la base, corriger d’éventuelles erreurs et mettre à jour les données selon l’évolution de l’état du patient. Ils seront autorisés à modifier seulement les entrées dont ils sont propriétaires.

L'interface de management a été implémentée en langage Java en accord avec les technologies J2EE et l'utilisation des framework ‘Strust’ et ‘Hibernate’. (cf. chapitre 2.3.1 Implémentation et architecture de l'interface Web).

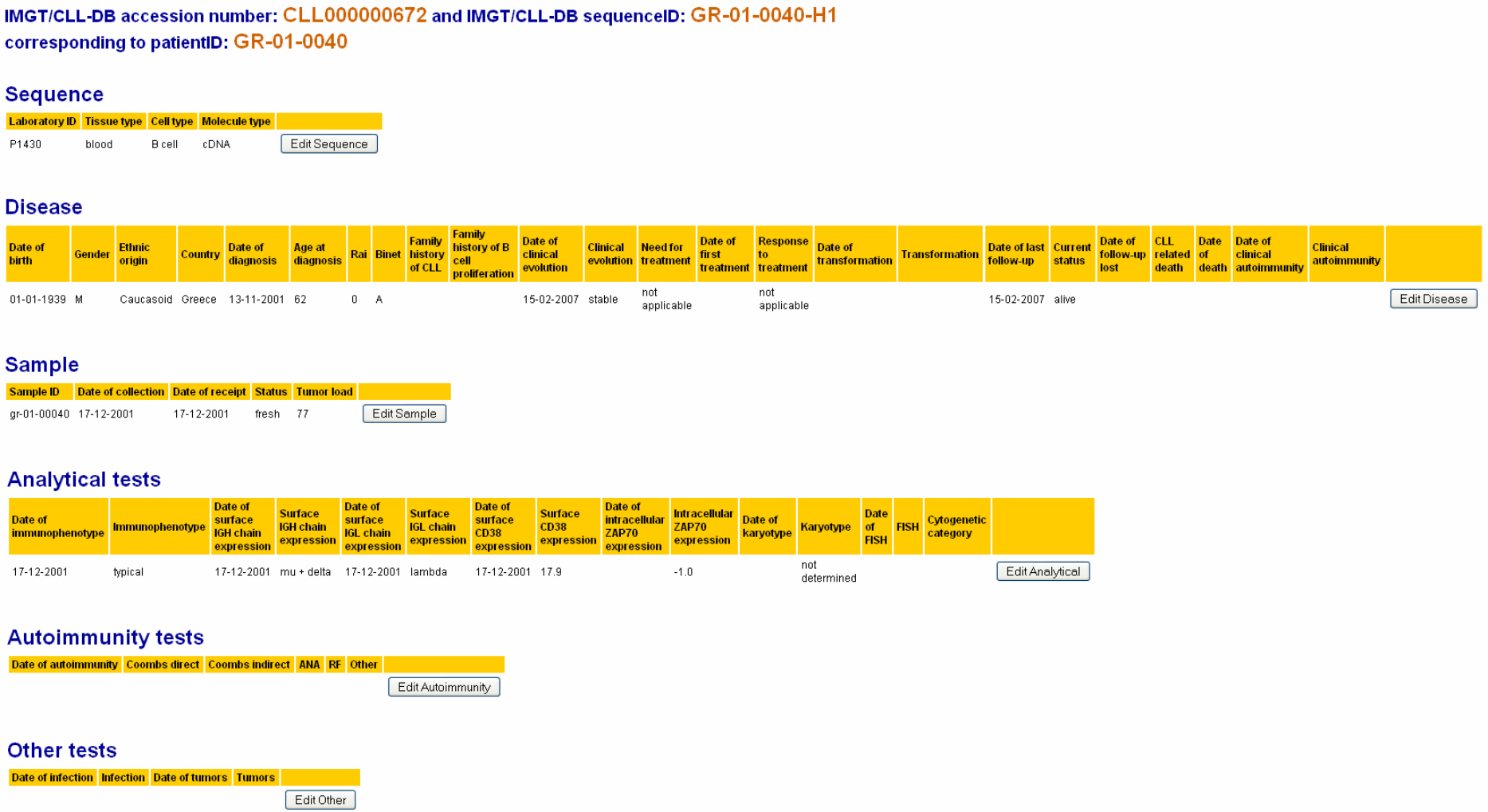


Figure 4.2: Les données de la séquence sélectionnée par le gestionnaire de la base sont affichées selon 6 catégories (Sequence, Disease, Sample, Analytical tests, Autoimmunity tests et other tests). Chaque catégorie peut être éditée**.**

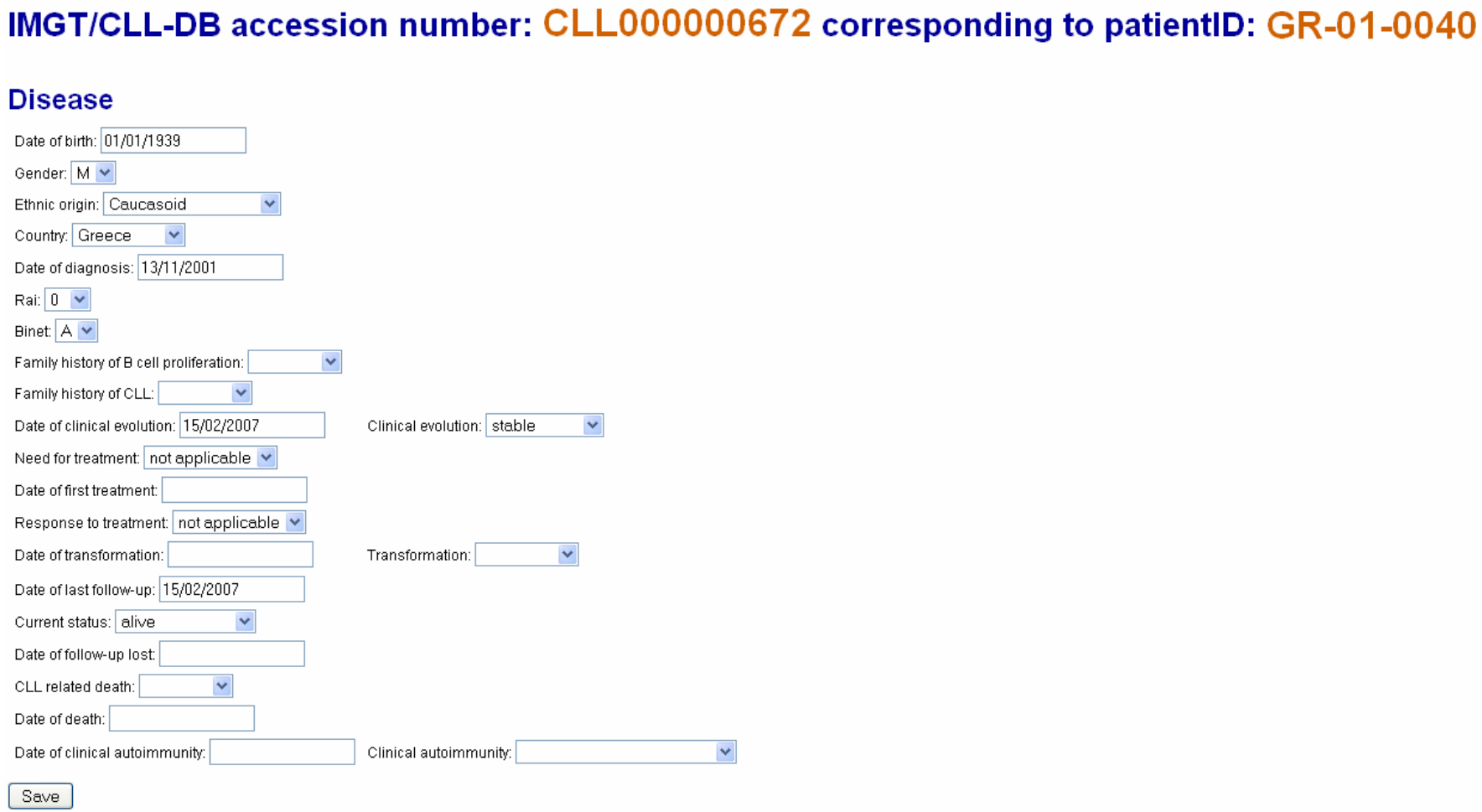


Figure 4.3: Edition des informations de la catégorie 'Disease' pour la séquence CLL000000672. Les données sont éditées par défaut avec les valeurs précédemment enregistrées. Pour chaque donnée soumise aux contraintes du vocabulaire contrôlé, le choix des valeurs s'affiche dans la fenêtre sélectionnée.

4.3 Interface utilisateur

La base de données IMGT/CLL-DB est accessible via une interface web. Son accessibilité est restreinte aux seuls partenaires du projet. L'interface permet l'interrogation de la base de façon simple et conviviale selon différents niveaux de précision (recherche simple et recherche avancée) et offre deux formes (vues) de présentation des données ('IMGT/CLL-DB Search Results' et le 'IMGT/CLL-DB Patient card'). Enfin, elle intègre des outils de traitement des séquences (IMGT/V-QUEST Synthesis view) et de récupération des données sous différents formats (FASTA et Excel).

4.3.1 Implémentation et architecture de l'interface Web

Toutes les applications Web au sein d'IMGT® étaient jusqu'à ce jour développées sous une interface de programmation (Application Programming Interface ou API) qui correspond à un ensemble de fonctions, procédures et de classes mises à disposition des programmes informatiques sous forme de bibliothèques logicielles, développées au sein d'IMGT® et antérieures aux standards actuels. Lors de la mise en œuvre du système d'information et suite à une veille technologique, nous avons implémenté l'interface utilisateur du système selon une architecture standard utilisant des technologies récentes et performantes dans le but d'offrir une plus grande souplesse et une simplification des développements et de la maintenance.

L'interface utilisateur de la base de données IMGT/CLL-DB a été implémentée selon l'architecture Modèle-Vue-Contrôleur (MVC) (qui est un 'design pattern') c'est-à-dire une architecture qui organise l'interface Homme-Machine d'une application logicielle. Elle offre un cadre pour structurer une application Web. L'architecture MVC se divise en trois entités distinctes ayant chacune un rôle précis dans l'interface:

- Le Modèle, contient les données manipulées par le programme. Dans le cas particulier d'une base de données, c'est le Modèle qui la contient. Il assure la gestion des données de la base et garantit leur intégrité. Le Modèle permet l'interaction entre la base de données et le traitement des données. Il offre des méthodes pour mettre à jour les données (insertion, suppression, changement de valeur) et des méthodes pour récupérer ces données. Les résultats renvoyés par le Modèle sont dénués de toute présentation.

- La Vue correspond à l'interface avec laquelle l'utilisateur interagit. Sa première tâche est d'afficher les données qu'elle a récupérées auprès du Modèle. Sa seconde tâche est de répertorier l'ensemble des actions de l'utilisateur (clic de souris, sélection d'une entrée, boutons, etc…) et de les renvoyer au contrôleur. La Vue n'effectue aucun traitement, elle se contente d'afficher les résultats des traitements effectués par le Modèle. La Vue peut aussi offrir la possibilité à l'utilisateur de changer de vue pour, par exemple, présenter les mêmes données sous d'autres formats. Elle peut être conçue en html ou tout autre « langage » de présentation tel que les JSP (Java Server Page; technologie basée sur Java qui permet aux développeurs de générer dynamiquement des pages web).
- Le Contrôleur est chargé de synchroniser le Modèle et la Vue. Il reçoit tous les événements de l'utilisateur et enclenche les actions à effectuer. Si une action nécessite un changement des données, le Contrôleur demande la modification des données au Modèle et ensuite avertit la Vue que les données ont changé afin que celle-ci se mette à jour. Certains événements de l'utilisateur ne concernent pas les données mais la Vue. Dans ce cas, le Contrôleur demande à la Vue de se modifier. Le Contrôleur n'effectue aucun traitement, ne modifie aucune donnée. Il analyse la requête du client et se contente d'appeler le Modèle adéquat et de renvoyer la Vue correspondante à la demande.

L'interface Web du système d'information développé avec une architecture MVC est subdivisée sous forme de composants. Cela permet de développer de manière indépendante chacune de ses parties pour ensuite les associer. Chaque partie étant un projet indépendant, il est plus aisé de modifier, d'ajouter ou de supprimer une fonctionnalité dans un projet sans pour autant modifier les autres. Quatre couches de composants ont été définies:

- La partie Présentation se concentre sur la manière d'afficher les informations pour l'utilisateur et sur la manière dont il récupère ces informations.
- La partie Métier se concentre sur la partie fonctionnelle de l'application sur ce dont l'utilisateur a besoin pour travailler, ce qu'il manipule dans le cadre de son travail. Dans notre cas, nous manipulons des données de séquences, des données sur les patients etc…

- La partie Persistance se concentre sur la manière d'enregistrer, de mettre à jour ou de supprimer les informations dans une source de données en l'occurrence la base de données IMGT/CLL-DB.
- La partie Service se concentre sur la communication d'une couche à une autre, par exemple le transfert d'informations de la couche de persistance vers la couche de présentation.

Pour chaque couche, il existe des boîtes à outils (ensembles de bibliothèques, d'outils et de conventions) appelés dans le jargon informatique des Frameworks, permettant le développement d'applications Web.

Le framework Struts est un standard pour la couche de présentation, basé sur les principes d'application Web Java J2EE. Il respecte le modèle MVC et la séparation du code de présentation, du code de contrôle, du code métier et du code de persistance. Le framework Hibernate (open source) gère la persistance des objets en base de données relationnelle. C'est un outil de mapping objet/relationnel (ORM) qui permet de faire le lien entre la représentation objet des données et sa représentation relationnelle. Hibernate s'occupe donc du transfert des classes Java dans les tables de la base de données, il permet également de requêter les données et propose des moyens de les récupérer. En clair il a pour but de gérer l'accès aux données contenues dans une base de données.
La figure 4.4 est une représentation schématique de l'architecture de l'interface Web mise en place pour la base de données IMGT/CLL-DB. Le Modèle est représenté par des objets créés à partir d'une connexion à une source de données; dans notre application, la source de données est la base de données IMGT/CLL-DB. Il décrit ou contient les données manipulées par l'application. Il assure la gestion de ces données et garantit leur intégrité.
Dans notre application le Modèle contient la couche dite de persistance implémentée via le framework Hibernate qui prend en charge l'ensemble des traitements concernant la gestion des données de la base. Chaque table de la base de données est ainsi mappée telle quelle ou divisée en différentes classes Java (13 classes). Toute la gestion des données se fait alors par l'intermédiaire de ces classes. La couche de persistance gère également la gestion des connexions à la base (utilisation du pool de connexion Cp30).

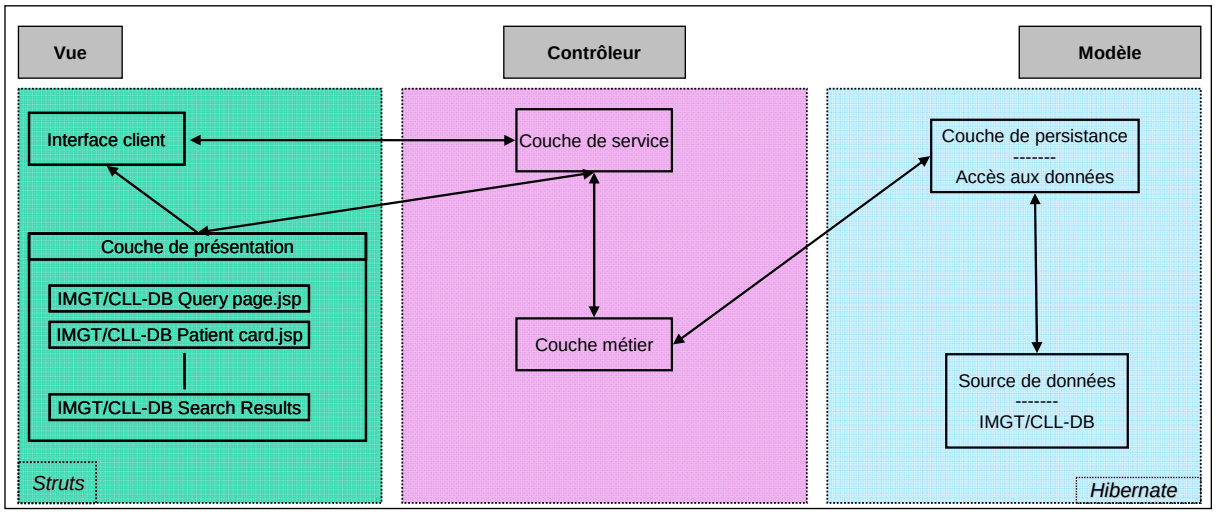


Figure 4.4: Architecture du système d'information sous un modèle MVC. Il divise l'interface en un Modèle (modèle de données), une Vue (présentation, interface utilisateur) et un Contrôleur (logique de contrôle, gestion des événements, synchronisation). Dans notre application le 'Modèle' contient la couche dite de persistance implémentée via le framework Hibernate qui prend en charge l'ensemble des traitements concernant la gestion des données de la base. Le framework Struts est un standard pour la couche de présentation, basé sur les principes d'application Web Java J2EE.

La Vue correspond à l'interface avec laquelle l'utilisateur interagit. Sa première tâche est de présenter les résultats renvoyés par le Modèle. Sa seconde tâche est de recevoir toutes les actions de l'utilisateur (requête, sélection d'une entrée, lancement d'une application interactive (IMGT:V-QUEST synthesis view…). Ces différents événements sont envoyés au Contrôleur. La Vue n'effectue aucun traitement, elle se contente d'afficher les résultats des traitements effectués par le Modèle. Les résultats des traitements sont conçus dans cette application uniquement avec des JSP.

Le Contrôleur, administre tous les composants de l'application. C'est le Contrôleur qui distribue le travail à la Vue et au Modèle. La couche dite de service gère l'ensemble des événements et les relations et le transfert d'informations entre les autres couches du système, la couche de présentation, la couche métier et la couche de persistance. Il analyse la requête du client et se contente d'appeler le Modèle adéquat et de renvoyer la Vue correspondante à la demande. La couche métier constitue l'ensemble des méthodes Java qui manipulent les données à la suite de leur récupération via la couche de persistance, telle que l'analyse des séquences par le programme IMGT/V-QUEST Synthesis view.

4.3.2 IMGT/CLL-DB Query page

L'interface d'interrogation 'IMGT/CLL-DB Query page' (Figure 4.5) est une interface Web qui permet d'interroger la base de données IMGT/CLL-DB de façon simple et précise. L'interface propose deux types d'interrogations: une recherche par identifiant ou numéro d'accès et une recherche par critères multiples.

Search by ID

La recherche 'Search by ID' (Figure 4.5A) permet à l'utilisateur de faire des recherches (i) le numéro d'accession, (ii) l'identifiant de la séquence, (ii) l'identifiant du patient ou (iv) le numéro d'accession de IMGT/LIGM-DB [244]. L'interface offre la possibilité de construire une requête constituée de plusieurs numéros d'accès ou de plusieurs identifiants séparés par des 'virgules'. Dans ce cas précis, les numéros d'accès ou les identifiants utilisés doivent être obligatoirement de même type, uniquement des numéros d'accès IMGT/CLL-DB ou uniquement des identifiants Sequence ID, etc. L'interface 'Search by ID' accepte l'utilisation d'un 'caractère de remplacement' ('wildcard') dans la formulation de la requête. Ce caractère de remplacement est symbolisé par le sigle '*' et permet de faire des recherches à un niveau plus général. Son utilisation offre la possibilité de requêter la base non plus sur une séquence ou un patient précis mais sur toutes les séquences provenant d'un laboratoire ou d'un pays en particulier. Par exemple la requête 'GR-01*' recherche toutes les séquences provenant d'un laboratoire situé en Grèce et défini par le code 01 ou de façon encore plus générale 'GR*' pour toutes les séquences provenant de tous les laboratoires situés en Grèce. Enfin l'interface de recherche 'Search by ID' permet d'obtenir toutes les séquences stockées dans la base de données en utilisant le seul caractère de remplacement '*'.

L'ensemble des numéros d'accès ou identifiants acceptés par l'interface de recherche 'Search by ID' est résumé dans le tableau 4.1.

Tableau 4.1: Ensemble des numéros d'accès ou identifiants acceptés par l'outil de recherche 'Search by ID' et exemple d'utilisation.

Numéro d'accès ou Identifiant accepté	Format de requête multiple	Caractère de remplacement
IMGT/CLL-DB accession number	CLL000000001,CLL000000002,CLL000000003...	Non
Sequence ID	NY-01-0001-H1,NY-01-0002-H1,NY-01-0003-H1…	Oui. Ex: NY-01-0001*
Patient ID	NY-01-0001,NY-01-0002,NY-01-0003…	Oui. Ex: NY*
IMGT/LIGM-DB accession number	DQ100880, DQ100832, DQ100903…	Non

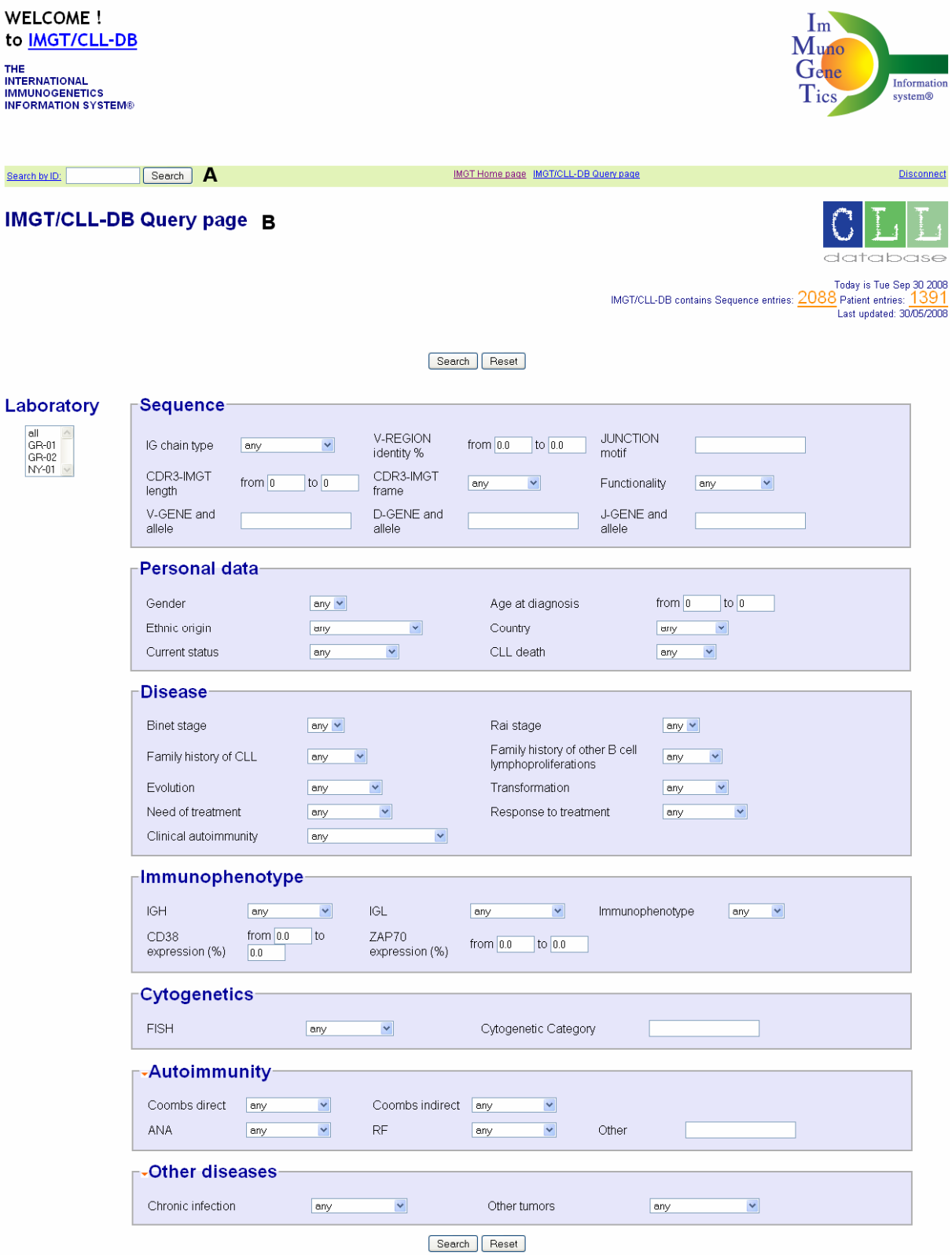


Figure 4.5: L'interface de recherche de la base IMGT/CLL-DB. **A**: Interface simple qui permet la recherche par les numéros d'accès ou par les identifiants de patients ou de séquences. **B**: Interface avancée qui permet des recherches complexes par l'intermédiaire de 39 critères, divisés en 8 catégories (Laboratory, Sequence, Personal data, Disease, Immunophenotype, Cytogenetics, Autoimmunity et Other Diseases).

IMGT/CLL-DB Search

L'interface IMGT/CLL-DB Query page (Figure 4.5 B) est également un outil de recherche avancé qui offre la possibilité à l'utilisateur de constituer une requête par l'utilisation de critères multiples, soit:

39 critères répartis en 8 catégories: ‘Laboratory’ (1 critère), ‘Sequence’ (9 critères), ‘Personal data’ (6 critères), ‘Disease (9 critères), ‘Immunophenotype’ (5 critères), ‘Cytogenetics’ (2 critères), ‘Autoimmunity’ (5 critères), ‘Other diseases’ (2 critères). A chaque catégorie correspond un thème spécifique qui regroupe un ou plusieurs critères de recherche qu'il est possible ou non de renseigner selon les souhaits de l’utilisateur. Tous les critères de recherche peuvent être combinés pour obtenir la requête souhaitée la plus précise possible.

Laboratory

La catégorie ‘Laboratory’ (Figure 4.5B) permet de rechercher des séquences selon le laboratoire propriétaire de la séquence. Il est possible de choisir pour cette catégorie un ou plusieurs laboratoires comme critère de recherche. Par défaut, la recherche se fait sur l'ensemble des laboratoires.

Sequence

La catégorie ‘Sequence’ (Figure 4.5B) regroupe les critères de recherche relatifs aux principales caractéristiques des séquences d’IG. Elle offre la possibilité à l’utilisateur de choisir le type de chaîne (chaîne lourde (IGH) ou légère kappa (IGK) ou lamdba (IGL)), le pourcentage d’identité de la V-REGION comparé aux séquences germline les plus proches. L’utilisateur a également la possibilité de rechercher des séquences contenant un motif caractéristique de la jonction exprimé en acides aminés, de définir la taille du CDR3-IMGT, le cadre de lecture du CDR3, si le CDR3 est dans un cadre de lecture ouvert (in-frame) ou non (out-of-frame), la fonctionnalité de la séquence (‘productive’ ou ‘unproductive’), et enfin le nom du gène V ou de l’allèle ou du sous-groupe (ex: IGHV1-69, IHV1-69*01, IGHV1), le nom du gène D ou de l’allèle ou du groupe et le nom du gène J ou de l’allèle ou du groupe en accord avec les règles de nomenclature IMGT.

Personal Data

La catégorie ‘Personal Data’ (Figure 4.5B) regroupe les critères de recherche relatifs au patient, sexe, l’âge, date du diagnostic, origine ethnique, pays de résidence, statut actuel et si décès, lien avec la LLC ou non.

Disease

La catégorie ‘Disease’ (Figure 4.5B) regroupe les critères de recherche relatifs à la description du stade de la maladie et de son évolution. L’utilisateur peut interroger la base sur le stade clinique du patient selon la classification Binet, ou la classification Rai. L’utilisateur peut également rechercher les patients sur leur histoire familiale (de la LLC, ou d’autres syndromes lymphoprolifératifs B), sur l’évolution de la maladie (progressive, stable, régressive), ou sur sa transformation (Richter, PLL, etc.).

L’utilisateur peut également rechercher les patients ayant besoin impérativement d’un traitement, la réponse au traitement (positive ou négative), les patients qui ont des complications autoimmunes (telles que ‘anemia’, ‘Thrombocytopenia’ etc…).

Immunophenotype

La catégorie ‘Immunophenotype’ (Figure 4.5B) regroupe les critères de recherche relatifs à l'immunophénotypage. L'utilisateur a la possibilité de faire des requêtes selon l’isotypes des chaînes lourdes et des chaînes légères exprimés à la surface des cellules B, l’immunophénotype des cellules B des patients (typique c’est-à-dire caractéristique d'un patient atteint de LLC ou atypiques). Enfin l'utilisateur peut définir sa recherche selon la valeur des facteurs de pronostic: CD38 et ZAP 70.

Cytogenetics

La catégorie ‘Cytogenetics’ (Figure 4.5B) regroupe les critères de recherche relatifs aux données cytogénétiques des patients. L’utilisateur peut rechercher des patients par le résultat des analyses ‘fluorescence in situ hybridation’ (‘FISH’) qui permet de retrouver des anomalies chromosomiques chez les patients et par leur catégorie cytogénétique, qui constitue un code décrivant le caryotype des patients.

Autoimmunity

La catégorie ‘Autoimmunity’ (Figure 4.5B) regroupe les critères relatifs aux manifestations de phénomènes autoimmuns chez les patients. L'utilisateur a la possibilité d'interroger la base de données selon les résultats d'une série de tests classiques pour détecter les phénomènes d'autoimmunité. 5 critères de recherche sont disponibles: les tests de Coombs, ‘Coombs direct’ et ‘Coombs indirect’, les tests de recherche d’autoanticorps avec le dépistage des anticorps antinucléaires (‘ANA’) et des facteurs rhumatoïdes (‘RF’). Le dernier critère de

recherche ‘Other’ est un champ libre, dans lequel l’utilisateur est libre de spécifier des informations liées à la recherche de manifestations d’autoimmunité.

Other diseases

La catégorie ‘Other diseases’ (Figure 4.5B) regroupe les critères de recherche relatifs aux complications infectieuses ou aux cancers associés qui peuvent survenir au cours de l’évolution de la maladie.

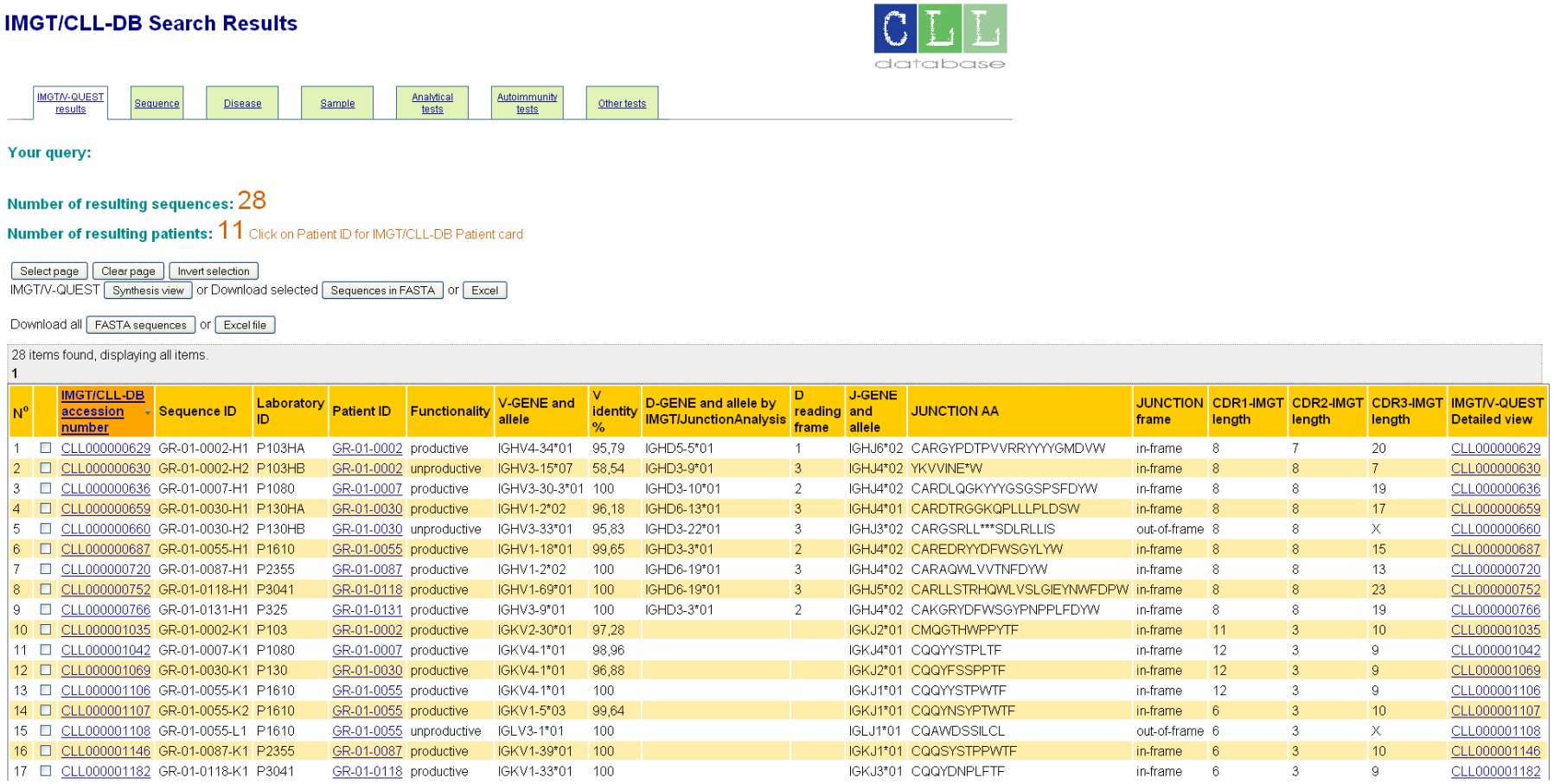
4.3.3 IMGT/CLL-DB results

Deux présentations différentes de visualisation des résultats sont proposées par l’interface Web de la base de données:

- la page ‘Search Results’ présente les résultats d’une recherche de multiples séquences appartenant à de multiples patients. Elle est présentée sous forme de liste dans un tableau constituée de 5 onglets. Chaque onglet regroupe les données stockées dans la base par thème. Cette vue permet alors une comparaison aisée des données entre elles.
- le ‘Patient card’ présente les informations pour un seul patient.

4.3.3.1 IMGT/CLL-DB Search Results

La sortie ‘Search Results’ (Figure 4.6) affiche les données de plusieurs séquences. Les résultats sont présentés sous forme de tableaux composés de 7 onglets, chacun présentant les données regroupées par thème: IMGT/V-QUEST results, Sequence, Disease, Sample, Analytical tests, Autoimmunity tests, et Other tests. Le tableau est présenté sous forme de liste paginée et affiche jusqu'à 50 entrées par page; chaque ligne correspondant à une entrée (soit une séquence, soit un patient selon l'onglet visualisé) et chaque colonne à un résultat donné.



IMGT/CLL-DB Search Results

IMGT/V-QUEST results | Sequence | Disease | Sample | Analytical tests | Autoimmunity tests | Other tests

Your query:

Number of resulting sequences: 28

Number of resulting patients: 11 Click on Patient ID for IMGT/CLL-DB Patient card

Select page | Clear page | Invert selection

IMGT/V-QUEST Synthesis view or Download selected Sequences in FASTA or Excel

Download all FASTA sequences or Excel file

28 items found, displaying all items.

1

N°		IMGT/CLL-DB accession number	Sequence ID	Laboratory ID	Patient ID	Functionality	V-GENE and allele	V identity %	D-GENE and allele by IMGT/JunctionAnalysis	D reading frame	J-GENE and allele	JUNCTION AA	JUNCTION frame	CDR1-IMGT length	CDR2-IMGT length	CDR3-IMGT length	IMGT/V-QUEST Detailed view
1	☐	CLL000000629	GR-01-0002-H1	P103HA	GR-01-0002	productive	IGHV4-34*01	95,79	IGHD5-5*01	1	IGHJ6*02	CARGYPDTPVVRRYYYYGMDVW	in-frame	8	7	20	CLL000000629
2	☐	CLL000000630	GR-01-0002-H2	P103HB	GR-01-0002	unproductive	IGHV3-15*07	58,54	IGHD3-9*01	3	IGHJ4*02	YKVVINE*W	in-frame	8	8	7	CLL000000630
3	☐	CLL000000636	GR-01-0007-H1	P1080	GR-01-0007	productive	IGHV3-30-3*01	100	IGHD3-10*01	2	IGHJ4*02	CARDLQGKYYYGSGSPSFDYW	in-frame	8	8	19	CLL000000636
4	☐	CLL000000659	GR-01-0030-H1	P130HA	GR-01-0030	productive	IGHV1-2*02	96,18	IGHD6-13*01	3	IGHJ4*01	CARDTRGGKQPLLLPLDSW	in-frame	8	8	17	CLL000000659
5	☐	CLL000000660	GR-01-0030-H2	P130HB	GR-01-0030	unproductive	IGHV3-33*01	95,83	IGHD3-22*01	3	IGHJ3*02	CARGSRLL***SDLRLLIS	out-of-frame	8	8	X	CLL000000660
6	☐	CLL000000687	GR-01-0055-H1	P1610	GR-01-0055	productive	IGHV1-18*01	99,65	IGHD3-3*01	2	IGHJ4*02	CAREDRYYDFWSGYLYW	in-frame	8	8	15	CLL000000687
7	☐	CLL000000720	GR-01-0087-H1	P2355	GR-01-0087	productive	IGHV1-2*02	100	IGHD6-19*01	3	IGHJ4*02	CARAQWLVVTNFDYW	in-frame	8	8	13	CLL000000720
8	☐	CLL000000752	GR-01-0118-H1	P3041	GR-01-0118	productive	IGHV1-69*01	100	IGHD6-19*01	3	IGHJ5*02	CARLLSTRHQWLVSLGIEYNWFDPW	in-frame	8	8	23	CLL000000752
9	☐	CLL000000766	GR-01-0131-H1	P325	GR-01-0131	productive	IGHV3-9*01	100	IGHD3-3*01	2	IGHJ4*02	CAKGRYDFWSGYPNPPLFDYW	in-frame	8	8	19	CLL000000766
10	☐	CLL000001035	GR-01-0002-K1	P103	GR-01-0002	productive	IGKV2-30*01	97,28			IGKJ2*01	CMQGTHWPPYTF	in-frame	11	3	10	CLL000001035
11	☐	CLL000001042	GR-01-0007-K1	P1080	GR-01-0007	productive	IGKV4-1*01	98,96			IGKJ4*01	CQQYYSTPLTF	in-frame	12	3	9	CLL000001042
12	☐	CLL000001069	GR-01-0030-K1	P130	GR-01-0030	productive	IGKV4-1*01	96,88			IGKJ2*01	CQQYFSSPPTF	in-frame	12	3	9	CLL000001069
13	☐	CLL000001106	GR-01-0055-K1	P1610	GR-01-0055	productive	IGKV4-1*01	100			IGKJ1*01	CQQYYSTPWTF	in-frame	12	3	9	CLL000001106
14	☐	CLL000001107	GR-01-0055-K2	P1610	GR-01-0055	productive	IGKV1-5*03	99,64			IGKJ1*01	CQQYNSYPTWTF	in-frame	6	3	10	CLL000001107
15	☐	CLL000001108	GR-01-0055-L1	P1610	GR-01-0055	unproductive	IGLV3-1*01	100			IGLJ1*01	CQAWDSSILCL	out-of-frame	6	3	X	CLL000001108
16	☐	CLL000001146	GR-01-0087-K1	P2355	GR-01-0087	productive	IGKV1-39*01	100			IGKJ1*01	CQQSYSTPPWTF	in-frame	6	3	10	CLL000001146
17	☐	CLL000001182	GR-01-0118-K1	P3041	GR-01-0118	productive	IGKV1-33*01	100			IGKJ3*01	CQQYDNPLFTF	in-frame	6	3	9	CLL000001182

Figure 4.6: 'IMGT/CLL-DB Search Results'. Le 'Search Results' affiche les entrées de IMGT/CLL-DB, sous forme d'un tableau composé de 7 onglets. Chaque onglet présente des données regroupées par thème, avec 'IMGT/V-QUEST results' (résultats de l'analyse de IMGT/V-QUEST), 'Sequence' (description de la séquence), 'Disease' (description du patient et de sa pathologie), 'Sample' (description de l'échantillon), 'Analytical tests' (description de l'immunophénotype et de facteur de pronostic), 'Autoimmunity tests' (test pour la recherche de signes autoimmuns) et 'Other tests' (complications infectieuses et cancéreuses). Pour chaque onglet, la requête et le nombre de résultats (nombre de séquences et de patients) sont affichés. L'onglet 'IMGT/V-QUEST results' offre la possibilité d'analyser jusqu'à 50 séquences avec 'IMGT/V-QUEST Synthesis view', de récupérer les séquences en format FASTA et toutes les données sous format Excel. Pour chaque entrée, le 'IMGT/V-QUEST results' montre le numéro d'accès de IMGT/CLL-DB, les identifiants avec la Sequence ID, le Laboratory ID, le Patient ID, la fonctionnalité, le nom du gène V et allèle associé au pourcentage d'identité, le nom du gène D et allèle déterminés par IMGT/JunctionAnalysis, le cadre de lecture du D, le nom du gène J et allèle, la jonction en AA, le cadre de lecture de la jonction, la taille des 3 CDR-IMGT et un lien vers la fiche complète des résultats de IMGT/V-QUEST.

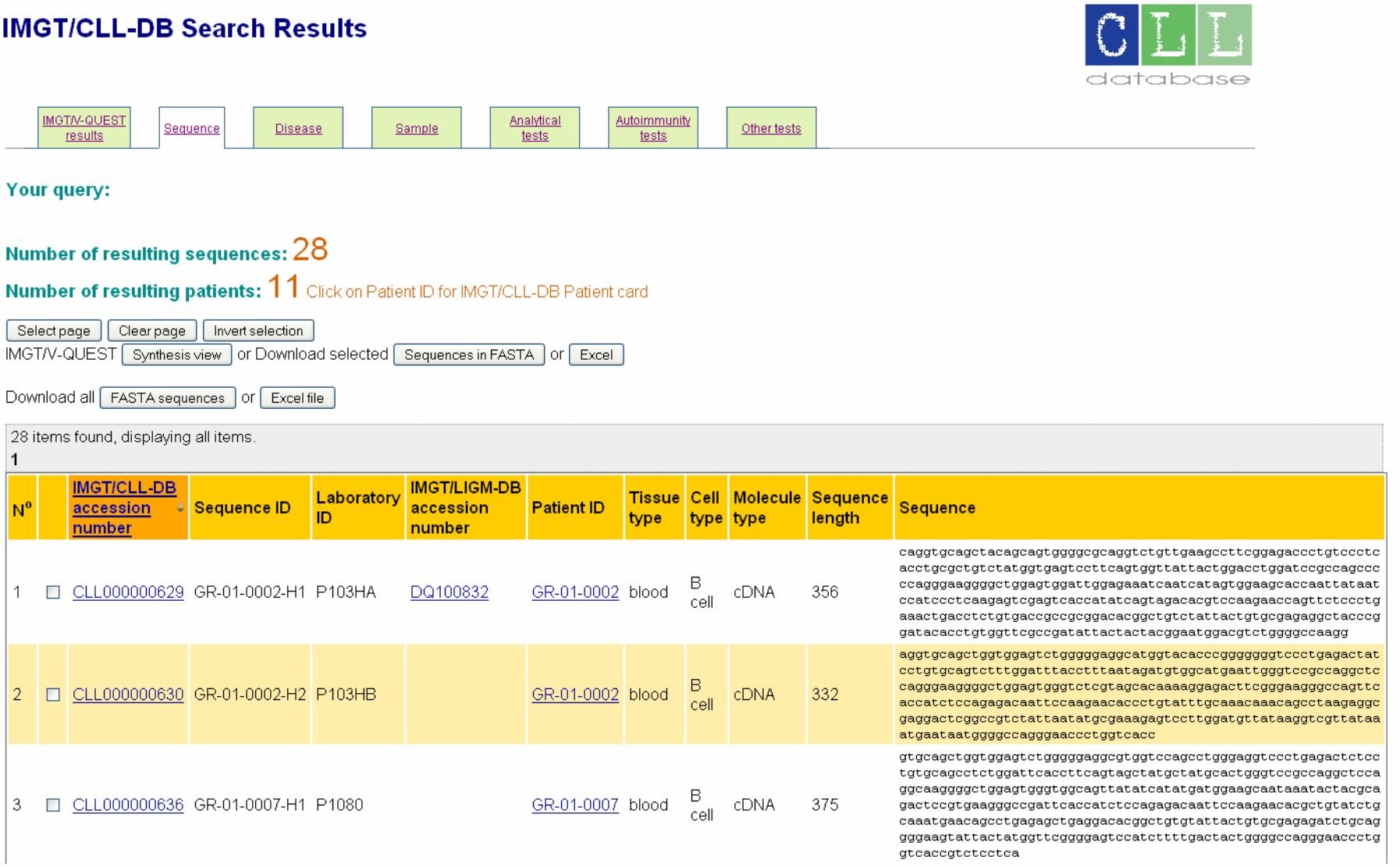
IMGT/V-QUEST Search results

L'onglet ‘IMGT/V-QUEST results’ (Figure 4.6) présente pour chaque séquence les principaux résultats de l'analyse par IMGT/V-QUEST. Chaque séquence est identifiée par son numéro d'accès (IMGT/CLL-DB accession number), par son identifiant (Sequence ID), par l'identifiant du laboratoire propriétaire de la séquence (Laboratory ID), et par l'identifiant du patient (Patient ID) qui est également un lien vers la sortie ‘Patient card’.

Les principaux résultats de l'analyse par IMGT/V-QUEST sont aux nombre de 16, avec l'évaluation de la fonctionnalité: ‘productive’ ou ‘unproductive’ (Functionality), le nom du gène V et allèle germline (V-GENE and allele) le plus proche associé au pourcentage d'identité (V identity %), le nom du gène D et allèle germline (D-GENE and allele), le cadre de lecture du D (D reading frame), le nom du gène J et allèle germline (J-GENE and allele), la jonction en AA (JUNCTION AA), le cadre de lecture de la jonction, 'in-frame' ou 'out-of-frame' (JUNCTION frame), la taille des trois CDR-IMGT, le CRD1-IMGT (CDR1-IMGT length), le CDR2-IMGT (CDR2-IMGT length), le CDR3-IMGT (CDR3-IMGT length) et un lien vers les résultats complets et détaillés de l'analyse de la séquence par IMGT/V-QUEST (cf. paragraphe 3.5.2.1 Detailed view).

Sequence

L'onglet ‘Sequence’ (Figure 4.7) présente pour chaque séquence 10 types de résultats. Chaque séquence est définie par ses identifiants uniques (IMGT/CLL-DB accession number et Sequence ID), par l'identifiant du laboratoire propriétaire de la séquence (Laboratory ID), et si la séquence est publique, par le numéro d’accès IMGT/LIGM-DB [244]. Cet identifiant est un lien direct de l'entrée dans la base IMGT/LIGM-DB [244]. De plus, à chaque séquence est associé à l’identifiant du patient (Patient ID), le type de tissu (blood, bone marrow...) (Tissu type), le type de cellule (B cell) (Cell type), le type de molécule gDNA, ou cDNA (Molecule type), la taille de la séquence (Sequence length), et enfin la séquence en nucléotides (Sequence).



IMGT/CLL-DB Search Results

IMGT/V-QUEST results | Sequence | Disease | Sample | Analytical tests | Autoimmunity tests | Other tests

Your query:

Number of resulting sequences: 28

Number of resulting patients: 11 Click on Patient ID for IMGT/CLL-DB Patient card

Select page | Clear page | Invert selection

IMGT/V-QUEST Synthesis view or Download selected Sequences in FASTA or Excel

Download all FASTA sequences or Excel file

28 items found, displaying all items.

1

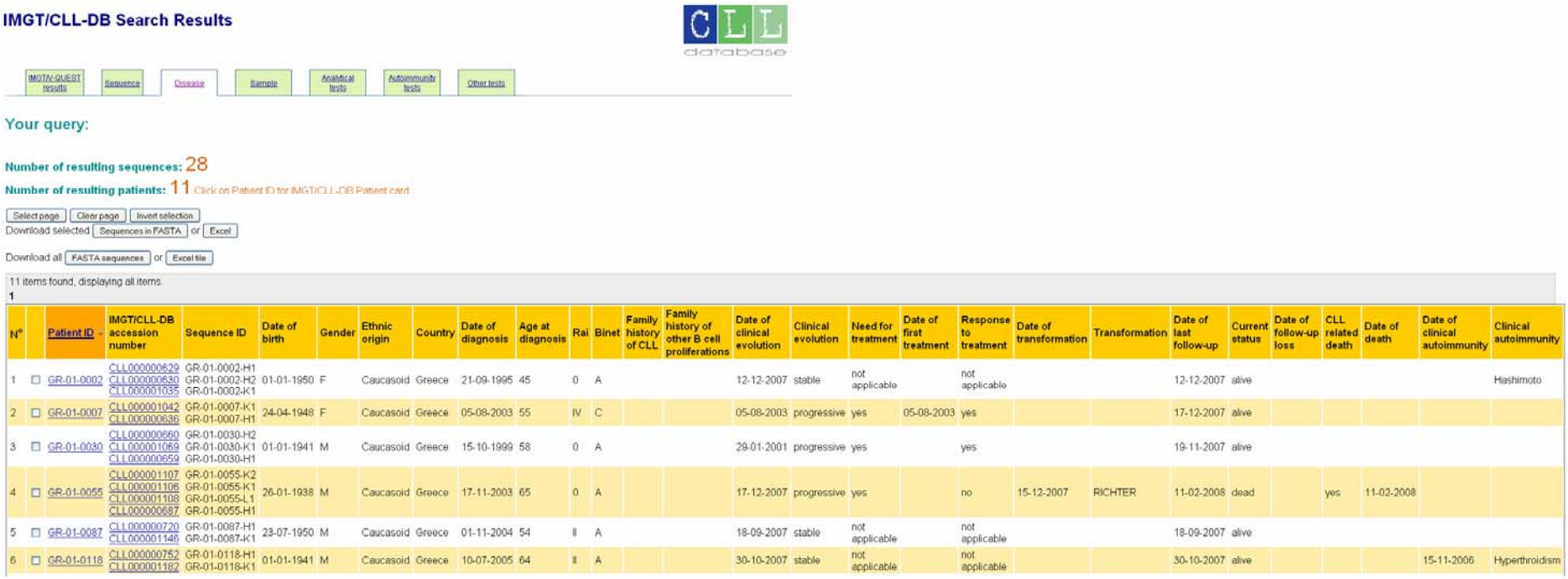
N°		IMGT/CLL-DB accession number	Sequence ID	Laboratory ID	IMGT/LIGM-DB accession number	Patient ID	Tissue type	Cell type	Molecule type	Sequence length	Sequence
1	☐	CLL000000629	GR-01-0002-H1	P103HA	DQ100832	GR-01-0002	blood	B cell	cDNA	356	caggtgcagctacagcagtggggcgcaggtctgttgaagccttcggagaccctgtccctcacctgcgctgtctatggtgagtccttcagtggttattactggacctggatccgccagccccagggaaggggctggagtggattggagaaatcaatcatagtggaagcaccaattataatccatccctcaagagtcgagtcaccatatcagtagacacgtccaagaaccagtttctccctgaaactgacctctgtgaccgccgcggacacggctgtctattactgtgcgagaggctacccggatacacctgtggttcgccgatattactactacggaatggacgtctggggccaagg
2	☐	CLL000000630	GR-01-0002-H2	P103HB		GR-01-0002	blood	B cell	cDNA	332	aggtgcagctggtggagtctgggggaggcatggtacacccgggggggtccctgagactatcctgtgcagtctttggatttacctttaatagatgtggcatgaattgggtccgccaggctccagggaaggggctggagtggtctcgtagcacaaaaggagacttcgggaagggccagttccaccatctccagagacaattccaagaacaccctgtatttgcaaacaaacagcctaagaggcgaggactcggccgtctattaatatgcgaaagagtccttggatgttataaggtcgttataaatgaataatggggccagggaaccctggtcacc
3	☐	CLL000000636	GR-01-0007-H1	P1080		GR-01-0007	blood	B cell	cDNA	375	gtgcagctggtggagtctgggggaggcgtggtccagcctgggaggtccctgagactctcctgtgcagcctctggattcaccttcagtagctatgctatgcactgggtccgccaggctccaggcaaggggctggagtgggtggcagttatatcatatgatggaagcaataaatactacgcagactccgtgaagggccgattcaccatctccagagacaattccaagaacacgctgtatctgcaaatgaacagcctgagagctgaggacacggctgtgtattactgtgcgagagatctgcagggaagtattactatggttcggggagtccatcttttgactactggggccagggaaccctggtcaccgtctcctca

Figure 4.7: L'onglet 'Sequence' de la page 'IMGT/CLL-DB Search Results', présente pour chaque entrée, le numéro d'accès IMGT/CLL-DB, le Sequence ID, le Laboratory ID, le numéro d'accès IMGT/LIGMDB, le Patient ID, le type tissulaire, le type cellulaire, le type moléculaire, la taille de la séquence et la séquence.

Disease

L'onglet ‘Disease’ (Figure 4.8) présente pour chaque patient 27 résultats décrivant le patient, sa maladie et son évolution.

Pour chaque patient (ligne) est défini l'identifiant du patient (Patient ID) qui se trouve être également un lien vers le ‘Patient card’, le numéro d'accés (IMGT/CLL-DB accession number) et l'identifiant (Sequence ID) des séquences associées au patient. De plus cet onglet détaille les données en rapport aux patients et à sa pathologie avec: (i) la date de naissance (Date of birth), (ii) le sexe (Gender), (iii) l'origine ethnique (Ethnic origin), (iv) le pays de résidence (Country), (v) la date de diagnostic de la maladie (Date of diagnosis), (vi) l’âge du patient au diagnostic (Age at diagnosis), (vii et viii) le stade de la maladie selon les deux classifications Rai et Binet (Rai, Binet), (viii et ix) l'histoire familiale de la LLC (Family history of CLL) et d’autres syndromes lymphoprolifératifs B (Family history of other B cell proliferations), (x) l’évolution de la LLC associée à une date (Date of clinical evolution, Clinical evolution), (xi) le besoin d'un traitement (Need for treatment), (xii) la date du premier traitement (Date of first treatment), (xii) la réponse au traitement (Response to treatment), (xiii) la date de complication vers une forme agressive de la maladie (Date of transformation), (xiv) le type de transformation (Transformation), (xv) la date de la dernière consultation du patient (Date of last follow-up), (xvi) le statut du patient (Current status), (xvii) la date de la perte du suivi du patient (Date of follow-up loss), (xviii) le lien ou non du décès avec la LLC (CLL related death), la date du décès (Date of death), (xix) la date d'apparition de complications autoimmunes (Date of clinical autoimmunity) et (xx) le type de complications autoimmunes (Clinical autoimmunity).



IMGT/CLL-DB Search Results

IMGT/V-QUEST results | Sequence | Disease | Sample | Analytical tests | Autoimmunity tests | Other tests

Your query:

Number of resulting sequences: 28

Number of resulting patients: 11 Click on Patient ID for IMGT/CLL-DB Patient card

Select page | Clear page | Invert selection

Download selected Sequences in FASTA or Excel

Download all FASTA sequences or Excel file

11 items found, displaying all items.

1

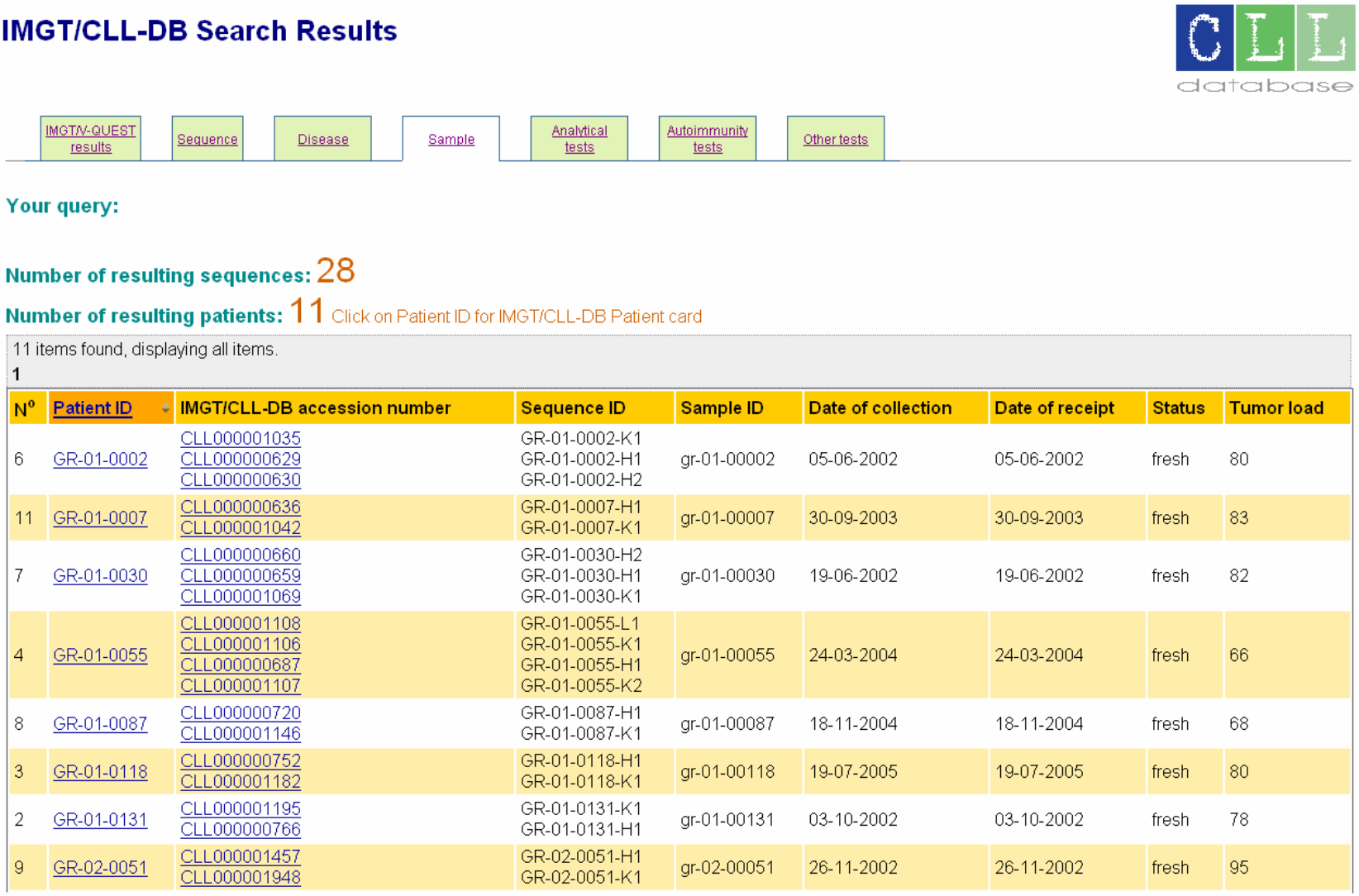
N°	Patient ID	IMGT/CLL-DB accession number	Sequence ID	Date of birth	Gender	Ethnic origin	Country	Date of diagnosis	Age at diagnosis	Rai	Binet	Family history of CLL	Family history of other B cell proliferations	Date of clinical evolution	Clinical evolution	Need for treatment	Date of first treatment	Response to treatment	Date of transformation	Transformation	Date of last follow-up	Current status	Date of follow-up loss	CLL related death	Date of death	Date of clinical autoimmunity	Clinical autoimmunity
1	GR-01-0002	CLL000000629 CLL000000630 CLL000001035	GR-01-0002-H1 GR-01-0002-H2 GR-01-0002-K1	01-01-1950	F	Caucasoid	Greece	21-09-1995	45	0	A			12-12-2007	stable	not applicable		not applicable			12-12-2007	alive					Hashimoto
2	GR-01-0007	CLL000001042 CLL000000636	GR-01-0007-K1 GR-01-0007-H1	24-04-1948	F	Caucasoid	Greece	05-08-2003	55	IV	C			05-08-2003	progressive	yes	05-08-2003	yes			17-12-2007	alive					
3	GR-01-0030	CLL000000660 CLL000001069 CLL000000659	GR-01-0030-H2 GR-01-0030-K1 GR-01-0030-H1	01-01-1941	M	Caucasoid	Greece	15-10-1999	58	0	A			29-01-2001	progressive	yes		yes			19-11-2007	alive					
4	GR-01-0055	CLL000001107 CLL000001106 CLL000001108 CLL000000687	GR-01-0055-K2 GR-01-0055-K1 GR-01-0055-L1 GR-01-0055-H1	26-01-1938	M	Caucasoid	Greece	17-11-2003	65	0	A			17-12-2007	progressive	yes		no	15-12-2007	RICHTER	11-02-2008	dead		yes	11-02-2008		
5	GR-01-0087	CLL000000720 CLL000001146	GR-01-0087-H1 GR-01-0087-K1	23-07-1950	M	Caucasoid	Greece	01-11-2004	54	II	A			18-09-2007	stable	not applicable		not applicable			18-09-2007	alive					
6	GR-01-0118	CLL000000752 CLL000001182	GR-01-0118-H1 GR-01-0118-K1	01-01-1941	M	Caucasoid	Greece	10-07-2005	64	II	A			30-10-2007	stable	not applicable		not applicable			30-10-2007	alive				15-11-2006	Hyperthroidism

Figure 4.8: L'onglet 'Disease' de la page IMGT/CLL-DB Search Results, présente pour chaque entrée, le Patient ID, le numéro d'accès IMGT/CLL-DB, le Sequence ID, la date de naissance, le sexe, l'origine ethnique, le pays de résidence du patient, la date du diagnostic, l'âge au moment du diagnostic, la classification selon Rai et selon Binet, l'histoire familiale de la LLC et des autres syndromes prolifératifs, l'évolution de la maladie et la date correspondant à cette évolution, le besoin d'un traitement, la date du premier traitement, la réponse au traitement, la transformation de la maladie et la date de cette transformation, la date du dernier suivi médical, le statut actuel du patient, la date de la perte du suivi du patient, le lien ou non du décès avec la LLC, la date du décès, la date d'apparition de complications autoimmunes et le type de complications autoimmunes.

Dans l'extrait représenté, nous avons 6 patients différents, pour lesquels 16 séquences d'IG sont disponibles. Les séquences en format FASTA peuvent être récupérées en sélectionnant les séquence (colonne 'N°'du tableau en partant de la gauche) et en cliquant sur le bouton 'Sequences in FASTA' au-dessus du tableau. Les données sous format Excel peuvent être récupéreées en sélectionnant les entrées (colonne 'N°' du tableau en partant de la gauche) et en cliquant sur le bouton 'Excel' au-dessus du tabbleau.

Sample

L'onglet ‘Sample’ (Figure 4.9) présente pour chaque patient 8 types de résultats: (i) L'identifiant du patient (Patient ID) qui se trouve être également un lien vers le ‘Patient card’, (ii) les numéros d'accès IMGT/CLL-DB (IMGT/CLL-DB accession number) et (iii) les identifiants des séquences associées au patient (Sequence ID), (iv) l'identifiant de l'échantillon (Sample ID), (v) la date du prélèvement de l'échantillon (Date of collection), (vi) la date de réception de l'échantillon par le laboratoire (Date of receipt), (vii) le statut physique (frais, congelé etc.) de l'échantillon (Status) et enfin (viii) le pourcentage de cellules cancéreuses dans l'échantillon (Tumor load).



IMGT/CLL-DB Search Results

IMGT/V-QUEST results | Sequence | Disease | Sample | Analytical tests | Autoimmunity tests | Other tests

Your query:

Number of resulting sequences: 28

Number of resulting patients: 11 Click on Patient ID for IMGT/CLL-DB Patient card

11 items found, displaying all items.

1

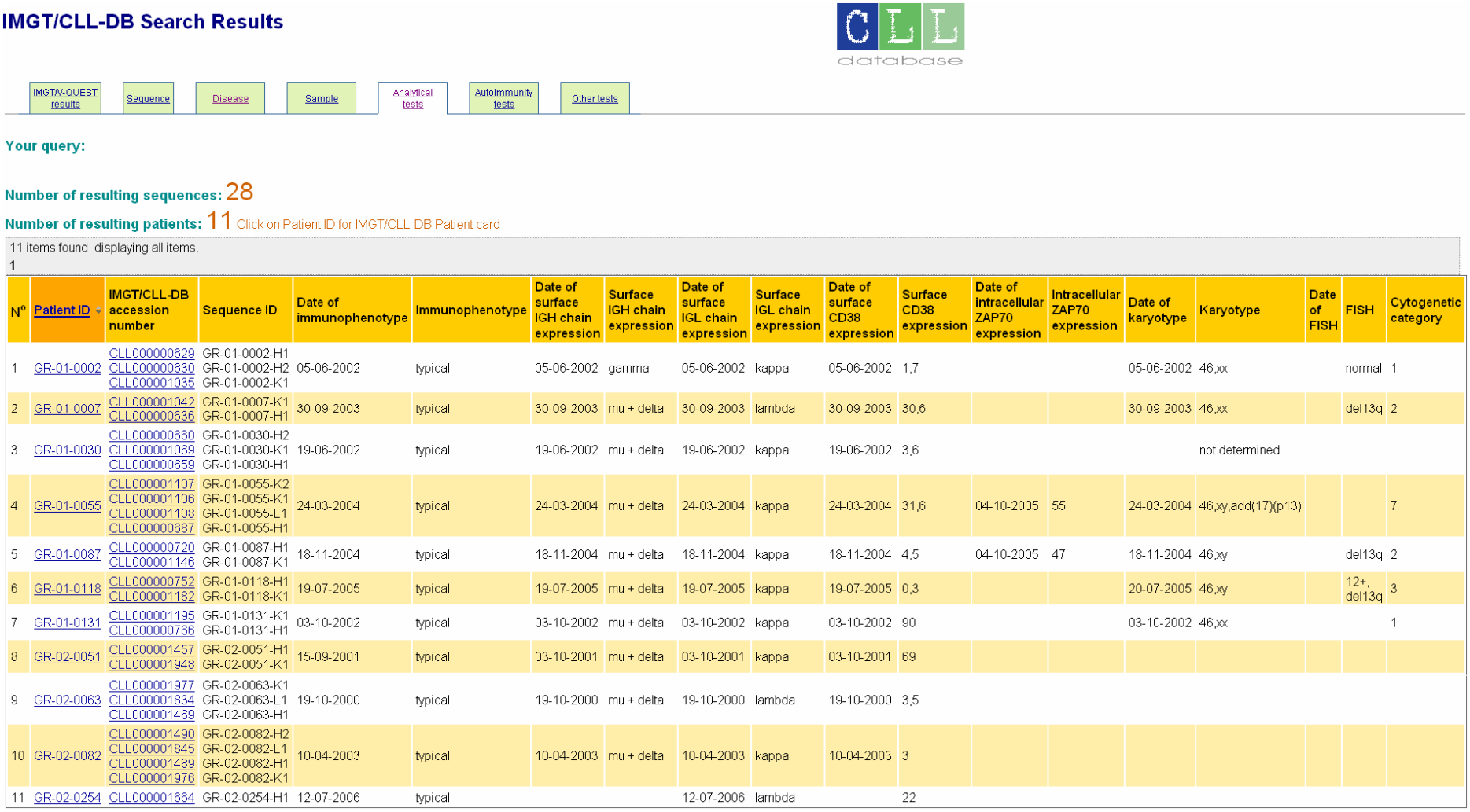
N°	Patient ID	IMGT/CLL-DB accession number	Sequence ID	Sample ID	Date of collection	Date of receipt	Status	Tumor load
6	GR-01-0002	CLL000001035 CLL000000629 CLL000000630	GR-01-0002-K1 GR-01-0002-H1 GR-01-0002-H2	gr-01-00002	05-06-2002	05-06-2002	fresh	80
11	GR-01-0007	CLL000000636 CLL000001042	GR-01-0007-H1 GR-01-0007-K1	gr-01-00007	30-09-2003	30-09-2003	fresh	83
7	GR-01-0030	CLL000000660 CLL000000659 CLL000001069	GR-01-0030-H2 GR-01-0030-H1 GR-01-0030-K1	gr-01-00030	19-06-2002	19-06-2002	fresh	82
4	GR-01-0055	CLL000001108 CLL000001106 CLL000000687 CLL000001107	GR-01-0055-L1 GR-01-0055-K1 GR-01-0055-H1 GR-01-0055-K2	gr-01-00055	24-03-2004	24-03-2004	fresh	66
8	GR-01-0087	CLL000000720 CLL000001146	GR-01-0087-H1 GR-01-0087-K1	gr-01-00087	18-11-2004	18-11-2004	fresh	68
3	GR-01-0118	CLL000000752 CLL000001182	GR-01-0118-H1 GR-01-0118-K1	gr-01-00118	19-07-2005	19-07-2005	fresh	80
2	GR-01-0131	CLL000001195 CLL000000766	GR-01-0131-K1 GR-01-0131-H1	gr-01-00131	03-10-2002	03-10-2002	fresh	78
9	GR-02-0051	CLL000001457 CLL000001948	GR-02-0051-H1 GR-02-0051-K1	gr-02-00051	26-11-2002	26-11-2002	fresh	95

Figure 4.9: L'onglet 'Sample' de la page 'IMGT/CLL-DB Search Results' présente pour chaque entrée, le Patient ID, le numéro d'accès IMGT/CLL-DB, le Sequence ID, le Sample ID, la date de collection, la date de réception de l'échantillon par le laboratoire (Date of receipt), le statut de l'échantillon et le pourcentage de cellules cancéreuses dans l'échantillon (Tumor load).

Analytical tests

L'onglet ‘Analytical tests’ (Figure 4.10) présente pour chaque patient 18 types de résultats:

En plus de (i) l'identifiant du patient (Patient ID), (ii) les numéros d'accès IMGT/CLL-DB (IMGT/CLL-DB accession number) et (iii) les identifiants des séquences associées au patient (Sequence ID). Le tableau ‘Analytic tests’ fournit associés à une date, l’immunophénotype des cellules B (typique ou atypique) ((iv) Date of immunophenotype et (v) Immunophenotype), l’isotype des chaînes lourdes et des chaînes légères des IG exprimées à la surface des cellules B, ((vi) Date of surface IGH chain expression, (vii) Surface IGH chain expression, (viii) Date of surface IGL chain expression, (ix) Surface IGL chain expression), les marqueurs de pronostic CD38 et ZAP70 ((x) Date of surface CD38 expression, (xi) Surface CD38 expression, (xii) Date of intracellular ZAP70 expression, (xiii) Intracellular ZAP70 expression), le caryotype ((xiv) Karyotype, (xv) Date of Karyotype), le résultat du FISH ((xvi) FISH), (xvii) la date associée et (xviii) la classification cytogénétique (Cytogenetic category).



IMGT/CLL-DB Search Results

IMGT/V-QUEST results | Sequence | Disease | Sample | Analytical tests | Autoimmunity tests | Other tests

Your query:

Number of resulting sequences: 28

Number of resulting patients: 11 Click on Patient ID for IMGT/CLL-DB Patient card

11 items found, displaying all items.

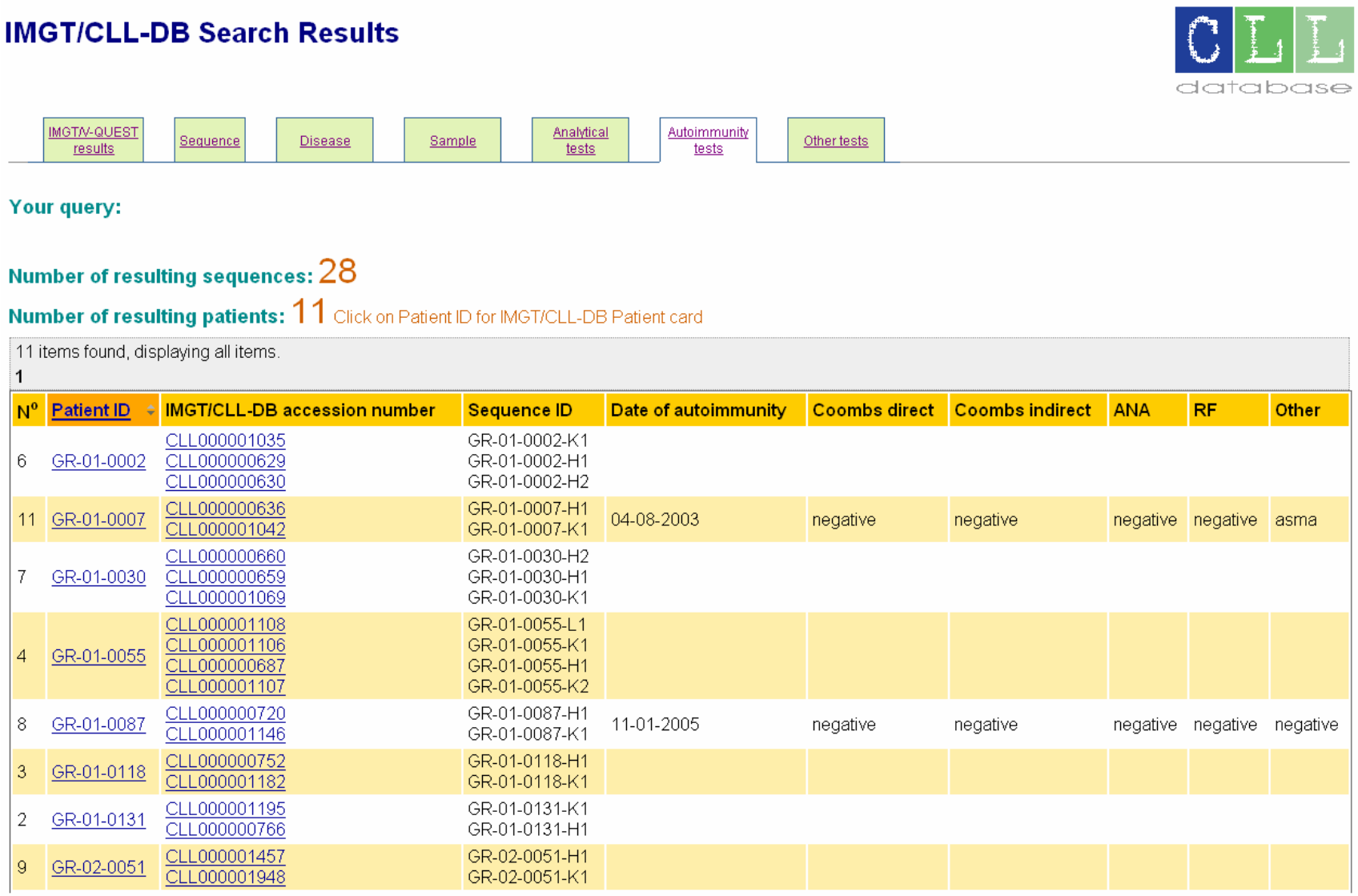
1

N°	Patient ID	IMGT/CLL-DB accession number	Sequence ID	Date of immunophenotype	Immunophenotype	Date of surface IGH chain expression	Surface IGH chain expression	Date of surface IGL chain expression	Surface IGL chain expression	Date of surface CD38 expression	Surface CD38 expression	Date of intracellular ZAP70 expression	Intracellular ZAP70 expression	Date of karyotype	Karyotype	Date of FISH	FISH	Cytogenetic category
1	GR-01-0002	CLL000000629 CLL000000630 CLL000001035	GR-01-0002-H1 GR-01-0002-H2 GR-01-0002-K1	05-06-2002	typical	05-06-2002	gamma	05-06-2002	kappa	05-06-2002	1,7			05-06-2002	46,xx		normal	1
2	GR-01-0007	CLL000001042 CLL000000636	GR-01-0007-K1 GR-01-0007-H1	30-09-2003	typical	30-09-2003	mu + delta	30-09-2003	lambda	30-09-2003	30,6			30-09-2003	46,xx		del13q	2
3	GR-01-0030	CLL000000660 CLL000001069 CLL000000659	GR-01-0030-H2 GR-01-0030-K1 GR-01-0030-H1	19-06-2002	typical	19-06-2002	mu + delta	19-06-2002	kappa	19-06-2002	3,6				not determined			
4	GR-01-0055	CLL000001107 CLL000001106 CLL000001108 CLL000000687	GR-01-0055-K2 GR-01-0055-K1 GR-01-0055-L1 GR-01-0055-H1	24-03-2004	typical	24-03-2004	mu + delta	24-03-2004	kappa	24-03-2004	31,6	04-10-2005	55	24-03-2004	46,xy,add(17)(p13)			7
5	GR-01-0087	CLL000000720 CLL000001146	GR-01-0087-H1 GR-01-0087-K1	18-11-2004	typical	18-11-2004	mu + delta	18-11-2004	kappa	18-11-2004	4,5	04-10-2005	47	18-11-2004	46,xy		del13q	2
6	GR-01-0118	CLL000000752 CLL000001182	GR-01-0118-H1 GR-01-0118-K1	19-07-2005	typical	19-07-2005	mu + delta	19-07-2005	kappa	19-07-2005	0,3			20-07-2005	46,xy		12+, del13q	3
7	GR-01-0131	CLL000001195 CLL000000766	GR-01-0131-K1 GR-01-0131-H1	03-10-2002	typical	03-10-2002	mu + delta	03-10-2002	kappa	03-10-2002	90			03-10-2002	46,xx			1
8	GR-02-0051	CLL000001457 CLL000001948	GR-02-0051-H1 GR-02-0051-K1	15-09-2001	typical	03-10-2001	mu + delta	03-10-2001	kappa	03-10-2001	69							
9	GR-02-0063	CLL000001977 CLL000001834 CLL000001469	GR-02-0063-K1 GR-02-0063-L1 GR-02-0063-H1	19-10-2000	typical	19-10-2000	mu + delta	19-10-2000	lambda	19-10-2000	3,5							
10	GR-02-0082	CLL000001490 CLL000001845 CLL000001489 CLL000001976	GR-02-0082-H2 GR-02-0082-L1 GR-02-0082-H1 GR-02-0082-K1	10-04-2003	typical	10-04-2003	mu + delta	10-04-2003	kappa	10-04-2003	3							
11	GR-02-0254	CLL000001664	GR-02-0254-H1	12-07-2006	typical			12-07-2006	lambda		22							

Figure 4.10: L'onglet 'Analytical tests' de la page 'IMGT/CLL-DB Search Results' présente, pour chaque entrée, le Patient ID, le numéro d'accès IMGT/CLL-DB, le Sequence ID, et associés chaque fois à une date, l'immunophenotypage, l'isotype des chaînes lourdes et légères exprimée à la surface des cellules B, valeur de l'expression de CD38 et de ZAP70, le caryotype, les résultats de FISH et la catégorie cytogénétique.

Autoimmunity tests

L'onglet ‘Autoimmunity tests’ (Figure 4.11) présente pour chaque patient 9 types de résultats: (i) l'identifiant du patient (Patient ID), (ii) le numéro d'accès IMGT/CLL-DB (IMGT/CLL-DB accession number), (iii) les identifiants des séquences associées au patient (Sequence ID), (iv) la date associée aux tests (Date of immunity) (v) de Coombs direct et (vi), Coombs indirect, la recherche d’autoanticorps (vii), anticorps antinucléaires (ANA), (viii) facteurs rhumatoïdes (RF) et (ix) un champ libre (Other).



IMGT/CLL-DB Search Results

IMGT/V-QUEST results | Sequence | Disease | Sample | Analytical tests | Autoimmunity tests | Other tests

Your query:

Number of resulting sequences: 28

Number of resulting patients: 11 Click on Patient ID for IMGT/CLL-DB Patient card

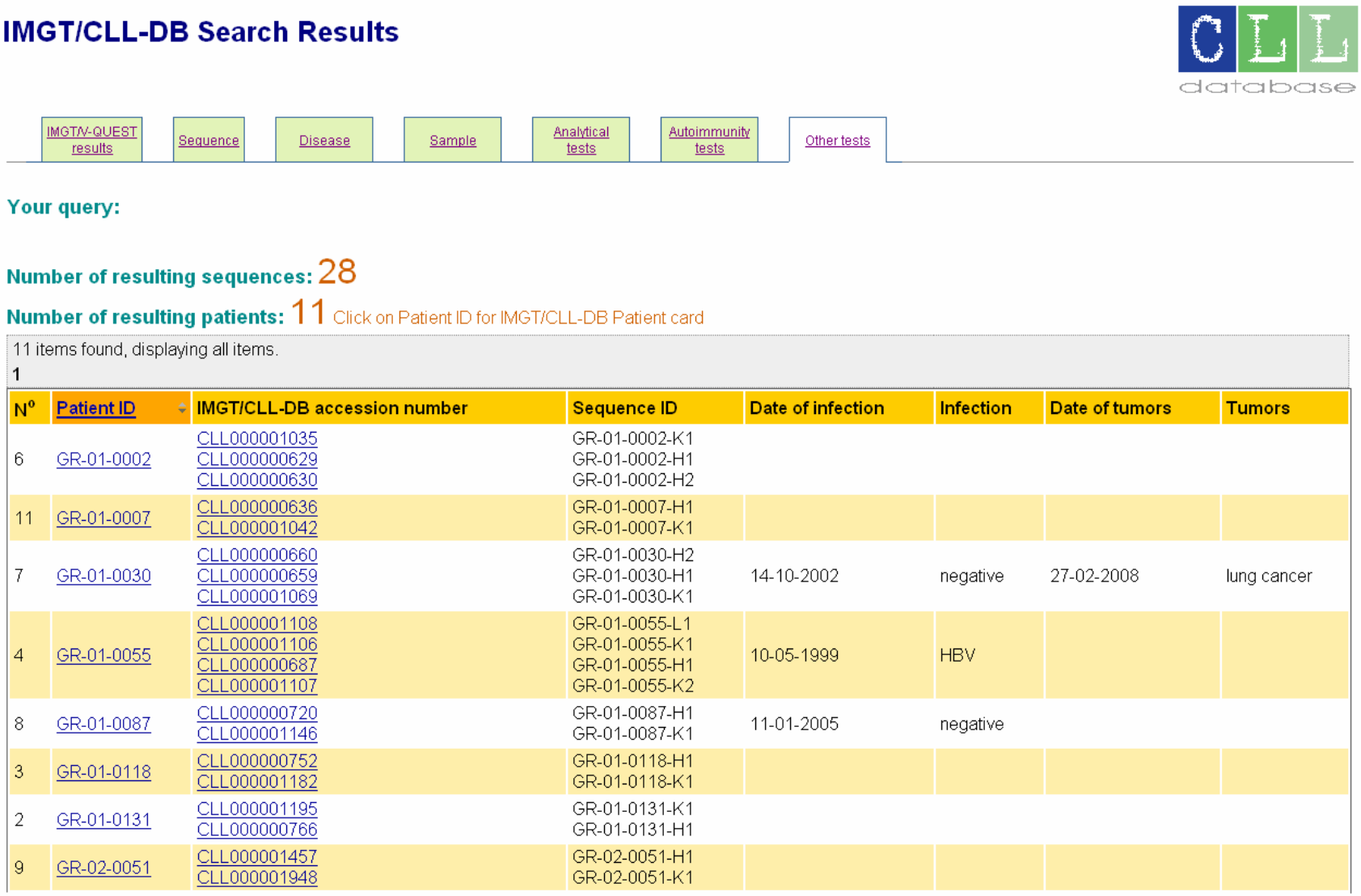
11 items found, displaying all items.
1

N°	Patient ID	IMGT/CLL-DB accession number	Sequence ID	Date of autoimmunity	Coombs direct	Coombs indirect	ANA	RF	Other
6	GR-01-0002	CLL000001035 CLL000000629 CLL000000630	GR-01-0002-K1 GR-01-0002-H1 GR-01-0002-H2						
11	GR-01-0007	CLL000000636 CLL000001042	GR-01-0007-H1 GR-01-0007-K1	04-08-2003	negative	negative	negative	negative	asma
7	GR-01-0030	CLL000000660 CLL000000659 CLL000001069	GR-01-0030-H2 GR-01-0030-H1 GR-01-0030-K1						
4	GR-01-0055	CLL000001108 CLL000001106 CLL000000687 CLL000001107	GR-01-0055-L1 GR-01-0055-K1 GR-01-0055-H1 GR-01-0055-K2						
8	GR-01-0087	CLL000000720 CLL000001146	GR-01-0087-H1 GR-01-0087-K1	11-01-2005	negative	negative	negative	negative	negative
3	GR-01-0118	CLL000000752 CLL000001182	GR-01-0118-H1 GR-01-0118-K1						
2	GR-01-0131	CLL000001195 CLL000000766	GR-01-0131-K1 GR-01-0131-H1						
9	GR-02-0051	CLL000001457 CLL000001948	GR-02-0051-H1 GR-02-0051-K1						

Figure 4.11: L'onglet 'Autoimmunity tests' de la page 'IMGT/CLL-DB Search Results' présente, pour chaque entrée, le Patient ID, le numéro d'accès IMGT/CLL-DB, le Sequence ID, la date de la série de tests et les résultats de la série de tests: Coombs direct, Coombs indirect, ANA, RF, Other**.**

Other tests

L'onglet ‘Other tests’ (Figure 4.12) présente pour chaque patient 7 types de résultats: (i) L'identifiant du patient (Patient ID), (ii) les numéros d'accès IMGT/CLL-DB (IMGT/CLL-DB accession number), (iii) les identifiants des séquences associées au patient (Sequence ID) et les complications infectieuses et les cancers qui peuvent survenir au cours de l’évolution de la maladie, (iv) la date de diagnostic d'une infection (Date of infection), (v) le nom de l'infection (infection), (vi) la date de diagnostic d’un cancer (Date of tumors) et (vii) le nom du cancer (Tumors).



IMGT/CLL-DB Search Results

IMGT/V-QUEST results | Sequence | Disease | Sample | Analytical tests | Autoimmunity tests | Other tests

Your query:

Number of resulting sequences: 28

Number of resulting patients: 11 Click on Patient ID for IMGT/CLL-DB Patient card

11 items found, displaying all items.

1

N°	Patient ID	IMGT/CLL-DB accession number	Sequence ID	Date of infection	Infection	Date of tumors	Tumors
6	GR-01-0002	CLL000001035 CLL000000629 CLL000000630	GR-01-0002-K1 GR-01-0002-H1 GR-01-0002-H2				
11	GR-01-0007	CLL000000636 CLL000001042	GR-01-0007-H1 GR-01-0007-K1				
7	GR-01-0030	CLL000000660 CLL000000659 CLL000001069	GR-01-0030-H2 GR-01-0030-H1 GR-01-0030-K1	14-10-2002	negative	27-02-2008	lung cancer
4	GR-01-0055	CLL000001108 CLL000001106 CLL000000687 CLL000001107	GR-01-0055-L1 GR-01-0055-K1 GR-01-0055-H1 GR-01-0055-K2	10-05-1999	HBV		
8	GR-01-0087	CLL000000720 CLL000001146	GR-01-0087-H1 GR-01-0087-K1	11-01-2005	negative		
3	GR-01-0118	CLL000000752 CLL000001182	GR-01-0118-H1 GR-01-0118-K1				
2	GR-01-0131	CLL000001195 CLL000000766	GR-01-0131-K1 GR-01-0131-H1				
9	GR-02-0051	CLL000001457 CLL000001948	GR-02-0051-H1 GR-02-0051-K1				

Figure 4.12: L'onglet 'Other tests' de la page 'IMGT/CLL-DB Search Results' présente, pour chaque entrée, le Patient ID, le numéro d'accès IMGT/CLL-DB, le Sequence ID, la date d'apparition d'une infection, le nom de l'infection, la date de diagnostic d'un cancer, et le nom du cancer (Tumors).

4.3.3.2 IMGT/CLL-DB Patient card

La sortie ‘IMGT/CLL-DB Patient card’ (Figure 4.13) affiche l'ensemble des données d'un patient. Les données sont présentées sous la forme de 6 tableaux regroupant les informations concernant la ou les séquence(s) (Sequence), la maladie (Disease), l’échantillon (Sample), les tests analytiques (Analytical tests), les tests autoimmuns (Autoimmunity tests) et enfin un dernier tableau qui concerne les complications infectieuses ou cancéreuses (Other tests).

Sequence

Le tableau 'Sequence' (Figure 4.13) présente pour chaque séquence d’un patient, les données concernant les séquences associées aux principaux résultats de l'analyse par IMGT/V-QUEST. Chaque séquence est identifiée par son numéro d'accès IMGT/CLL-DB (IMGT/CLL-DB accession number), par son identifiant (Sequence ID), par l'identifiant du laboratoire propriétaire de la séquence (Laboratory ID) et si la séquence est publique, par le numéro d’accès IMGT/LIGM-DB [244]. Cet identifiant est un lien direct de l'entrée dans la base IMGT/LIGM-DB [244]. De plus, à chaque séquence est associé à l’identifiant du patient (Patient ID), le type de tissu (blood, bone marrow...) (Tissu type), le type de cellule (B cell) (Cell type), le type de molécule gDNA, ou cDNA (Molecule type), la taille de la séquence (Sequence length), et enfin la séquence en nucléotides (Sequence).
Les principaux résultats de l'analyse par IMGT/V-QUEST sont aux nombre de 12, avec l'évaluation de la fonctionnalité: ‘productive’ ou ‘unproductive’ (Functionality), le nom du gène V et allèle germline (V-GENE and allele) le plus proche associé au pourcentage d'identité (V identity %), le nom du gène D et allèle germline (D-GENE and allele), le cadre de lecture du D (D reading frame), le nom du gène J et allèle germline (J-GENE and allele), la jonction en AA (JUNCTION AA), le cadre de lecture de la jonction, 'in-frame' ou 'out-of-frame' (JUNCTION frame), la taille des trois CDR-IMGT, le CRD1-IMGT (CDR1-IMGT length), le CDR2-IMGT (CDR2-IMGT length), le CDR3-IMGT (CDR3-IMGT length) et un lien vers les résultats complets et détaillés de l'analyse de la séquence par IMGT/V-QUEST (cf. paragraphe 3.5.2.1 Detailed view).

IGT/CLL-DB Patient card: GR-01-0103

Sequences Select all Clear all
IMGT/V-QUEST Synthesis view Download Sequences in FASTA or Excel

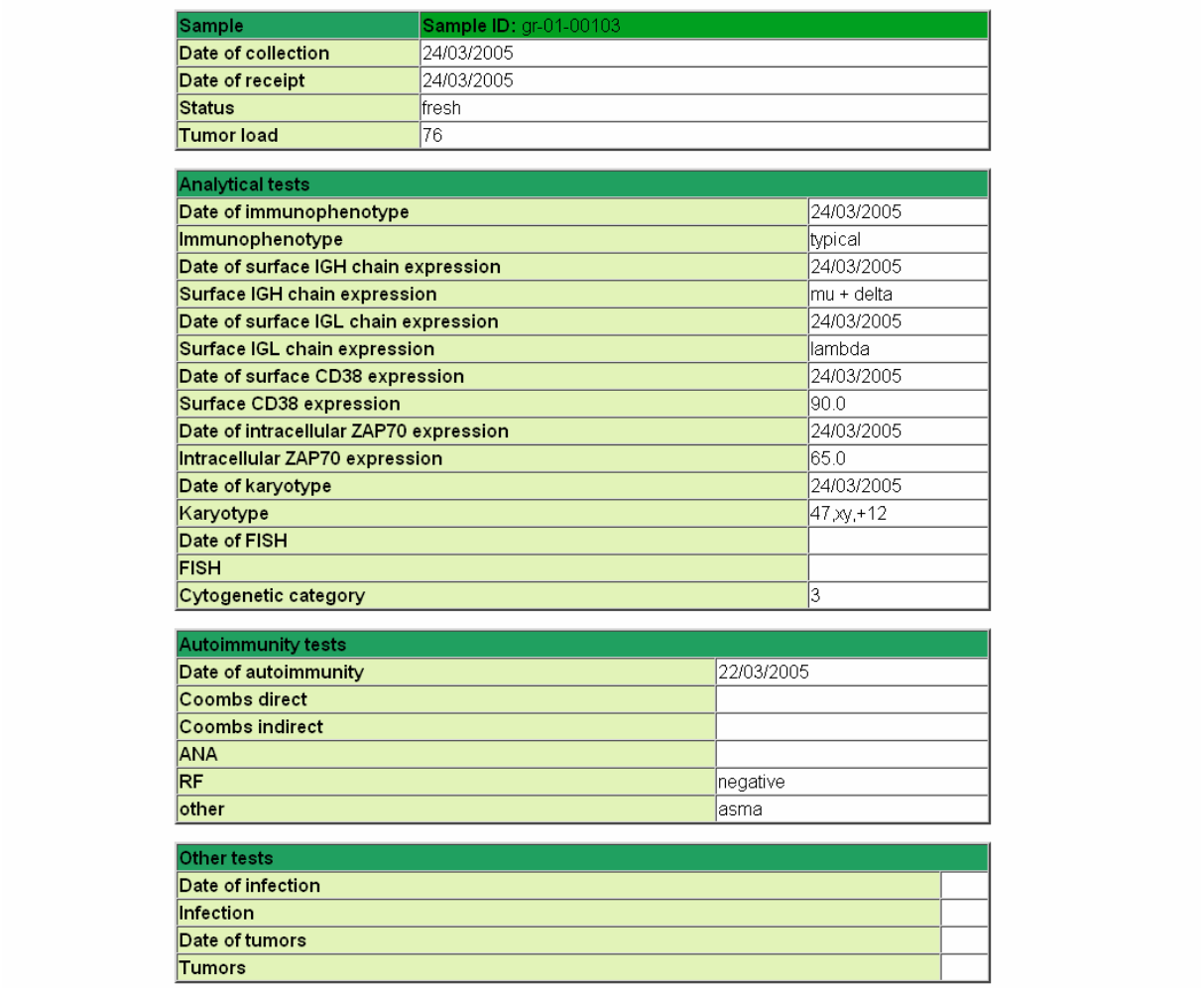
Sequence	IMGT/CLL-DB accession number: CLL000000736	Sequence ID: GR-01-0103-H1
Laboratory ID	P2703	
IMGT/LIGM-DB		
Tissue type	blood	
Cell type	B cell	
Molecule type	cDNA	
Functionality	productive	
V-GENE and allele	IGHV1-69*09	
V identity %	99.31% (286/288nt)	
D-GENE and allele by IMGT/JunctionAnalysis	IGHD3-22*01	
D reading frame	2	
J-GENE and allele	IGHJ5*01	
JUNCTION AA	CLGYYDSSGYYSLLAWW	
JUNCTION frame	in-frame	
CDR1-IMGT length	8	
CDR2-IMGT length	8	
CDR3-IMGT length	15	
IMGT/V-QUEST Detailed view	CLL000000736	

Sequence:

```
tcaggtgcagctggtgcagtctggggctgaggtgaagaagcctgggtcctcggtgaaggt
ctcctgcaaggcttctggaggcaccttcagcagctatactatcagctgggtgcgacaggc
ccctggacaagggcttgagtggatgggaaggatcatccctatccttggtatagcaaacta
cgcacagaagttccagggcagagtcacgattaccgcggacaaatccacgagcacagccta
catggagctgagcagcctgagatctgaggacacggccgtgtattactgcttggggtacta
tgatagtagtggttattacagtcttttagcctggtggggccagggaaccctgttcaccgt
ctcctcaa
```

Length: 368

Disease	Patient ID: GR-01-0103
Date of birth	12/02/1938
Gender	M
Ethnic origin	Caucasoid
Country	Greece
Date of diagnosis	24/03/2005
Age at diagnosis	67
Rai	II
Binet	B
Family history of CLL	
Family history of other B cell proliferations	
Date of clinical evolution	04/12/2007
Clinical evolution	stable
Need for treatment	not applicable
Date of first treatment	
Response to treatment	not applicable
Date of transformation	
Transformation	
Date of last follow-up	04/12/2007
Current status	alive
Date of follow-up loss	
CLL related death	
Date of death	
Date of clinical autoimmunity	
Clinical autoimmunity	



Sample	Sample ID: gr-01-00103
Date of collection	24/03/2005
Date of receipt	24/03/2005
Status	fresh
Tumor load	76

Analytical tests	
Date of immunophenotype	24/03/2005
Immunophenotype	typical
Date of surface IGH chain expression	24/03/2005
Surface IGH chain expression	mu + delta
Date of surface IGL chain expression	24/03/2005
Surface IGL chain expression	lambda
Date of surface CD38 expression	24/03/2005
Surface CD38 expression	90.0
Date of intracellular ZAP70 expression	24/03/2005
Intracellular ZAP70 expression	65.0
Date of karyotype	24/03/2005
Karyotype	47,xy,+12
Date of FISH	
FISH	
Cytogenetic category	3

Autoimmunity tests	
Date of autoimmunity	22/03/2005
Coombs direct	
Coombs indirect	
ANA	
RF	negative
other	asma

Other tests	
Date of infection	
Infection	
Date of tumors	
Tumors	

Figure 4.13: IMGT/CLL-DB Patient card. Cette carte affiche l'ensemble des données d'un patient. Les données sont présentées sous la forme de 6 tableaux regroupant les informations concernant la (ou les) séquence(s) (Sequence), la maladie (Disease), l'échantillon (Sample), les tests analytiques (Analytical tests), les tests autoimmuns (Autoimmunity tests) et un dernier tableau qui concerne les complications infectieuses ou cancéreuses (Other tests).

Disease

Le tableau 'Disease' (Figure 4.13) présente 24 résultats décrivant le patient, sa maladie et son évolution.

(i) Le patient est identifié par l'identifiant du patient (Patient ID) et décrit par (ii) la date de naissance (Date of birth), (iii) le sexe (Gender), (iv) l'origine ethnique (Ethnic origin), (v) le pays de résidence (Country), (vi) la date de diagnostic de la maladie (Date of diagnosis), (vii) l'âge du patient au diagnostic (Age at diagnosis), (viii et ix) le stade de la maladie selon les deux classifications Rai et Binet (Rai, Binet), (x et xi) l'histoire familiale de la LLC (Family history of CLL) et d'autres syndromes lymphoprolifératifs B (Family history of other B cell proliferations), (xii et xiii) l'évolution de la LLC associée à une date (Date of clinical evolution, Clinical evolution), (xiv) le besoin d'un traitement (Need for treatment), (xv) la date du premier traitement (Date of first treatment), (xvi) la réponse au traitement (Response

to treatment), (xvii) la date de complication vers une forme agressive de la maladie (Date of transformation), (xviii) le type de transformation (Transformation), (xix) la date de la dernière consultation du patient (Date of last follow-up), (xx) le statut du patient (Current status), (xxi) la date de la perte du suivi du patient (Date of follow-up loss), (xxii) le lien ou non du décès avec la LLC (CLL related death), la date du décès (Date of death), (xxiii) la date d'apparition de complications autoimmunes (Date of clinical autoimmunity) et (xxiv) le type de complications autoimmunes (Clinical autoimmunity).

Sample

Le tableau 'Sample' (Figure 4.13) présente 5 types de résultats: (i) l'identifiant de l'échantillon (Sample ID), (ii) la date du prélèvement de l'échantillon (Date of collection), (iii) la date de réception de l'échantillon par le laboratoire (Date of receipt), (iv) le statut physique (frais, congelé etc.) de l'échantillon (Status) et enfin (v) le pourcentage de cellules cancéreuses dans l'échantillon (Tumor load).

Analytical tests

Le tableau ‘Analytical tests’ (Figure 4.13) présente 15 types de résultats:

Le tableau ‘Analytic tests’ fournit associés à une date, l’immunophénotype des cellules B (typique ou atypique) ((i) Date of immunophenotype et (ii) Immunophenotype), l’isotype des chaînes lourdes et des chaînes légères des IG exprimées à la surface des cellules B, ((iii) Date of surface IGH chain expression, (iv) Surface IGH chain expression, (v) Date of surface IGL chain expression, (vi) Surface IGL chain expression), les marqueurs de pronostic CD38 et ZAP70 ((vii) Date of surface CD38 expression, (viii) Surface CD38 expression, (ix) Date of intracellular ZAP70 expression, (x) Intracellular ZAP70 expression), le caryotype ((xi) Date of Karyotype, (xii) Karyotype), le résultat du FISH ((xiiii) FISH), (xiv) la date associée et (xv) la classification cytogénétique (Cytogenetic category).

Autoimmunity tests

Le tableau ‘Autoimmunity tests’ (Figure 4.13) présente pour chaque patient 6 types de résultats:

Les résultats des tests de signe autoimmun associé à une date (i) Date of autoimmunity, test de Coombs (ii) Coombs direct et (iii) Coombs indirect, et la recherche d’autoanticorps (iv) anticorps antinucléaires (ANA) et (v) facteurs rhumatoïdes (RF) et (vi) un champ libre (Other).

Other tests

Le tableau 'Other tests' (Figure 4.13) présente 4 types de résultats:
Les complications infectieuses et les cancers qui peuvent survenir au cours de l'évolution de la maladie, (i) la date de diagnostic d'une infection (Date of infection), (ii) le nom de l'infection (infection), (iii) la date de diagnostic d'un cancer (Date of tumors) et (iv) le nom du cancer (Tumors).

4.3.4 Accès aux résultats détaillés d'IMGT/V-QUEST

A partir des pages de résultat de IMGT/CLL-DB, que ce soit 'IMGT/CLL-DB Search results' ou 'IMGT/CLL-DB Patient card', il est possible de visualiser les résultats complets de IMGT/V-QUEST.

4.3.4.1 IMGT/V-QUEST Detailed view

Le tableau de résultats présenté dans l'onglet 'IMGT/V-QUEST results' de 'IMGT/CLL-DB Search Results' (Figure 4.6) contient une colonne 'IMGT/V-QUEST Detailed view'qui permet pour une séquence donnée de visualiser les résultats de IMGT/V-QUEST sous le type 'IMGT/V-QUEST Detailed view' (cf. Chapitre 3 IMGT/V-QUEST).
La même option est accessible à partir de 'IMGT/CLL-DB Patient card' dans l'avant dernière ligne du tableau 'Sequence' (Figure 4.13).

4.3.4.2 IMGT/V-QUEST Synthesis view

Les onglets 'IMGT/V-QUEST Results' et 'Sequence' de 'IMGT/CLL-DB Search Results' contiennent un bouton 'Synthesis view', situé au-dessus du tableau de résultat (Figure 4.6 et Figure 4.7) qui permet de visualiser, pour les séquences sélectionnées, les résultats de l'analyse 'IMGT/V-QUEST Synthesis view' (cf. Chapitre 3 IMGT/V-QUEST). Cette option permet d'analyser jusqu'à 50 séquences simultanément.

La même option est accessible à partir de 'IMGT/CLL-DB Patient card', le bouton étant situé en haut de la page de résultat (Figure 4.13).

4.3.5 Téléchargement des données de la base IMGT/CLL-DB

A partir de l'interface IMGT/CLL-DB l'utilisateur peut télécharger les résultats d'une requête sous 2 formats différents:

- les séquences en format FASTA dans un fichier texte.
- les séquences et les informations associées dans un fichier Excel.

4.3.5.1 Téléchargement des séquences en format FASTA

Deux boutons, présents en haut des pages de résulats permettent d'accéder à cette fonctionnalité.

- le bouton 'Sequences in FASTA' permet de télécharger les séquences qui ont été sélectionnées dans une page de résultat (une page contient au maximun les résultats de 50 séquences). Ce bouton est accessible dans la page 'IMGT/CLL-DB Patient card' (Figure 4.13) et dans les onglets 'IMGTV-QUEST results' (Figure 4.6), 'Sequence' (Figure 4.7) et 'Disease' (Figure 4.8) de 'IMGT/CLL-DB Search Results'.
- le bouton 'FASTA sequences' permet de télécharger l'ensemble des séquences résultant d'une requête, quelque soit leur nombre, en format FASTA. Ce bouton est accessible à partir des onglets 'IMGTV-QUEST results' (Figure 4.6), 'Sequence' (Figure 4.7) et 'Disease' (Figure 4.8) de la page 'IMGT/CLL-DB Search Results'.

4.3.5.2 Téléchargement des séquences et des informations associées dans un fichier Excel

L'utilisateur à la possibilité de télécharger les séquences et leurs informations associées dans un fichier Excel dont la présentation est en tout point identique à celle des fichiers Excel utilisés pour la sélection des nouvelles données (cf. Chapitre 4.2.1 Sélection des nouvelles données) avec en plus les résultats de l'analyse des séquences par IMGT/V-QUEST.

- le bouton 'Excel' permet de télécharger les séquences et les données associées qui ont été sélectionnée ans une page de résultat. Ce bouton est accessible dans la page 'IMGT/CLL-DB Patient card' (Figure 4.13) et dans les onglets 'IMGTV-QUEST results' (Figure 4.6), 'Sequence' (Figure 4.7) et 'Disease' (Figure 4.8) de 'IMGT/CLL-DB Search Results'.
- le bouton 'Excel file' permet de télécharger l'ensemble des séquences et des données associées résultant d'une requête, quelque soit leur nombre. Ce bouton est accessible à partir des onglets 'IMGTV-QUEST results' (Figure 4.6), 'Sequence' (Figure 4.7) et 'Disease' (Figure 4.8) de la page 'IMGT/CLL-DB Search Results'.

Conclusion

La base de données IMGT/CLL-DB associée au programme spécialisé IMGT/V-QUEST forme un système d'information cohérent dédié à l'analyse et à la gestion des récepteurs d'antigènes dans le cadre de pathologies. Ce système a été implémenté dans le cadre d'une collaboration internationale avec des équipes cliniques spécialistes de la LLC. La mise en oeuvre de ce système d'information a constitué la majeure partie de mon travail de thèse. L'organisation de la base de données IMGT/CLL-DB permet de prendre en compte les différentes informations structurées selon 3 thématiques, (1) les informations concernant les séquences des IG et leur analyse par IMGT/V-QUEST, (2) les informations concernant la le patient et la pathologie (évolution, tests, suivi) et (3) les données concernant les échantillons prélevés sur les patients (date du prélèvement, date de réception, type de tissu…). Les séquences d'IG sont décrites en détail avec l'ensemble des résultats de l'analyse par IMGT/V-QUEST, tandis que les données associées aux patients sont regroupées en '54' classes d'informations qui ont été sélectionnées en accord avec les partenaires du projet.

Afin de maintenir et de gérer les données de la base IMGT/CLL-DB, nous avons implémenté un programme JAVA semi-automatique qui gère le chargement des nouvelles entrées et des mises à jour. Le contrôle des données est obtenu par la mise en place d'un système d'intégrité constitué de 3 niveaux de sécurité, (1) un vocabulaire contrôlé défini pour 28 classes d'informations, (2) des règles intrinsèques à la base de données qui définissent le type et la taille des informations pouvant être stockées pour un champ donné, et (3) des programmes informatiques qui déclenchent des actions de vérification du format des données lors d'insertions ou de mises à jour.

Le système d'information IMGT/CLL-DB a été finalisé par la mise en place d'une interface Web développée pour rendre l'ensemble des données de la base accessible de façon simple et conviviale aux seuls partenaires du projet. L'interface offre un système d'interrogation avancé qui permet aux utilisateurs de rechercher les informations selon 39 critères de recherche divisés en 8 thématiques. Deux types de sorties ont été implémentés pour afficher les résultats. Le 'Patient card' présente toutes les données concernant un patient, et 'IMGT/CLL-DB Search Results' présente pour de multiples séquences et/ou patients, l'intégralité des informations regroupées par thème grâce à un système d'onglet permettant une comparaison aisée entre les données. L'interface Web offre des outils dynamiques (1) pour récupérer les séquences sous format FASTA, (2) pour récupérer l'ensemble des données sous format Excel et (3) pour utiliser 'IMGT/V-QUEST Synthesis view' de façon dynamique.

Les améliorations récentes et les limites de l'outil IMGT/V-QUEST nous amènent à développer de nouveaux projets dans le but de perfectionner le système d'information. Dans un premier temps nous souhaitons intégrer systématiquement l'option de détection automatique des insertions et des délétions potentielles à l'analyse des séquences d'IG par IMGT/V-QUEST lors du chargement de nouvelles séquences. Lors d'un second projet nous souhaitons faire évoluer l'option 'IMGT/V-QUEST Synthesis view' pour nous affranchir de la limite du nombre de séquences visualisées.
Finalement, nous avons développé un prototype d'interface Web pour la réalisation des mises à jour ponctuelles. Nous souhaitons finaliser cette interface pour que ces mises à jour puissent être effectuées directement par les propriétaires des séquences.

Ce système d'information résulte d'une volonté des laboratoires impliqués dans le projet de partager et de fédérer les données issues de la recherche médicale et fondamentale. Les partenaires du projet (équipes cliniques et équipe IMGT) ont défini les facteurs, les informations de patients atteints de la LLC et les caractéristiques génétiques des IG à corréler pour mieux cerner les mécanismes impliqués dans la pathologie. Notamment, il est possible de comparer les marqueurs de pronostic tels que le CD38 ou le ZAP70, au taux de mutations des chaînes lourdes des IG. Nous avons été conduit à standardiser les valeurs associées à chaque information avec la mise en place d'un vocabulaire contrôlé.
Ce système d'information pourra facilement être adapté à d'autres pathologies ou à la gestion de récepteurs d'antigènes (IG et TR) répondant à une même spécificité (reconnaissance de même antigène) avec une modification minimale du modèle de base.

DISCUSSION ET CONCLUSION

IMGT® est le système d'information international en ImMunoGénétique®, spécialisé dans la gestion des séquences et des structures 3D des IG, TR et MCH des vertébrés. Ces protéines assurent la reconnaissance antigénique et la spécificité du système immunitaire adaptatif. Compte tenu de la nécessité d'avoir des structures spécialisées dans la gestion des données biologiques et des standards de haute qualité utilisés dans la recherche aussi bien fondamentale que médicale, nous avons développé en partenariat avec des laboratoires cliniques au sein de IMGT®, un système d'information dédié à l'analyse et à la gestion des récepteurs d'antigènes en relation avec des pathologies du système immunitaire: constitué de l'outil IMGT/V-QUEST [236] spécialisé dans l'analyse des séquences réarrangées des récepteurs d'antigènes et de la base de donnée IMGT/CLL-DB en prenant l'exemple de la LLC. Cette approche s'avère original car c'est à notre connaissance le premier système d'information dont la donnée principale est la séquence à laquelle sont rattachés des informations cliniques de patients.

La première partie du travail réalisée au cours de cette thèse a porté sur la réécriture du programme cœur de l'outil IMGT/V-QUEST. Nous avons d'une part réorganisé, homogénéisé et standardisé le programme dans le but de faciliter l'intégration de l'outil dans de nouvelles applications. Des améliorations significatives des performances de l'outil nous permettent d'analyser les séquences par lots, la limite étant actuellement de 50 séquences pour une analyse standard, sans recherche des insertions et délétions, et de 10 séquences dans le cas où l'option de recherche des insertions et délétions a été sélectionnée.
D'autre part, nous avons apporté de nombreux perfectionnements significatifs en incorporant de nouvelles fonctionnalités, améliorant la précision de la description des séquences réarrangées des utilisateurs. L'outil a maintenant la capacité d'évaluer systématiquement le taux de mutations dans la V-REGION des chaînes lourdes des IG (pourcentage d'identité entre la V-REGION de la séquence utilisateur et la V-REGION du gène germline le plus proche) considéré comme le marqueur pronostic le plus pertinent actuellement pour la LLC. Les mutations présentes dans les V-REGION sont localisées et caractérisées comme étant silencieuses ou non silencieuses et comme des transitions ou des transversions. Les mutations provoquant des changements d'acides aminés dans la protéine sont qualifiées en accord avec les classes AA d'IMGT [241] basées sur l'hydropathie, le volume et les

propriétés physicochimiques des AA. L'outil identifie également la position des nucléotides préférentiellement mutés (décrit dans la littérature) que l'on appelle les positions hot spot dans le gène et allèle V germline le plus proche de la séquence utilisateur. Une option permet de localiser les insertions et/ou délétions potentielles qui peuvent survenir dans la V-REGION des séquences réarrangées lors de la synthèse des IG aussi bien dans les cellules normales que dans les cellules malignes [242]. Une évaluation de la fonctionnalité de la séquence utilisateur est proposée comme étant 'productive' ou 'unproductive'. Enfin, il en résulte une annotation complète des V-J et V-D-J-REGION. Parallèlement, nous avons fait évolué l'interface Web utilisateur dans le but d'offrir un outil hautement paramétrable répondant aux besoins des utilisateurs.

Dans la seconde partie de ma thèse et ce grâce aux améliorations apportées à l'outil IMGT/V-QUEST, nous avons développé dans le cadre d'une collaboration avec des laboratoires de recherche clinique, un système d'information dédié à l'analyse et à la gestion des séquences réarrangées des IG de patients atteints de la LLC. La base de données IMGT/CLL-DB est conforme aux règles et au vocabulaire de IMGT-ONTOLOGY, ce qui lui confère un haut niveau de standardisation, s'étendant à la description de tous les récepteurs d'antigènes. Cette standardisation apparaît comme un élément crucial en bioinformatique depuis l'avènement de techniques expérimentales produisant des quantités importantes de données génomiques. L'architecture de IMGT/CLL-DB, associée au programme IMGT/V-QUEST répond à ces principes. L'annotation automatique de IMGT/V-QUEST apporte également l'avantage de contribuer à la cohérence des données dans la base de données. IMGT/CLL-DB a été conçue dans le cadre d'une collaboration avec des équipes cliniques, nous offrant la possibilité d'un premier développement collaboratif et intégratif des domaines médicaux, immunoinformatiques, et immunogénétiques. Les données de séquences correspondent à l'annotation détaillée et précise de la séquence réarrangée par l'analyse de IMGT/V-QUEST. Les données médicales relatives aux patients et leur pathologie gérées dans la base ont été déterminées par les partenaires du projet. Toutes les informations ont été, dans la mesure du possible, standardisées par la mise en place d'un vocabulaire contrôlé. Le système d'information permet de corréler les données médicales relatives aux patients avec les données génétiques des séquences d'IG, pour une meilleure caractérisation des mécanismes moléculaires impliqués dans la pathologie.

Le système d'information, appliqué à la LLC, est un premier modèle proposé pour l'organisation, et la création d'un système d'information destiné à intégrer et à fédérer des

données issues de la recherche fondamentale et la recherche clinique. L'architecture du système d'information mise en œuvre au sein d'IMGT®, devrait pouvoir s'appliquer avec succès à d'autres pathologies du système immunitaire (maladies autoimmunes, SIDA...). J'espère que les travaux effectués durant cette thèse pourront servir à leur développement et apporteront une aide non négligeable aux travaux de recherche des équipes cliniques partenaires du projet.

BIBLIOGRAPHIE

1. Gelbart, W.M. Databases in genomic research. Science 282, 659-661 (1998).
2. Tonegawa, S. Somatic generation of antibody diversity. Nature 302, 575-581 (1983).
3. Lefranc, M.-P. and Lefranc, G. *The immunoglobulin FactsBook.* , ed. A. Press. 2001.
4. Lefranc, M.-P. and Lefranc, G. *The T cell receptor FactsBook*, ed. L. Academic Press, UK. 2001.
5. Magdelaine-Beuzelin, C., Kaas, Q., Wehbi, V., Ohresser, M., Jefferis, R., Lefranc, M.P. and Watier, H. Structure-function relationships of the variable domains of monoclonal antibodies approved for cancer treatment. Crit Rev Oncol Hematol 64, 210-225 (2007).
6. Fais, F., Ghiotto, F., Hashimoto, S., Sellars, B., Valetto, A., Allen, S.L., Schulman, P., Vinciguerra, V.P., Rai, K., Rassenti, L.Z., Kipps, T.J., Dighiero, G., Schroeder, H.W., Jr., Ferrarini, M. and Chiorazzi, N. Chronic lymphocytic leukemia B cells express restricted sets of mutated and unmutated antigen receptors. J Clin Invest 102, 1515-1525 (1998).
7. Stamatopoulos, K., Belessi, C., Moreno, C., Boudjograh, M., Guida, G., Smilevska, T., Belhoul, L., Stella, S., Stavroyianni, N., Crespo, M., Hadzidimitriou, A., Sutton, L., Bosch, F., Laoutaris, N., Anagnostopoulos, A., Montserrat, E., Fassas, A., Dighiero, G., Caligaris-Cappio, F., Merle-Béral, H., Ghia, P. and Davi, F. Over 20% of patients with chronic lymphocytic leukemia carry stereotyped receptors: Pathogenetic implications and clinical correlations. Blood 109, 259-270 (2007).
8. Kostareli, E., Smilevska, T., Stamatopoulos, K., Kouvatsi, A. and Anagnostopoulos, A. Chronic lymphocytic leukaemia: an immunobiology approach. Srp Arh Celok Lek 136, 319-323 (2008).
9. Lefranc, M.-P., Giudicelli, V., Ginestoux, C., Jabado-Michaloud, J., Folch, G., Wu, Y., Bellahcene, F., Gemrot, E., Brochet, X., Lane, J., Regnier, L., Ehrenmann, F., Lefranc, G. and Duroux, P. IMGT®, the international ImMunoGeneTics information system®. Nucleic Acids Res (2008).
10. Lefranc, M.-P., Giudicelli, V., Kaas, Q., Duprat, E., Jabado-Michaloud, J., Scaviner, D., Ginestoux, C., Clement, O., Chaume, D. and Lefranc, G. IMGT, the international ImMunoGeneTics information system. Nucleic Acids Res 33, D593-D597 (2005).
11. Giudicelli, V. and Lefranc, M.-P. Ontology for immunogenetics: the IMGT-ONTOLOGY. Bioinformatics 15, 1047-1054 (1999).
12. Lefranc, M.-P., Giudicelli, V., Ginestoux, C., Bosc, N., Folch, G., Guiraudou, D., Jabado-Michaloud, J., Magris, S., Scaviner, D., Thouvenin, V., Combres, K., Girod, D., Jeanjean, S., Protat, C., Yousfi-Monod, M., Duprat, E., Kaas, Q., Pommie, C., Chaume, D. and Lefranc, G. IMGT-ONTOLOGY for immunogenetics and immunoinformatics. In Silico Biol 4, 17-29 (2004).
13. Duroux, P., Kaas, Q., Brochet, X., Lane, J., Ginestoux, C., Lefranc, M.-P. and Giudicelli, V. IMGT-Kaleidoscope, the formal IMGT-ONTOLOGY paradigm. Biochimie 90, 570-583 (2008).
14. Lefranc, M.-P., Pommié, C., Ruiz, M., Giudicelli, V., Foulquier, E., Truong, L., Thouvenin-Contet, V. and Lefranc, G. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. Dev Comp Immunol 27, 55-77 (2003).
15. Lefranc, M.-P., Pommié, C., Kaas, Q., Duprat, E., Bosc, N., Guiraudou, D., Jean, C., Ruiz, M., Da Piédade, I., Rouard, M., Foulquier, E., Thouvenin, V. and Lefranc, G.

IMGT unique numbering for immunoglobulin and T cell receptor constant domains and Ig superfamily C-like domains. Dev Comp Immunol 29, 185-203 (2005).
16. Lefranc, M.-P., Duprat, E., Kaas, Q., Tranne, M., Thiriot, A. and Lefranc, G. IMGT unique numbering for MHC groove G-DOMAIN and MHC superfamily (MhcSF) G-LIKE-DOMAIN. Dev Comp Immunol 29, 917-938 (2005).
17. Lefranc, M.-P. WHO-IUIS Nomenclature Subcommittee for immunoglobulins and T cell receptors report. Immunogenetics 59, 899-902 (2007).
18. Lefranc, M.-P. WHO-IUIS Nomenclature Subcommittee for immunoglobulins and T cell receptors report August 2007, 13th International Congress of Immunology, Rio de Janeiro, Brazil. Dev Comp Immunol 32, 461-463 (2008).
19. Brack, C., Hirama, M., Lenhard-Schuller, R. and Tonegawa, S. A complete immunoglobulin gene is created by somatic recombination. Cell 15, 1-14 (1978).
20. Weigert, M., Perry, R., Kelley, D., Hunkapiller, T., Schilling, J. and Hood, L. The joining of V and J gene segments creates antibody diversity. Nature 283, 497-499 (1980).
21. Sakano, H., Huppi, K., Heinrich, G. and Tonegawa, S. Sequences at the somatic recombination sites of immunoglobulin light-chain genes. Nature 280, 288-294 (1979).
22. Early, P., Huang, H., Davis, M., Calame, K. and Hood, L. An immunoglobulin heavy chain variable region gene is generated from three segments of DNA: VH, D and JH. Cell 19, 981-992 (1980).
23. Schatz, D.G., Oettinger, M.A. and Baltimore, D. The V(D)J recombination activating gene, RAG-1. Cell 59, 1035-1048 (1989).
24. Oettinger, M.A., Schatz, D.G., Gorka, C. and Baltimore, D. RAG-1 and RAG-2, adjacent genes that synergistically activate V(D)J recombination. Science 248, 1517-1523 (1990).
25. Max, E.E., Seidman, J.G. and Leder, P. Sequences of five potential recombination sites encoded close to an immunoglobulin kappa constant region gene. Proc. Natl Acad. Sci. USA 76, 3450-3454 (1979).
26. Alt, F.W. and Baltimore, D. Joining of immunoglobulin heavy chain gene segments: implications from a chromosome with evidence of three D-JH fusions. Proc. Natl Acad. Sci. USA 79, 4118-4122 (1982).
27. Landau, N.R., St John, T.P., Weissman, I.L., Wolf, S.C., Silverstone, A.E. and Baltimore, D. Cloning of terminal transferase cDNA by antibody screening. Proc. Natl Acad. Sci. USA 81, 5836-5840 (1984).
28. Lafaille, J.J., DeCloux, A., Bonneville, M., Takagaki, Y. and Tonegawa, S. Junctional sequences of T cell receptor gamma delta genes: implications for gamma delta T cell lineages and for a novel intermediate of V-(D)-J joining. Cell 59, 859-870 (1989).
29. Lewis, S.M. P nucleotide insertions and the resolution of hairpin DNA structures in mammalian cells. Proc. Natl Acad. Sci. USA 91, 1332-1336 (1994).
30. Yannone, S.M., Khan, I.S., Zhou, R.Z., Zhou, T., Valerie, K. and Povirk, L.F. Coordinate 5' and 3' endonucleolytic trimming of terminally blocked blunt DNA double-strand break ends by Artemis nuclease and DNA-dependent protein kinase. Nucleic Acids Res. 36, 3354-3365 (2008).
31. Niewolik, D., Pannicke, U., Lu, H., Ma, Y., Wang, L.C., Kulesza, P., Zandi, E., Lieber, M.R. and Schwarz, K. DNA-PKcs dependence of Artemis endonucleolytic activity, differences between hairpins and 5' or 3' overhangs. J Biol Chem 281, 33900-33909 (2006).

32. Thai, T.H., Purugganan, M.M., Roth, D.B. and Kearney, J.F. Distinct and opposite diversifying activities of terminal transferase splice variants. Nat Immunol 3, 457-462 (2002).
33. Thai, T.H. and Kearney, J.F. Isoforms of terminal deoxynucleotidyltransferase: developmental aspects and function. Adv Immunol 86, 113-136 (2005).
34. Doyen, N., Boule, J.B., Rougeon, F. and Papanicolaou, C. Evidence that the long murine terminal deoxynucleotidyltransferase isoform plays no role in the control of V(D)J junctional diversity. J Immunol 172, 6764-6767 (2004).
35. Gearhart, P.J., Johnson, N.D., Douglas, R. and Hood, L. IgG antibodies to phosphorylcholine exhibit more diversity than their IgM counterparts. Nature 291, 29-34 (1981).
36. Teng, G. and Papavasiliou, F.N. Immunoglobulin somatic hypermutation. Annu Rev Genet 41, 107-120 (2007).
37. Fukita, Y., Jacobs, H. and Rajewsky, K. Somatic hypermutation in the heavy chain locus correlates with transcription. Immunity 9, 105-114 (1998).
38. Bachl, J., Carlson, C., Gray-Schopfer, V., Dessing, M. and Olsson, C. Increased transcription levels induce higher mutation rates in a hypermutating cell line. J Immunol 166, 5051-5057 (2001).
39. Peled, J.U., Kuang, F.L., Iglesias-Ussel, M.D., Roa, S., Kalis, S.L., Goodman, M.F. and Scharff, M.D. The biochemistry of somatic hypermutation. Annu Rev Immunol 26, 481-511 (2008).
40. Di Noia, J.M. and Neuberger, M.S. Molecular mechanisms of antibody somatic hypermutation. Annu. Rev. Biochem. 76, 1-22 (2007).
41. Petersen-Mahrt, S.K., Harris, R.S. and Neuberger, M.S. AID mutates E. coli suggesting a DNA deamination mechanism for antibody diversification. Nature 418, 99-103 (2002).
42. Di Noia, J. and Neuberger, M.S. Altering the pathway of immunoglobulin hypermutation by inhibiting uracil-DNA glycosylase. Nature 419, 43-48 (2002).
43. Knapp, M.R., Liu, C.P., Newell, N., Ward, R.B., Tucker, P.W., Strober, S. and Blattner, F. Simultaneous expression of immunoglobulin mu and delta heavy chains by a cloned B-cell lymphoma: a single copy of the VH gene is shared by two adjacent CH genes. Proc. Natl Acad. Sci. USA 79, 2996-3000 (1982).
44. Maki, R., Roeder, W., Traunecker, A., Sidman, C., Wabl, M., Raschke, W. and Tonegawa, S. The role of DNA rearrangement and alternative RNA processing in the expression of immunoglobulin delta genes. Cell 24, 353-365 (1981).
45. Kerr, W.G., Hendershot, L.M. and Burrows, P.D. Regulation of IgM and IgD expression in human B-lineage cells. J Immunol 146, 3314-3321 (1991).
46. Lefranc, M.-P. and Lefranc, G. *Molecular genetics of immunoglobulin allotype expression, In: The human IgG subclasses*. F. Shakib, ed. P. Press. 1990, Oxford.
47. Padlan, E.A. Anatomy of the antibody molecule. Mol Immunol 31, 169-217 (1994).
48. Cory, S. and Adams, J.M. Deletions are associated with somatic rearrangement of immunoglobulin heavy chain genes. Cell 19, 37-51 (1980).
49. Honjo, T. and Kataoka, T. Organization of immunoglobulin heavy chain genes and allelic deletion model. Proc. Natl Acad. Sci. USA 75, 2140-2144 (1978).
50. Rabbitts, T.H., Forster, A., Dunnick, W. and Bentley, D.L. The role of gene deletion in the immunoglobulin heavy chain switch. Nature 283, 351-356 (1980).
51. Iwasato, T., Shimizu, A., Honjo, T. and Yamagishi, H. Circular DNA is excised by immunoglobulin class switch recombination. Cell 62, 143-149 (1990).

52. Matsuoka, M., Yoshida, K., Maeda, T., Usuda, S. and Sakano, H. Switch circular DNA formed in cytokine-treated mouse splenocytes: evidence for intramolecular DNA deletion in immunoglobulin class switching. Cell 62, 135-142 (1990).
53. Kluin, P.M., Kayano, H., Zani, V.J., Kluin-Nelemans, H.C., Tucker, P.W., Satterwhite, E. and Dyer, M.J. IgD class switching: identification of a novel recombination site in neoplastic and normal B cells. Eur J Immunol 25, 3504-3508 (1995).
54. Arpin, C., de Bouteiller, O., Razanajaona, D., Fugier-Vivier, I., Briere, F., Banchereau, J., Lebecque, S. and Liu, Y.J. The normal counterpart of IgD myeloma cells in germinal center displays extensively mutated IgVH gene, Cmu-Cdelta switch, and lambda light chain expression. J Exp Med 187, 1169-1178 (1998).
55. Liu, Y.J., de Bouteiller, O., Arpin, C., Briere, F., Galibert, L., Ho, S., Martinez-Valdez, H., Banchereau, J. and Lebecque, S. Normal human IgD+IgM- germinal center B cells can express up to 80 mutations in the variable region of their IgD transcripts. Immunity 4, 603-613 (1996).
56. Kehry, M., Ewald, S., Douglas, R., Sibley, C., Raschke, W., Fambrough, D. and Hood, L. The immunoglobulin mu chains of membrane-bound and secreted IgM molecules differ in their C-terminal segments. Cell 21, 393-406 (1980).
57. Rabbitts, T.H., Forster, A. and Milstein, C.P. Human immunoglobulin heavy chain genes: evolutionary comparisons of C mu, C delta and C gamma genes and associated switch sequences. Nucleic Acids Res. 9, 4509-4524 (1981).
58. Nelson, K.J., Haimovich, J. and Perry, R.P. Characterization of productive and sterile transcripts from the immunoglobulin heavy-chain locus: processing of micron and muS mRNA. Mol Cell Biol 3, 1317-1332 (1983).
59. Blattner, F.R. and Tucker, P.W. The molecular biology of immunoglobulin D. Nature 307, 417-422 (1984).
60. Hieter, P.A., Korsmeyer, S.J., Waldmann, T.A. and Leder, P. Human immunoglobulin kappa light-chain genes are deleted or rearranged in lambda-producing B cells. Nature 290, 368-372 (1981).
61. Korsmeyer, S.J., Hieter, P.A., Ravetch, J.V., Poplack, D.G., Waldmann, T.A. and Leder, P. Developmental hierarchy of immunoglobulin gene rearrangements in human leukemic pre-B-cells. Proc Natl Acad Sci U S A 78, 7096-7100 (1981).
62. Hollis, G.F., Evans, R.J., Stafford-Hollis, J.M., Korsmeyer, S.J. and McKearn, J.P. Immunoglobulin lambda light-chain-related genes 14.1 and 16.1 are expressed in pre-B cells and may encode the human immunoglobulin omega light-chain protein. Proc Natl Acad Sci U S A 86, 5552-5556 (1989).
63. Schiff, C., Bensmana, M., Guglielmi, P., Milili, M., Lefranc, M.P. and Fougereau, M. The immunoglobulin lambda-like gene cluster (14.1, 16.1 and F lambda 1) contains gene(s) selectively expressed in pre-B cells and is the human counterpart of the mouse lambda 5 gene. Int Immunol 2, 201-207 (1990).
64. Meffre, E., Casellas, R. and Nussenzweig, M.C. Antibody regulation of B cell development. Nat Immunol 1, 379-385 (2000).
65. Croce, C.M., Shander, M., Martinis, J., Cicurel, L., D'Ancona, G.G., Dolby, T.W. and Koprowski, H. Chromosomal location of the genes for human immunoglobulin heavy chains. Proc. Natl Acad. Sci. USA 76, 3416-3419 (1979).
66. Kirsch, I.R., Morton, C.C., Nakahara, K. and Leder, P. Human immunoglobulin heavy chain genes map to a region of translocations in malignant B lymphocytes. Science 216, 301-303 (1982).
67. McBride, O.W., Battey, J., Hollis, G.F., Swan, D.C., Siebenlist, U. and Leder, P. Localization of human variable and constant region immunoglobulin heavy chain

genes on subtelomeric band q32 of chromosome 14. Nucleic Acids Res. 10, 8155-8170 (1982).
68. Shin, E.K., Matsuda, F., Nagaoka, H., Fukita, Y., Imai, T., Yokoyama, K., Soeda, E. and Honjo, T. Physical map of the 3' region of the human immunoglobulin heavy chain locus: clustering of autoantibody-related variable segments in one haplotype. Embo J. 10, 3641-3645 (1991).
69. Matsuda, F., Shin, E.K., Nagaoka, H., Matsumura, R., Haino, M., Fukita, Y., Taka-ishi, S., Imai, T., Riley, J.H., Anand, R. and et al. Structure and physical map of 64 variable segments in the 3'0.8-megabase region of the human immunoglobulin heavy-chain locus. Nature Genetics 3, 88-94 (1993).
70. Cook, G.P., Tomlinson, I.M., Walter, G., Riethman, H., Carter, N.P., Buluwela, L., Winter, G. and Rabbitts, T.H. A map of the human immunoglobulin VH locus completed by analysis of the telomeric region of chromosome 14q. Nature Genetics 7, 162-168 (1994).
71. Cook, G.P. and Tomlinson, I.M. The human immunoglobulin VH repertoire. Immunol. Today 16, 237-242 (1995).
72. Matsuda, F., Ishii, K., Bourvagnet, P., Kuma, K., Hayashida, H., Miyata, T. and Honjo, T. The complete nucleotide sequence of the human immunoglobulin heavy chain variable region locus. J. Exp. Med. 188, 2151-2162 (1998).
73. Pallarès, N., Lefebvre, S., Contet, V., Matsuda, F. and Lefranc, M.P. The human immunoglobulin heavy variable genes. Exp. Clin. Immunogenet. 16, 36-60 (1999).
74. Siebenlist, U., Ravetch, J.V., Korsmeyer, S., Waldmann, T. and Leder, P. Human immunoglobulin D segments encoded in tandem multigenic families. Nature 294, 631-635 (1981).
75. Buluwela, L., Albertson, D.G., Sherrington, P., Rabbitts, P.H., Spurr, N. and Rabbitts, T.H. The use of chromosomal translocations to study human immunoglobulin gene organization: mapping DH segments within 35 kb of the C mu gene and identification of a new DH locus. Embo J. 7, 2003-2010 (1988).
76. Corbett, S.J., Tomlinson, I.M., Sonnhammer, E.L., Buck, D. and Winter, G. Sequence of the human immunoglobulin diversity (D) segment locus: a systematic analysis provides no evidence for the use of DIR segments, inverted D segments, "minor" D segments or D-D recombination. J. Mol. Biol. 270, 587-597 (1997).
77. Ruiz, M., Pallares, N., Contet, V., Barbi, V. and Lefranc, M.P. The human immunoglobulin heavy diversity (IGHD) and joining (IGHJ) segments. Exp. Clin. Immunogenet. 16, 173-184 (1999).
78. Ravetch, J.V., Siebenlist, U., Korsmeyer, S., Waldmann, T. and Leder, P. Structure of the human immunoglobulin mu locus: characterization of embryonic and rearranged J and D genes. Cell 27, 583-591 (1981).
79. Flanagan, J.G. and Rabbitts, T.H. Arrangement of human immunoglobulin heavy chain constant region genes implies evolutionary duplication of a segment containing gamma, epsilon and alpha genes. Nature 300, 709-713 (1982).
80. Lefranc, M.P., Lefranc, G. and Rabbitts, T.H. Inherited deletion of immunoglobulin heavy chain constant region genes in normal human individuals. Nature 300, 760-762 (1982).
81. Lefranc, M.P., Lefranc, G., de Lange, G., Out, T.A., van den Broek, P.J., van Nieuwkoop, J., Radl, J., Helal, A.N., Chaabani, H., van Loghem, E. and et al. Instability of the human immunoglobulin heavy chain constant region locus indicated by different inherited chromosomal deletions. Mol. Biol. Med. 1, 207-217 (1983).

82. White, M.B., Shen, A.L., Word, C.J., Tucker, P.W. and Blattner, F.R. Human immunoglobulin D: genomic sequence of the delta heavy chain. Science 228, 733-737 (1985).
83. Huck, S., Fort, P., Crawford, D.H., Lefranc, M.P. and Lefranc, G. Sequence of a human immunoglobulin gamma 3 heavy chain constant region gene: comparison with the other human C gamma genes. Nucleic Acids Res. 14, 1779-1789 (1986).
84. Huck, S., Lefranc, G. and Lefranc, M.P. A human immunoglobulin IGHG3 allele (Gmb0,b1,c3,c5,u) with an IGHG4 converted region and three hinge exons. Immunogenetics 30, 250-257 (1989).
85. Ellison, J. and Hood, L. Linkage and sequence homology of two human immunoglobulin gamma heavy chain constant region genes. Proc. Natl Acad. Sci. USA 79, 1984-1988 (1982).
86. Bensmana, M., Huck, S., Lefranc, G. and Lefranc, M.P. The human immunoglobulin pseudo-gamma IGHGP gene shows no major structural defect. Nucleic Acids Res. 16, 3108 (1988).
87. Ellison, J.W., Berson, B.J. and Hood, L.E. The nucleotide sequence of a human immunoglobulin C gamma1 gene. Nucleic Acids Res. 10, 4071-4079 (1982).
88. Ellison, J., Buxbaum, J. and Hood, L. Nucleotide sequence of a human immunoglobulin C gamma 4 gene. DNA 1, 11-18 (1981).
89. Flanagan, J.G., Lefranc, M.P. and Rabbitts, T.H. Mechanisms of divergence and convergence of the human immunoglobulin alpha 1 and alpha 2 constant region gene sequences. Cell 36, 681-688 (1984).
90. Malcolm, S., Barton, P., Murphy, C., Ferguson-Smith, M.A., Bentley, D.L. and Rabbitts, T.H. Localization of human immunoglobulin kappa light chain variable region genes to the short arm of chromosome 2 by in situ hybridization. Proc. Natl Acad. Sci. USA 79, 4957-4961 (1982).
91. McBride, O.W., Hieter, P.A., Hollis, G.F., Swan, D., Otey, M.C. and Leder, P. Chromosomal location of human kappa and lambda immunoglobulin light chain constant region genes. J. Exp. Med. 155, 1480-1490 (1982).
92. Scaviner, D., Barbie, V., Ruiz, M. and Lefranc, M.P. Protein displays of the human immunoglobulin heavy, kappa and lambda variable and joining regions. Exp. Clin. Immunogenet. 16, 234-240 (1999).
93. Zachau, H.G. The immunoglobulin kappa locus-or-what has been learned from looking closely at one-tenth of a percent of the human genome. Gene 135, 167-173 (1993).
94. Huber, C., Schable, K.F., Huber, E., Klein, R., Meindl, A., Thiebe, R., Lamm, R. and Zachau, H.G. The V kappa genes of the L regions and the repertoire of V kappa gene sequences in the human germ line. Eur. J. Immunol. 23, 2868-2875 (1993).
95. Schable, K.F. and Zachau, H.G. The variable genes of the human immunoglobulin kappa locus. Biol. Chem. Hoppe Seyler 374, 1001-1022 (1993).
96. Schable, K., Thiebe, R., Flugel, A., Meindl, A. and Zachau, H.G. The human immunoglobulin kappa locus: pseudogenes, unique and repetitive sequences. Biol. Chem. Hoppe Seyler 375, 189-199 (1994).
97. Cox, J.P., Tomlinson, I.M. and Winter, G. A directory of human germ-line V kappa segments reveals a strong bias in their usage. Eur. J. Immunol. 24, 827-836 (1994).
98. Barbié, V. and Lefranc, M.P. The human immunoglobulin kappa variable (IGKV) genes and joining (IGKJ) segments. Exp. Clin. Immunogenet. 15, 171-183 (1998).
99. Hieter, P.A., Maizel, J.V., Jr. and Leder, P. Evolution of human immunoglobulin kappa J region genes. J. Biol. Chem. 257, 1516-1522 (1982).

100. Hieter, P.A., Max, E.E., Seidman, J.G., Maizel, J.V., Jr. and Leder, P. Cloned human and mouse kappa immunoglobulin constant and J region genes conserve homology in functional segments. Cell 22, 197-207 (1980).
101. Erikson, J., Martinis, J. and Croce, C.M. Assignment of the genes for human lambda immunoglobulin chains to chromosome 22. Nature 294, 173-175 (1981).
102. Emanuel, B.S., Cannizzaro, L.A., Magrath, I., Tsujimoto, Y., Nowell, P.C. and Croce, C.M. Chromosomal orientation of the lambda light chain locus: V lambda is proximal to C lambda in 22q11. Nucleic Acids Res. 13, 381-387 (1985).
103. Frippiat, J.P., Williams, S.C., Tomlinson, I.M., Cook, G.P., Cherif, D., Le Paslier, D., Collins, J.E., Dunham, I., Winter, G. and Lefranc, M.P. Organization of the human immunoglobulin lambda light-chain locus on chromosome 22q11.2. Hum. Mol. Genet. 4, 983-991 (1995).
104. Kawasaki, K., Minoshima, S., Schooler, K., Kudoh, J., Asakawa, S., de Jong, P.J. and Shimizu, N. The organization of the human immunoglobulin lambda gene locus. Genome Res. 5, 125-135 (1995).
105. Williams, S.C., Frippiat, J.P., Tomlinson, I.M., Ignatovich, O., Lefranc, M.P. and Winter, G. Sequence and evolution of the human germline V lambda repertoire. J. Mol. Biol. 264, 220-232 (1996).
106. Kawasaki, K., Minoshima, S., Nakato, E., Shibuya, K., Shintani, A., Schmeits, J.L., Wang, J. and Shimizu, N. One-megabase sequence analysis of the human immunoglobulin lambda gene locus. Genome Res 7, 250-261 (1997).
107. Pallarès, N., Frippiat, J.P., Giudicelli, V. and Lefranc, M.P. The human immunoglobulin lambda variable (IGLV) genes and joining (IGLJ) segments. Exp. Clin. Immunogenet 15, 8-18 (1998).
108. Hieter, P.A., Hollis, G.F., Korsmeyer, S.J., Waldmann, T.A. and Leder, P. Clustered arrangement of immunoglobulin lambda constant region genes in man. Nature 294, 536-540 (1981).
109. Taub, R.A., Hollis, G.F., Hieter, P.A., Korsmeyer, S., Waldmann, T.A. and Leder, P. Variable amplification of immunoglobulin lambda light-chain genes in human populations. Nature 304, 172-174 (1983).
110. Dariavach, P., Lefranc, G. and Lefranc, M.P. Human immunoglobulin C lambda 6 gene encodes the Kern+Oz-lambda chain and C lambda 4 and C lambda 5 are pseudogenes. Proc. Natl Acad. Sci. USA 84, 9074-9078 (1987).
111. Vasicek, T.J. and Leder, P. Structure and expression of the human immunoglobulin lambda genes. J. Exp. Med. 172, 609-620 (1990).
112. Edelman, G.M., Cunningham, B.A., Gall, W.E., Gottlieb, P.D., Rutishauser, U. and Waxdal, M.J. The covalent structure of an entire gammaG immunoglobulin molecule. Proc Natl Acad Sci U S A 63, 78-85 (1969).
113. Lefranc, M.P. Unique database numbering system for immunogenetic analysis. Immunol Today 18, 509 (1997).
114. Lefranc, M.P. The IMGT unique numbering for Immunoglobulins, T cell receptors and Ig-like domains. The Immunologist 7, 132-136 (1999).
115. Ruiz, M. and Lefranc, M.-P. IMGT gene identification and Colliers de Perles of human immunoglobulins with known 3D structures. Immunogenetics 53, 857-883 (2002).
116. Lefranc, M.-P., Giudicelli, V., Regnier, L. and Duroux, P. IMGT, a system and an ontology that bridge biological and computational spheres in bioinformatics. Brief Bioinform 9, 263-275 (2008).
117. Chiorazzi, N., Rai, K.R. and Ferrarini, M. Chronic lymphocytic leukemia. N Engl J Med 352, 804-815 (2005).

118. Rozman, C. and Montserrat, E. Chronic lymphocytic leukemia. N Engl J Med 333, 1052-1057 (1995).
119. Morton, L.M., Wang, S.S., Devesa, S.S., Hartge, P., Weisenburger, D.D. and Linet, M.S. Lymphoma incidence patterns by WHO subtype in the United States, 1992-2001. Blood 107, 265-276 (2006).
120. Zent, C.S., Kyasa, M.J., Evans, R. and Schichman, S.A. Chronic lymphocytic leukemia incidence is substantially higher than estimated from tumor registry data. Cancer 92, 1325-1330 (2001).
121. de Lima, M., O'Brien, S., Lerner, S. and Keating, M.J. Chronic lymphocytic leukemia in the young patient. Semin Oncol 25, 107-116 (1998).
122. Weiss, N.S. Geographical variation in the incidence of the leukemias and lymphomas. Natl Cancer Inst Monogr, 139-142 (1979).
123. Boggs, D.R., Chen, S.C., Zhang, Z.N. and Zhang, A. Chronic lymphocytic leukemia in China. Am J Hematol 25, 349-354 (1987).
124. Linet, M.S., Van Natta, M.L., Brookmeyer, R., Khoury, M.J., McCaffrey, L.D., Humphrey, R.L. and Szklo, M. Familial cancer history and chronic lymphocytic leukemia. A case-control study. Am J Epidemiol 130, 655-664 (1989).
125. Neuland, C.Y., Blattner, W.A., Mann, D.L., Fraser, M.C., Tsai, S. and Strong, D.M. Familial chronic lymphocytic leukemia. J Natl Cancer Inst 71, 1143-1150 (1983).
126. Yuille, M.R., Matutes, E., Marossy, A., Hilditch, B., Catovsky, D. and Houlston, R.S. Familial chronic lymphocytic leukaemia: a survey and review of published studies. Br J Haematol 109, 794-799 (2000).
127. Capalbo, S., Trerotoli, P., Ciancio, A., Battista, C., Serio, G. and Liso, V. Increased risk of lymphoproliferative disorders in relatives of patients with B-cell chronic lymphocytic leukemia: relevance of the degree of familial linkage. Eur J Haematol 65, 114-117 (2000).
128. Wiernik, P.H., Ashwin, M., Hu, X.P., Paietta, E. and Brown, K. Anticipation in familial chronic lymphocytic leukaemia. Br J Haematol 113, 407-414 (2001).
129. Goldin, L.R., Sgambati, M., Marti, G.E., Fontaine, L., Ishibe, N. and Caporaso, N. Anticipation in familial chronic lymphocytic leukemia. Am J Hum Genet 65, 265-269 (1999).
130. Horwitz, M., Goode, E.L. and Jarvik, G.P. Anticipation in familial leukemia. Am J Hum Genet 59, 990-998 (1996).
131. Guipaud, O., Deriano, L., Salin, H., Vallat, L., Sabatier, L., Merle-Béral, H. and Delic, J. B-cell chronic lymphocytic leukaemia: a polymorphic family unified by genomic features. Lancet Oncol 4, 505-514 (2003).
132. Matutes, E. and Polliack, A. Morphological and immunophenotypic features of chronic lymphocytic leukemia. Rev Clin Exp Hematol 4, 22-47 (2000).
133. Matutes, E., Owusu-Ankomah, K., Morilla, R., Garcia Marco, J., Houlihan, A., Que, T.H. and Catovsky, D. The immunological profile of B-cell disorders and proposal of a scoring system for the diagnosis of CLL. Leukemia 8, 1640-1645 (1994).
134. Moreau, E.J., Matutes, E., A'Hern, R.P., Morilla, A.M., Morilla, R.M., Owusu-Ankomah, K.A., Seon, B.K. and Catovsky, D. Improvement of the chronic lymphocytic leukemia scoring system with the monoclonal antibody SN8 (CD79b). Am J Clin Pathol 108, 378-382 (1997).
135. Boumsell, L., Bernard, A., Lepage, V., Degos, L., Lemerle, J. and Dausset, J. Some chronic lymphocytic leukemia cells bearing surface immunoglobulins share determinants with T cells. Eur J Immunol 8, 900-904 (1978).
136. Cheson, B.D., Bennett, J.M., Grever, M., Kay, N., Keating, M.J., O'Brien, S. and Rai, K.R. National Cancer Institute-sponsored Working Group guidelines for chronic

lymphocytic leukemia: revised guidelines for diagnosis and treatment. Blood 87, 4990-4997 (1996).
137. Deaglio, S., Vaisitti, T., Aydin, S., Ferrero, E. and Malavasi, F. In-tandem insight from basic science combined with clinical research: CD38 as both marker and key component of the pathogenetic network underlying chronic lymphocytic leukemia. Blood 108, 1135-1144 (2006).
138. Del Poeta, G., Maurillo, L., Venditti, A., Buccisano, F., Epiceno, A.M., Capelli, G., Tamburini, A., Suppo, G., Battaglia, A., Del Principe, M.I., Del Moro, B., Masi, M. and Amadori, S. Clinical significance of CD38 expression in chronic lymphocytic leukemia. Blood 98, 2633-2639 (2001).
139. Ibrahim, S., Keating, M., Do, K.A., O'Brien, S., Huh, Y.O., Jilani, I., Lerner, S., Kantarjian, H.M. and Albitar, M. CD38 expression as an important prognostic factor in B-cell chronic lymphocytic leukemia. Blood 98, 181-186 (2001).
140. Morabito, F., Mangiola, M., Oliva, B., Stelitano, C., Callea, V., Deaglio, S., Iacopino, P., Brugiatelli, M. and Malavasi, F. Peripheral blood CD38 expression predicts survival in B-cell chronic lymphocytic leukemia. Leuk Res 25, 927-932 (2001).
141. Binet, J.-L., Caligaris-Cappio, F., Catovsky, D., Cheson, B., Davis, T., Dighiero, G., Döhner, H., Hallek, M., Hillmen, P., Keating, M., Montserrat, E., Kipps, T.J. and Rai, K. Perspectives on the use of new diagnostic tools in the treatment of chronic lymphocytic leukemia. Blood 107, 859-861 (2006).
142. Korz, C., Pscherer, A., Benner, A., Mertens, D., Schaffner, C., Leupolt, E., Döhner, H., Stilgenbauer, S. and Lichter, P. Evidence for distinct pathomechanisms in B-cell chronic lymphocytic leukemia and mantle cell lymphoma by quantitative expression analysis of cell cycle and apoptosis-associated genes. Blood 99, 4554-4561 (2002).
143. Green, D.R. and Kroemer, G. The pathophysiology of mitochondrial cell death. Science 305, 626-629 (2004).
144. Kirkin, V., Joos, S. and Zörnig, M. The role of Bcl-2 family members in tumorigenesis. Biochim Biophys Acta 1644, 229-249 (2004).
145. Hanada, M., Delia, D., Aiello, A., Stadtmauer, E. and Reed, J.C. bcl-2 gene hypomethylation and high-level expression in B-cell chronic lymphocytic leukemia. Blood 82, 1820-1828 (1993).
146. Calin, G.A., Ferracin, M., Cimmino, A., Di Leva, G., Shimizu, M., Wojcik, S.E., Iorio, M.V., Visone, R., Sever, N.I., Fabbri, M., Iuliano, R., Palumbo, T., Pichiorri, F., Roldo, C., Garzon, R., Sevignani, C., Rassenti, L., Alder, H., Volinia, S., Liu, C.-g., Kipps, T.J., Negrini, M. and Croce, C.M. A MicroRNA signature associated with prognosis and progression in chronic lymphocytic leukemia. N Engl J Med 353, 1793-1801 (2005).
147. Cimmino, A., Calin, G.A., Fabbri, M., Iorio, M.V., Ferracin, M., Shimizu, M., Wojcik, S.E., Aqeilan, R.I., Zupo, S., Dono, M., Rassenti, L., Alder, H., Volinia, S., Liu, C.-G., Kipps, T.J., Negrini, M. and Croce, C.M. miR-15 and miR-16 induce apoptosis by targeting BCL2. Proc Natl Acad Sci U S A 102, 13944-13949 (2005).
148. Bouley, J., Deriano, L., Delic, J. and Merle-Beral, H. New molecular markers in resistant B-CLL. Leuk Lymphoma 47, 791-801 (2006).
149. Trbusek, M., Malcikova, J., Smardova, J., Kuhrova, V., Mentzlova, D., Francova, H., Bukovska, S., Svitakova, M., Kuglik, P., Linkova, V., Doubek, M., Brychtova, Y., Zacal, J., Kujickova, J., Pospisilova, S., Dvorakova, D., Vorlicek, J. and Mayer, J. Inactivation of p53 and deletion of ATM in B-CLL patients in relation to IgVH mutation status and previous treatment. Leukemia 20, 1159-1161 (2006).
150. Messmer, B.T., Messmer, D., Allen, S.L., Kolitz, J.E., Kudalkar, P., Cesar, D., Murphy, E.J., Koduru, P., Ferrarini, M., Zupo, S., Cutrona, G., Damle, R.N., Wasil,

T., Rai, K.R., Hellerstein, M.K. and Chiorazzi, N. In vivo measurements document the dynamic cellular kinetics of chronic lymphocytic leukemia B cells. J Clin Invest 115, 755-764 (2005).
151. Chiorazzi, N. and Ferrarini, M. B cell chronic lymphocytic leukemia: lessons learned from studies of the B cell antigen receptor. Annu Rev Immunol 21, 841-894 (2003).
152. Stevenson, F.K. and Caligaris-Cappio, F. Chronic lymphocytic leukemia: revelations from the B-cell receptor. Blood 103, 4389-4395 (2004).
153. Cavalli, F., Isaacson, P.G., Gascoyne, R.D. and Zucca, E. MALT Lymphomas. Hematology Am Soc Hematol Educ Program, 241-258 (2001).
154. Bröker, B.M., Klajman, A., Youinou, P., Jouquan, J., Worman, C.P., Murphy, J., Mackenzie, L., Quartey-Papafio, R., Blaschek, M., Collins, P. and et al. Chronic lymphocytic leukemic (CLL) cells secrete multispecific autoantibodies. J Autoimmun 1, 469-481 (1988).
155. Sthoeger, Z.M., Wakai, M., Tse, D.B., Vinciguerra, V.P., Allen, S.L., Budman, D.R., Lichtman, S.M., Schulman, P., Weiselberg, L.R. and Chiorazzi, N. Production of autoantibodies by CD5-expressing B lymphocytes from patients with chronic lymphocytic leukemia. J Exp Med 169, 255-268 (1989).
156. Borche, L., Lim, A., Binet, J.L. and Dighiero, G. Evidence that chronic lymphocytic leukemia B lymphocytes are frequently committed to production of natural autoantibodies. Blood 76, 562-569 (1990).
157. Schwartz, R.S. and Stollar, B.D. Heavy-chain directed B-cell maturation: continuous clonal selection beginning at the pre-B cell stage. Immunol Today 15, 27-32 (1994).
158. Michel, F., Merle-Béral, H., Legac, E., Michel, A., Debré, P. and Bismuth, G. Defective calcium response in B-chronic lymphocytic leukemia cells. Alteration of early protein tyrosine phosphorylation and of the mechanism responsible for cell calcium influx. J Immunol 150, 3624-3633 (1993).
159. Lankester, A.C., van Schijndel, G.M., van der Schoot, C.E., van Oers, M.H., van Noesel, C.J. and van Lier, R.A. Antigen receptor nonresponsiveness in chronic lymphocytic leukemia B cells. Blood 86, 1090-1097 (1995).
160. Zupo, S., Cutrona, G., Mangiola, M. and Ferrarini, M. Role of surface IgM and IgD on survival of the cells from B-cell chronic lymphocytic leukemia. Blood 99, 2277-2278 (2002).
161. Zupo, S., Isnardi, L., Megna, M., Massara, R., Malavasi, F., Dono, M., Cosulich, E. and Ferrarini, M. CD38 expression distinguishes two groups of B-cell chronic lymphocytic leukemias with different responses to anti-IgM antibodies and propensity to apoptosis. Blood 88, 1365-1374 (1996).
162. Chen, L., Widhopf, G., Huynh, L., Rassenti, L., Rai, K.R., Weiss, A. and Kipps, T.J. Expression of ZAP-70 is associated with increased B-cell receptor signaling in chronic lymphocytic leukemia. Blood 100, 4609-4614 (2002).
163. Lanham, S., Hamblin, T., Oscier, D., Ibbotson, R., Stevenson, F. and Packham, G. Differential signaling via surface IgM is associated with VH gene mutational status and CD38 expression in chronic lymphocytic leukemia. Blood 101, 1087-1093 (2003).
164. Cragg, M.S., Chan, H.T.C., Fox, M.D., Tutt, A., Smith, A., Oscier, D.G., Hamblin, T.J. and Glennie, M.J. The alternative transcript of CD79b is overexpressed in B-CLL and inhibits signaling for apoptosis. Blood 100, 3068-3076 (2002).
165. Vuillier, F., Dumas, G., Magnac, C., Prevost, M.-C., Lalanne, A.I., Oppezzo, P., Melanitou, E., Dighiero, G. and Payelle-Brogard, B. Lower levels of surface B-cell-receptor expression in chronic lymphocytic leukemia are associated with glycosylation and folding defects of the mu and CD79a chains. Blood 105, 2933-2940 (2005).

166. Payelle-Brogard, B., Magnac, C., Alcover, A., Roux, P. and Dighiero, G. Defective assembly of the B-cell receptor chains accounts for its low expression in B-chronic lymphocytic leukaemia. Br J Haematol 118, 976-985 (2002).
167. Bernal, A., Pastore, R.D., Asgary, Z., Keller, S.A., Cesarman, E., Liou, H.C. and Schattner, E.J. Survival of leukemic B cells promoted by engagement of the antigen receptor. Blood 98, 3050-3057 (2001).
168. Zupo, S., Massara, R., Dono, M., Rossi, E., Malavasi, F., Cosulich, M.E. and Ferrarini, M. Apoptosis or plasma cell differentiation of CD38-positive B-chronic lymphocytic leukemia cells induced by cross-linking of surface IgM or IgD. Blood 95, 1199-1206 (2000).
169. Caligaris-Cappio, F. Role of the microenvironment in chronic lymphocytic leukaemia. Br J Haematol 123, 380-388 (2003).
170. Panayiotidis, P., Jones, D., Ganeshaguru, K., Foroni, L. and Hoffbrand, A.V. Human bone marrow stromal cells prevent apoptosis and support the survival of chronic lymphocytic leukaemia cells in vitro. Br J Haematol 92, 97-103 (1996).
171. Tsukada, N., Burger, J.A., Zvaifler, N.J. and Kipps, T.J. Distinctive features of "nurselike" cells that differentiate in the context of chronic lymphocytic leukemia. Blood 99, 1030-1037 (2002).
172. Burger, J.A., Burger, M. and Kipps, T.J. Chronic lymphocytic leukemia B cells express functional CXCR4 chemokine receptors that mediate spontaneous migration beneath bone marrow stromal cells. Blood 94, 3658-3667 (1999).
173. Deaglio, S., Vaisitti, T., Bergui, L., Bonello, L., Horenstein, A.L., Tamagnone, L., Boumsell, L. and Malavasi, F. CD38 and CD100 lead a network of surface receptors relaying positive signals for B-CLL growth and survival. Blood 105, 3042-3050 (2005).
174. Chen, H., Treweeke, A.T., West, D.C., Till, K.J., Cawley, J.C., Zuzel, M. and Toh, C.H. In vitro and in vivo production of vascular endothelial growth factor by chronic lymphocytic leukemia cells. Blood 96, 3181-3187 (2000).
175. Lee, Y.K., Bone, N.D., Strege, A.K., Shanafelt, T.D., Jelinek, D.F. and Kay, N.E. VEGF receptor phosphorylation status and apoptosis is modulated by a green tea component, epigallocatechin-3-gallate (EGCG), in B-cell chronic lymphocytic leukemia. Blood 104, 788-794 (2004).
176. Granziero, L., Ghia, P., Circosta, P., Gottardi, D., Strola, G., Geuna, M., Montagna, L., Piccoli, P., Chilosi, M. and Caligaris-Cappio, F. Survivin is expressed on CD40 stimulation and interfaces proliferation and apoptosis in B-cell chronic lymphocytic leukemia. Blood 97, 2777-2783 (2001).
177. Pedersen, I.M., Kitada, S., Leoni, L.M., Zapata, J.M., Karras, J.G., Tsukada, N., Kipps, T.J., Choi, Y.S., Bennett, F. and Reed, J.C. Protection of CLL B cells by a follicular dendritic cell line is dependent on induction of Mcl-1. Blood 100, 1795-17801 (2002).
178. Hamblin, T.J., Davis, Z., Gardiner, A., Oscier, D.G. and Stevenson, F.K. Unmutated Ig V(H) genes are associated with a more aggressive form of chronic lymphocytic leukemia. Blood 94, 1848-1854 (1999).
179. Damle, R.N., Wasil, T., Fais, F., Ghiotto, F., Valetto, A., Allen, S.L., Buchbinder, A., Budman, D., Dittmar, K., Kolitz, J., Lichtman, S.M., Schulman, P., Vinciguerra, V.P., Rai, K.R., Ferrarini, M. and Chiorazzi, N. Ig V gene mutation status and CD38 expression as novel prognostic indicators in chronic lymphocytic leukemia. Blood 94, 1840-1847 (1999).
180. Ghia, P., Stamatopoulos, K., Belessi, C., Moreno, C., Stella, S., Guida, G., Michel, A., Crespo, M., Laoutaris, N., Montserrat, E., Anagnostopoulos, A., Dighiero, G., Fassas,

A., Caligaris-Cappio, F. and Davi, F. Geographic patterns and pathogenetic implications of IGHV gene usage in chronic lymphocytic leukemia: the lesson of the IGHV3-21 gene. Blood 105, 1678-1685 (2005).
181. Stamatopoulos, K., Belessi, C., Hadzidimitriou, A., Smilevska, T., Kalagiakou, E., Hatzi, K., Stavroyianni, N., Athanasiadou, A., Tsompanakou, A., Papadaki, T., Kokkini, G., Paterakis, G., Saloum, R., Laoutaris, N., Anagnostopoulos, A. and Fassas, A. Immunoglobulin light chain repertoire in chronic lymphocytic leukemia. Blood 106, 3575-3583 (2005).
182. Widhopf, G.F., 2nd, Rassenti, L.Z., Toy, T.L., Gribben, J.G., Wierda, W.G. and Kipps, T.J. Chronic lymphocytic leukemia B cells of more than 1% of patients express virtually identical immunoglobulins. Blood 104, 2499-2504 (2004).
183. Tobin, G., Thunberg, U., Johnson, A., Eriksson, I., Söderberg, O., Karlsson, K., Merup, M., Juliusson, G., Vilpo, J., Enblad, G., Sundström, C., Roos, G. and Rosenquist, R. Chronic lymphocytic leukemias utilizing the VH3-21 gene display highly restricted Vlambda2-14 gene use and homologous CDR3s: implicating recognition of a common antigen epitope. Blood 101, 4952-4957 (2003).
184. Thorselius, M., Krober, A., Murray, F., Thunberg, U., Tobin, G., Buhler, A., Kienle, D., Albesiano, E., Maffei, R., Dao-Ung, L.P., Wiley, J., Vilpo, J., Laurell, A., Merup, M., Roos, G., Karlsson, K., Chiorazzi, N., Marasca, R., Dohner, H., Stilgenbauer, S. and Rosenquist, R. Strikingly homologous immunoglobulin gene rearrangements and poor outcome in VH3-21-using chronic lymphocytic leukemia patients independent of geographic origin and mutational status. Blood 107, 2889-2894 (2006).
185. Murray, F., Darzentas, N., Hadzidimitriou, A., Tobin, G., Boudjogra, M., Scielzo, C., Laoutaris, N., Karlsson, K., Baran-Marzsak, F., Tsaftaris, A., Moreno, C., Anagnostopoulos, A., Caligaris-Cappio, F., Vaur, D., Ouzounis, C., Belessi, C., Ghia, P., Davi, F., Rosenquist, R. and Stamatopoulos, K. Stereotyped patterns of somatic hypermutation in subsets of patients with chronic lymphocytic leukemia: implications for the role of antigen selection in leukemogenesis. Blood 111, 1524-1533 (2008).
186. Tobin, G., Thunberg, U., Karlsson, K., Murray, F., Laurell, A., Willander, K., Enblad, G., Merup, M., Vilpo, J., Juliusson, G., Sundstrom, C., Soderberg, O., Roos, G. and Rosenquist, R. Subsets with restricted immunoglobulin gene rearrangement features indicate a role for antigen selection in the development of chronic lymphocytic leukemia. Blood 104, 2879-2885 (2004).
187. Hamblin, T.J., Orchard, J.A., Ibbotson, R.E., Davis, Z., Thomas, P.W., Stevenson, F.K. and Oscier, D.G. CD38 expression and immunoglobulin variable region mutations are independent prognostic variables in chronic lymphocytic leukemia, but CD38 expression may vary during the course of the disease. Blood 99, 1023-1029 (2002).
188. Dighiero, G. and Binet, J.L. When and how to treat chronic lymphocytic leukemia. N Engl J Med 343, 1799-1801 (2000).
189. Rai, K.R., Sawitsky, A., Cronkite, E.P., Chanana, A.D., Levy, R.N. and Pasternack, B.S. Clinical staging of chronic lymphocytic leukemia. Blood 46, 219-234 (1975).
190. Binet, J.L., Auquier, A., Dighiero, G., Chastang, C., Piguet, H., Goasguen, J., Vaugier, G., Potron, G., Colona, P., Oberling, F., Thomas, M., Tchernia, G., Jacquillat, C., Boivin, P., Lesty, C., Duault, M.T., Monconduit, M., Belabbes, S. and Gremy, F. A new prognostic classification of chronic lymphocytic leukemia derived from a multivariate survival analysis. Cancer 48, 198-206 (1981).
191. Montserrat, E., Sanchez-Bisono, J., Viñolas, N. and Rozman, C. Lymphocyte doubling time in chronic lymphocytic leukaemia: analysis of its prognostic significance. Br J Haematol 62, 567-575 (1986).

192. Molica, S., Reverter, J.C., Alberti, A. and Montserrat, E. Timing of diagnosis and lymphocyte accumulation patterns in chronic lymphocytic leukemia: analysis of their clinical significance. Eur J Haematol 44, 277-281 (1990).
193. Viñolas, N., Reverter, J.C., Urbano-Ispizua, A., Montserrat, E. and Rozman, C. Lymphocyte doubling time in chronic lymphocytic leukemia: an update of its prognostic significance. Blood Cells 12, 457-470 (1987).
194. Sarfati, M., Chevret, S., Chastang, C., Biron, G., Stryckmans, P., Delespesse, G., Binet, J.L., Merle-Beral, H. and Bron, D. Prognostic importance of serum soluble CD23 level in chronic lymphocytic leukemia. Blood 88, 4259-4264 (1996).
195. Molica, S., Levato, D., Dell'Olio, M., Matera, R., Minervini, M., Dattilo, A., Carotenuto, M. and Musto, P. Cellular expression and serum circulating levels of CD23 in B-cell chronic lymphocytic leukemia. Implications for prognosis. Haematologica 81, 428-433 (1996).
196. Gaidano, G., Ballerini, P., Gong, J.Z., Inghirami, G., Neri, A., Newcomb, E.W., Magrath, I.T., Knowles, D.M. and Dalla-Favera, R. p53 mutations in human lymphoid malignancies: association with Burkitt lymphoma and chronic lymphocytic leukemia. Proc Natl Acad Sci U S A 88, 5413-5417 (1991).
197. Keating, M.J. Chronic lymphocytic leukemia. Semin Oncol 26, 107-114 (1999).
198. Hallek, M., Langenmayer, I., Nerl, C., Knauf, W., Dietzfelbinger, H., Adorf, D., Ostwald, M., Busch, R., Kuhn-Hallek, I., Thiel, E. and Emmerich, B. Elevated serum thymidine kinase levels identify a subgroup at high risk of disease progression in early, nonsmoldering chronic lymphocytic leukemia. Blood 93, 1732-1737 (1999).
199. Kallander, C.F., Simonsson, B., Hagberg, H. and Gronowitz, J.S. Serum deoxythymidine kinase gives prognostic information in chronic lymphocytic leukemia. Cancer 54, 2450-2455 (1984).
200. Davi, F., Rosenquist, R., Ghia, P., Belessi, C. and Stamatopoulos, K. Determination of IGHV gene mutational status in chronic lymphocytic leukemia: bioinformatics advances meet clinical needs. Leukemia 22, 212-214 (2008).
201. Ghia, P., Ferreri, A.M. and Galigaris-Cappio, F. Chronic lymphocytic leukemia. Crit Rev Oncol Hematol 64, 234-246 (2007).
202. Tobin, G., Thunberg, U., Laurell, A., Karlsson, K., Aleskog, A., Willander, K., Soderberg, O., Merup, M., Vilpo, J., Hultdin, M., Sundstrom, C., Roos, G. and Rosenquist, R. Patients with chronic lymphocytic leukemia with mutated VH genes presenting with Binet stage B or C form a subgroup with a poor outcome. Haematologica 90, 465-469 (2005).
203. Tobin, G., Thunberg, U., Johnson, A., Thörn, I., Söderberg, O., Hultdin, M., Botling, J., Enblad, G., Sällström, J., Sundström, C., Roos, G. and Rosenquist, R. Somatically mutated Ig V(H)3-21 genes characterize a new subset of chronic lymphocytic leukemia. Blood 99, 2262-2264 (2002).
204. Wiestner, A., Rosenwald, A., Barry, T.S., Wright, G., Davis, R.E., Henrickson, S.E., Zhao, H., Ibbotson, R.E., Orchard, J.A., Davis, Z., Stetler-Stevenson, M., Raffeld, M., Arthur, D.C., Marti, G.E., Wilson, W.H., Hamblin, T.J., Oscier, D.G. and Staudt, L.M. ZAP-70 expression identifies a chronic lymphocytic leukemia subtype with unmutated immunoglobulin genes, inferior clinical outcome, and distinct gene expression profile. Blood 101, 4944-4951 (2003).
205. Crespo, M., Bosch, F., Villamor, N., Bellosillo, B., Colomer, D., Rozman, M., Marcé, S., López-Guillermo, A., Campo, E. and Montserrat, E. ZAP-70 expression as a surrogate for immunoglobulin-variable-region mutations in chronic lymphocytic leukemia. N Engl J Med 348, 1764-1775 (2003).

206. Orchard, J.A., Ibbotson, R.E., Davis, Z., Wiestner, A., Rosenwald, A., Thomas, P.W., Hamblin, T.J., Staudt, L.M. and Oscier, D.G. ZAP-70 expression and prognosis in chronic lymphocytic leukaemia. Lancet 363, 105-111 (2004).
207. Rassenti, L.Z., Huynh, L., Toy, T.L., Chen, L., Keating, M.J., Gribben, J.G., Neuberg, D.S., Flinn, I.W., Rai, K.R., Byrd, J.C., Kay, N.E., Greaves, A., Weiss, A. and Kipps, T.J. ZAP-70 compared with immunoglobulin heavy-chain gene mutation status as a predictor of disease progression in chronic lymphocytic leukemia. N Engl J Med 351, 893-901 (2004).
208. Oppezzo, P., Vasconcelos, Y., Settegrana, C., Jeannel, D., Vuillier, F., Legarff-Tavernier, M., Kimura, E.Y., Bechet, S., Dumas, G., Brissard, M., Merle-Béral, H., Yamamoto, M., Dighiero, G. and Davi, F. The LPL/ADAM29 expression ratio is a novel prognosis indicator in chronic lymphocytic leukemia. Blood 106, 650-657 (2005).
209. Le Garff-Tavernier, M., Ticchioni, M., Brissard, M., Salmon, C., Raynaud, S., Davi, F., Bernard, A., Merle-Beral, H., Ajchenbaum-Cymbalista, F. and Letestu, R. National standardization of ZAP-70 determination by flow cytometry: the French experience. Cytometry B Clin Cytom 72, 103-108 (2007).
210. Letestu, R., Rawstron, A., Ghia, P., Villamor, N., Boeckx, N., Boettcher, S., Buhl, A.M., Duerig, J., Ibbotson, R., Kroeber, A., Langerak, A., Le Garff-Tavernier, M., Mockridge, I., Morilla, A., Padmore, R., Rassenti, L., Ritgen, M., Shehata, M., Smolewski, P., Staib, P., Ticchioni, M., Walker, C. and Ajchenbaum-Cymbalista, F. Evaluation of ZAP-70 expression by flow cytometry in chronic lymphocytic leukemia: A multicentric international harmonization process. Cytometry B Clin Cytom 70, 309-314 (2006).
211. Ghia, P., Guida, G., Scielzo, C., Geuna, M. and Caligaris-Cappio, F. CD38 modifications in chronic lymphocytic leukemia: are they relevant? Leukemia 18, 1733-1735 (2004).
212. Ghia, P., Guida, G., Stella, S., Gottardi, D., Geuna, M., Strola, G., Scielzo, C. and Caligaris-Cappio, F. The pattern of CD38 expression defines a distinct subset of chronic lymphocytic leukemia (CLL) patients at risk of disease progression. Blood 101, 1262-1269 (2003).
213. Döhner, H., Stilgenbauer, S., Benner, A., Leupolt, E., Kröber, A., Bullinger, L., Döhner, K., Bentz, M. and Lichter, P. Genomic aberrations and survival in chronic lymphocytic leukemia. N Engl J Med 343, 1910-1916 (2000).
214. Austen, B., Powell, J.E., Alvi, A., Edwards, I., Hooper, L., Starczynski, J., Taylor, A.M.R., Fegan, C., Moss, P. and Stankovic, T. Mutations in the ATM gene lead to impaired overall and treatment-free survival that is independent of IGVH mutation status in patients with B-CLL. Blood 106, 3175-3182 (2005).
215. Stankovic, T., Weber, P., Stewart, G., Bedenham, T., Murray, J., Byrd, P.J., Moss, P.A. and Taylor, A.M. Inactivation of ataxia telangiectasia mutated gene in B-cell chronic lymphocytic leukaemia. Lancet 353, 26-29 (1999).
216. Geisler, C.H., Philip, P., Christensen, B.E., Hou-Jensen, K., Pedersen, N.T., Jensen, O.M., Thorling, K., Andersen, E., Birgens, H.S., Drivsholm, A., Ellegaard, J., Larsen, J.K., Plesner, T., Brown, P., Andersen, P.K. and Hansen, M.M. In B-cell chronic lymphocytic leukaemia chromosome 17 abnormalities and not trisomy 12 are the single most important cytogenetic abnormalities for the prognosis: a cytogenetic and immunophenotypic study of 480 unselected newly diagnosed patients. Leuk Res 21, 1011-1023 (1997).
217. Kröber, A., Seiler, T., Benner, A., Bullinger, L., Brückle, E., Lichter, P., Döhner, H. and Stilgenbauer, S. V(H) mutation status, CD38 expression level, genomic

aberrations, and survival in chronic lymphocytic leukemia. Blood 100, 1410-1416 (2002).
218. Keating, M.J., O'Brien, S., Lerner, S., Koller, C., Beran, M., Robertson, L.E., Freireich, E.J., Estey, E. and Kantarjian, H. Long-term follow-up of patients with chronic lymphocytic leukemia (CLL) receiving fludarabine regimens as initial therapy. Blood 92, 1165-1171 (1998).
219. Rai, K.R., Peterson, B.L., Appelbaum, F.R., Kolitz, J., Elias, L., Shepherd, L., Hines, J., Threatte, G.A., Larson, R.A., Cheson, B.D. and Schiffer, C.A. Fludarabine compared with chlorambucil as primary therapy for chronic lymphocytic leukemia. N Engl J Med 343, 1750-1757 (2000).
220. Leporrier, M., Chevret, S., Cazin, B., Boudjerra, N., Feugier, P., Desablens, B., Rapp, M.J., Jaubert, J., Autrand, C., Divine, M., Dreyfus, B., Maloum, K., Travade, P., Dighiero, G., Binet, J.L. and Chastang, C. Randomized comparison of fludarabine, CAP, and ChOP in 938 previously untreated stage B and C chronic lymphocytic leukemia patients. Blood 98, 2319-2325 (2001).
221. Yamauchi, T., Nowak, B.J., Keating, M.J. and Plunkett, W. DNA repair initiated in chronic lymphocytic leukemia lymphocytes by 4-hydroperoxycyclophosphamide is inhibited by fludarabine and clofarabine. Clin Cancer Res 7, 3580-3589 (2001).
222. O'Brien, S.M., Kantarjian, H.M., Cortes, J., Beran, M., Koller, C.A., Giles, F.J., Lerner, S. and Keating, M. Results of the fludarabine and cyclophosphamide combination regimen in chronic lymphocytic leukemia. J Clin Oncol 19, 1414-1420 (2001).
223. Flinn, I.W., Byrd, J.C., Morrison, C., Jamison, J., Diehl, L.F., Murphy, T., Piantadosi, S., Seifter, E., Ambinder, R.F., Vogelsang, G. and Grever, M.R. Fludarabine and cyclophosphamide with filgrastim support in patients with previously untreated indolent lymphoid malignancies. Blood 96, 71-75 (2000).
224. Eichhorst, B.F., Busch, R., Hopfinger, G., Pasold, R., Hensel, M., Steinbrecher, C., Siehl, S., Jäger, U., Bergmann, M., Stilgenbauer, S., Schweighofer, C., Wendtner, C.M., Döhner, H., Brittinger, G., Emmerich, B. and Hallek, M. Fludarabine plus cyclophosphamide versus fludarabine alone in first-line therapy of younger patients with chronic lymphocytic leukemia. Blood 107, 885-891 (2006).
225. O'Brien, S.M., Kantarjian, H., Thomas, D.A., Giles, F.J., Freireich, E.J., Cortes, J., Lerner, S. and Keating, M.J. Rituximab dose-escalation trial in chronic lymphocytic leukemia. J Clin Oncol 19, 2165-2170 (2001).
226. Wierda, W., O'Brien, S., Wen, S., Faderl, S., Garcia-Manero, G., Thomas, D., Do, K.-A., Cortes, J., Koller, C., Beran, M., Ferrajoli, A., Giles, F., Lerner, S., Albitar, M., Kantarjian, H. and Keating, M. Chemoimmunotherapy with fludarabine, cyclophosphamide, and rituximab for relapsed and refractory chronic lymphocytic leukemia. J Clin Oncol 23, 4070-4078 (2005).
227. Sutton, L., Maloum, K., Gonzalez, H., Zouabi, H., Azar, N., Boccaccio, C., Charlotte, F., Cosset, J.M., Gabarre, J., Leblond, V., Merle-Beral, H. and Binet, J.L. Autologous hematopoietic stem cell transplantation as salvage treatment for advanced B cell chronic lymphocytic leukemia. Leukemia 12, 1699-1707 (1998).
228. Pavletic, Z.S., Bierman, P.J., Vose, J.M., Bishop, M.R., Wu, C.D., Pierson, J.L., Kollath, J.P., Weisenburger, D.D., Kessinger, A. and Armitage, J.O. High incidence of relapse after autologous stem-cell transplantation for B-cell chronic lymphocytic leukemia or small lymphocytic lymphoma. Ann Oncol 9, 1023-1026 (1998).
229. Montserrat, E. Role of auto- and allotransplantation in B-cell chronic lymphocytic leukemia. Hematol Oncol Clin North Am 18, 915-926, x (2004).

230. Michallet, M., Archimbaud, E., Bandini, G., Rowlings, P.A., Deeg, H.J., Gahrton, G., Montserrat, E., Rozman, C., Gratwohl, A. and Gale, R.P. HLA-identical sibling bone marrow transplantation in younger patients with chronic lymphocytic leukemia. European Group for Blood and Marrow Transplantation and the International Bone Marrow Transplant Registry. Ann Intern Med 124, 311-315 (1996).
231. Doney, K.C., Chauncey, T. and Appelbaum, F.R. Allogeneic related donor hematopoietic stem cell transplantation for treatment of chronic lymphocytic leukemia. Bone Marrow Transplant 29, 817-823 (2002).
232. Rondón, G., Giralt, S., Huh, Y., Khouri, I., Andersson, B., Andreeff, M. and Champlin, R. Graft-versus-leukemia effect after allogeneic bone marrow transplantation for chronic lymphocytic leukemia. Bone Marrow Transplant 18, 669-672 (1996).
233. Moreno, C., Villamor, N., Colomer, D., Esteve, J., Martino, R., Nomdedeu, J., Bosch, F., Lopez-Guillermo, A., Campo, E., Sierra, J. and Montserrat, E. Allogeneic stem-cell transplantation may overcome the adverse prognosis of unmutated VH gene in patients with chronic lymphocytic leukemia. J Clin Oncol 23, 3433-3438 (2005).
234. Dreger, P., Corradini, P., Kimby, E., Michallet, M., Milligan, D., Schetelig, J., Wiktor-Jedrzejczak, W., Niederwieser, D., Hallek, M. and Montserrat, E. *Indications for allogeneic stem cell transplantation in chronic lymphocytic leukemia: the EBMT transplant consensus*. 2007. Department of Medicine V, University of Heidelberg, Heidelberg, Germany. peter.dreger@med.uni-heidelberg.de.
235. Giudicelli, V., Chaume, D. and Lefranc, M.-P. IMGT/V-QUEST, an integrated software program for immunoglobulin and T cell receptor V-J and V-D-J rearrangement analysis. Nucleic Acids Res 32, W435-W440 (2004).
236. Brochet, X., Lefranc, M.-P. and Giudicelli, V. IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. Nucleic Acids Res 36, W503-W508 (2008).
237. Giudicelli, V., Chaume, D. and Lefranc, M.-P. IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. Nucleic Acids Res 33, D256-D261 (2005).
238. Yousfi Monod, M., Giudicelli, V., Chaume, D. and Lefranc, M.-P. IMGT/JunctionAnalysis: the first tool for the analysis of the immunoglobulin and T cell receptor complex V-J and V-D-J JUNCTIONs. Bioinformatics 20 Suppl 1, i379-i385 (2004).
239. Cornish-Bowden, A. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. Nucleic Acids Res 13, 3021-3030 (1985).
240. Wain, H.M., Bruford, E.A., Lovering, R.C., Lush, M.J., Wright, M.W. and Povey, S. Guidelines for human gene nomenclature. Genomics 79, 464-470 (2002).
241. Pommié, C., Levadoux, S., Sabatier, R., Lefranc, G. and Lefranc, M.-P. IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties. J Mol Recognit 17, 17-32 (2004).
242. Belessi, C.J., Davi, F.B., Stamatopoulos, K.E., Degano, M., Andreou, T.M., Moreno, C., Merle-Beral, H., Crespo, M., Laoutaris, N.P., Montserrat, E., Caligaris-Cappio, F., Anagnostopoulos, A.Z. and Ghia, P. IGHV gene insertions and deletions in chronic lymphocytic leukemia: "CLL-biased" deletions in a subset of cases with stereotyped receptors. Eur J Immunol 36, 1963-1974 (2006).
243. Smith, T.F. and Waterman, M.S. Identification of common molecular subsequences. J Mol Biol 147, 195-197 (1981).
244. Giudicelli, V., Duroux, P., Ginestoux, C., Folch, G., Jabado-Michaloud, J., Chaume, D. and Lefranc, M.-P. IMGT/LIGM-DB, the IMGT comprehensive database of

immunoglobulin and T cell receptor nucleotide sequences. Nucleic Acids Res 34, D781-D784 (2006).
245. Elemento, O. and Lefranc, M.-P. IMGT/PhyloGene: an on-line tool for comparative analysis of immunoglobulin and T cell receptor genes. Dev Comp Immunol 27, 763-779 (2003).
246. Kaas, Q., Ruiz, M. and Lefranc, M.-P. IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data. Nucleic Acids Res 32, D208-D210 (2004).
247. Kaas, Q., Ehrenmann, F. and Lefranc, M.-P. IG, TR and IgSF, MHC and MhcSF: what do we learn from the IMGT Colliers de Perles? Brief Funct Genomic Proteomic 6, 253-264 (2007).
248. Souto-Carneiro, M.M., Longo, N.S., Russ, D.E., Sun, H.W. and Lipsky, P.E. Characterization of the human Ig heavy chain antigen binding complementarity determining region 3 using a newly developed software algorithm, JOINSOLVER. J Immunol 172, 6790-6802 (2004).
249. Ohm-Laursen, L., Nielsen, M., Larsen, S.R. and Barington, T. No evidence for the use of DIR, D-D fusions, chromosome 15 open reading frames or VH replacement in the peripheral repertoire was found on application of an improved algorithm, JointML, to 6329 human immunoglobulin H rearrangements. Immunology 119, 265-277 (2006).
250. Volpe, J.M., Cowell, L.G. and Kepler, T.B. SoDA: implementation of a 3D alignment algorithm for inference of antigen receptor recombinations. Bioinformatics 22, 438-444 (2006).
251. Mollova S, Retter I and W, M. *Visualising the immune repertoire*. in *BMC Systems Biology*. 2007.
252. Gaeta, B.A., Malming, H.R., Jackson, K.J., Bain, M.E., Wilson, P. and Collins, A.M. iHMMune-align: hidden Markov model-based alignment and identification of germline genes in rearranged immunoglobulin gene sequences. Bioinformatics 23, 1580-1587 (2007).
253. Retter, I., Althaus, H.H., Munch, R. and Muller, W. VBASE2, an integrative V gene database. Nucleic Acids Res 33, D671-4 (2005).
254. Bleakley, K., Giudicelli, V., Wu, Y., Lefranc, M.P. and Biau, G. IMGT standardization for statistical analyses of T cell receptor junctions: the TRAV-TRAJ example. In Silico Biol 6, 573-588 (2006).
255. Giudicelli, V., Protat, C. and Lefranc, M.-P. *The IMGT strategy for the automatic annotation of IG and TR cDNA sequences: IMGT/Automat*. Ed DISC/Spid DKB ed. Vol. 31. 2003: ECCB'2003, European Conference on Computational Biology. pp103-104.

ANNEXES

Annexe 1. Alphabet dégénéré de l'ADN selon le code IUPAC-IUB

Lorsque nous parlons de comparaison de séquences, nous devons envisager le fait qu'à une position donnée, il existe des incertitudes sur la détermination de la base. On peut par exemple vouloir autoriser à une position soit une adénine soit une guanine : on dit alors qu'à cette position, on est en présence d'une base dégénérée.

Nous avons utilisé comme notation pour les bases dégénérées, le code IUPAC-IUB, résumé dans la figure suivante :

Code	Base	Origine du choix de la lettre
A	A	Adenine
C	C	Cytosine
G	G	Guanine
T	T	Thymine
R	A ou G	puRine
Y	C ou T	pYrimidine
S	G ou C	Intéraction forte (3 ponts H) Strong
W	A ou T	Intéraction faible (2 ponts H) Weak
K	G ou T	Keto
M	A ou C	aMino
B	C ou G ou T	Pas de A
D	A ou G ou T	Pas de C
H	A ou C ou T	Pas de G
V	A ou C ou G	Pas de T
N	N'importe qu'elle base	aNy
.	gap	

Ce tableau se lit ainsi: un R correspond à une adénine ou à une guanine.

Annexe 2. Matrice de substitution utilisée pour les alignements sans insertions et délétions

Matrice de Substitution, utilisée pour le calcul du score d'alignement par IMGT/V-QUEST. Elle prend en compte le code dégénéré de l'ADN. Pour une substitution d'un nucléotide par un même nucléotide la valeur est de 0 (ex : A en A), pour une substitution d'un nucléotide par un autre nucléotide la valeur est de 2 (ex : T en C ou W en S), et pour une substitution d'un nucléotide dégénéré par un nucléotide dégénéré pouvant définir le même nucléotide la valeur est de 1 (ex : R en S).

	.	x	T	c	A	G	R	Y	K	M	S	W	B	D	H	V
.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
X	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T	0	0	0	2	2	2	2	0	0	2	2	0	0	0	0	2
C	0	0	2	0	2	2	2	0	2	0	0	2	0	2	0	0
A	0	0	2	2	0	2	0	2	2	0	2	0	2	0	0	0
G	0	0	2	2	2	0	0	2	0	2	0	2	0	0	2	0
R	0	0	2	2	0	0	0	2	1	1	1	1	1	0	1	0
Y	0	0	0	0	2	2	2	0	1	1	1	1	0	1	0	1
K	0	0	0	2	2	0	1	1	0	2	1	1	0	0	1	1
M	0	0	2	0	0	2	1	1	2	0	1	1	1	1	0	0
S	0	0	2	0	2	0	1	1	1	1	0	2	0	1	1	0
W	0	0	0	2	0	2	1	1	1	1	2	0	1	0	0	1
B	0	0	0	0	2	0	1	0	0	1	0	1	0	1	1	1
D	0	0	0	2	0	0	0	1	0	1	1	0	1	0	1	1
H	0	0	0	0	0	2	1	0	1	0	1	0	1	1	0	1
V	0	0	2	0	0	0	0	1	1	0	0	1	1	1	1	0

Annexe 3. Matrice de substitution utilisée pour les alignements Smith et Waterman

Matrice utilisée fréquemment dans l'algorithme Smith & Waterman dans biojava. Matrice NUC.4.4 ftp://ftp.ncbi.nlm.nih.gov/blast/matrices/NUC.4.4

	A	T	G	C	S	W	R	Y	K	M	B	V	H	D	N
A	5	-4	-4	-4	-4	1	1	-4	-4	1	-4	-1	-1	-1	-2
T	-4	5	-4	-4	-4	1	-4	1	1	-4	-1	-4	-1	-1	-2
G	-4	-4	5	-4	1	-4	1	-4	1	-4	-1	-1	-4	-1	-2
C	-4	-4	-4	5	1	-4	-4	1	-4	1	-1	-1	-1	-4	-2
S	-4	-4	1	1	-1	-4	-2	-2	-2	-2	-1	-1	-3	-3	-1
W	1	1	-4	-4	-4	-1	-2	-2	-2	-2	-3	-3	-1	-1	-1
R	1	-4	1	-4	-2	-2	-1	-4	-2	-2	-3	-1	-3	-1	-1
Y	-4	1	-4	1	-2	-2	-4	-1	-2	-2	-1	-3	-1	-3	-1
K	-4	1	1	-4	-2	-2	-2	-2	-1	-4	-1	-3	-3	-1	-1
M	1	-4	-4	1	-2	-2	-2	-2	-4	-1	-3	-1	-1	-3	-1
B	-4	-1	-1	-1	-1	-3	-3	-1	-1	-3	-1	-2	-2	-2	-1
V	-1	-4	-1	-1	-1	-3	-1	-3	-3	-1	-2	-1	-2	-2	-1
H	-1	-1	-4	-1	-3	-1	-3	-1	-3	-1	-2	-2	-1	-2	-1
D	-1	-1	-1	-4	-3	-1	-1	-3	-1	-3	-2	-2	-2	-1	-1
N	-2	-2	-2	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1

Annexe 4. Valeurs du seuil et de l'overlap utilisées pour l'alignement global utilisé par IMGT/V-QUEST

Valeurs des seuils utilisées par IMGT/V-QUEST lors de différentes étapes d'alignement par IMGT/V-QUEST. Une valeur seuil correspond au score d'alignement minimal pour que celui-ci soit considéré comme significatif.

Valeurs des overlaps utilisées lors des différentes étapes d'alignement par IMGT/V-QUEST. La valeur d'un overlap définit la longueur minimale de l'alignement pour qu'il soit considéré, en fonction de la longueur de la plus petite des deux séquences à aligner. Par exemple un overlap de 1/3 correspond à une longueur minimale égale à au moins 1/3 de la longueur de la plus petite des deux séquences à aligner.

Etape d'alignement	**seuil**	**overlap**
Identification du type de chaîne	100	1/2
Identification et description de la V-REGION	600	1/3
Identification et description de la J-REGION	600	1/4
Identification et description de la D-REGION	500	1/2

Annexe 5. Séquences d'IG utilisées pour la mise en place de la recherche des insertions/délétions par IMGT/V-QUEST

Séquences d'IG humaine, de patient atteint de LLC. Elle proviennent du consortium ERIC (http://www.ericll.org/igcllWorkshop/problematic.txt) et sont caractérisées par la présence d'insertions et/ou de délétions.

Swe-336-II	IG	Humain	cctctggattcaccttcagttactactatatgagcggggtccgc caggctcccgggaaggggctggaatgggtaggtttcattagaaa caaagctaatggtgggacaacagaatagaccacgtctgtgaaag gcagattcacaatctcaagagatgattccaaaagcatcacctat ctgcaaatgaagagcctgaaaaccgaggacacggccgtgtatta ctgttccagaggctggcgggacgatttttggatagaactggttc gacccctgg
Swe-505-II	IG	Humain	cctctggattcaccttcacttactagtacatgagcggggtccgc caggctcccgggaaggggctggaatgggtacatttcattagaaa caaagctagtggtgggacaatagaatagaccacgtctgtgacag gcagattcacaatctcaagagatgattccaaaagcatcagctat ctgcaaatgaagagcctgaaaagcgacgacacggccgtgtgttc ttgttccagagaccaccaacacgatttttggagtggtattacgc aacggacgtctgg
N2205	IG	Humain	aggtgcagctgttggagtctgggggaggcttggtacagcctggg gggtccctgagactctcctgtgcagcctctggattcacctttag cagctatgccatgagctgggtccgccaggctccagggaaggggc tggagtgggtctcagctattagtggtagtggtggtagcacatac tacgcagactccgtgaagggccggttcaccatctccagagacaa ttccaagaacacgctgtatctgcaaatgaacagcctgagagccg aggacacggccgtatattactgtgcgaaagcttacacggtgact accgtactttgactactggggccagggaaccctggtcaccgtct cctca
N950HB	IG	Humain	caggtgcagctgcaggagtcgggcccaggactggtgaagccttc ggagaccctgtccctcacctgcactgtctctggtggctccatca gtagttactactggagctggatccggcagcccccagggaagggga ctggagtggattgggtatatctattacagtgggagcaccaacta caaccctccctcaagagtcgagtcaccatatcagtagacacgt ccaagaaccagttctccctgaagctgagctctgtgaccgctgtg gacacggccgtgtattactgtgcgagattaccccaactacgtat tactatgatagtagtggttaattgcagacgaagttgactactgg ggccagggaaccctggtcaccgtctcctca
P130	IG	Humain	caggtgcagctggtggagtctgggggaggcgtggtccagcctgg gaggtccctgagactctcctgtgtagcgtctggattcaccttca gttgttatggcatacactgggtccgccaggctccaggcaagggg ctggagtgggtgacagttatatggtatgatggaagtaataaatg ctatgcagactccgtgaagggccgattcacaatctccacagaca attccaagaacacgctggatctgcaaatgaacagcttgagagcc gaggacacgactgtgtattagtgtgcgagagggagtcgattact atgatagtagtcggaccttagacttttgatatctggggccaagg taccctggtcac
Swe-308-I	IG	Humain	ttctggaggcaccttcagcagctatgctatcagctgggtgcgac aggcccctggacaagggcttgagtggatgggagggatcatccct atctttggtacagcaaactacgcacagaagttccagggcagagt cacgattaccgcggacgaatccacgagcacagcctacatggagc tgagcagcctgagatctgaggacacggccgtgtattactgtgcg atttccccctctcagccttagcagtggctggccctttactacta ctacggtatggacgtctgg

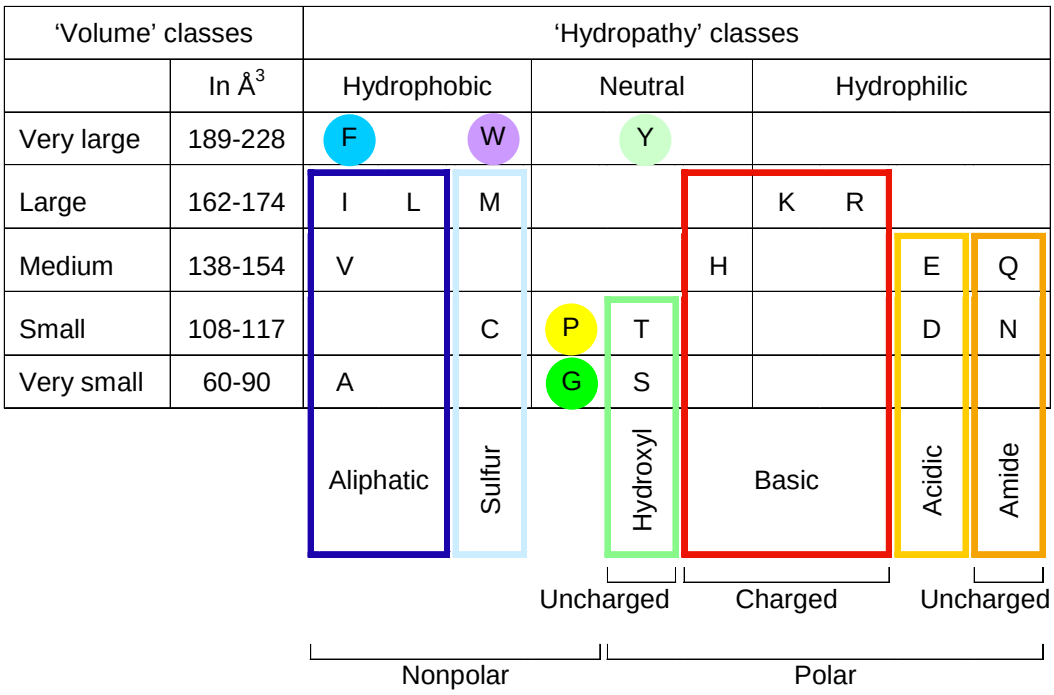
Swe-251-II	IG	Humain	gcctctggattcaccttcagtagctacgacatgcactgggtccg ccaagctacaggaaaaggtctggagtgggtctcagctattggta ctgctggtgacccatactatccaggctccgtgaagggccgattc accatctccagagaaaatgccaagaactccttgtatcttcaaat gaacagcctgagagccggggacacggctgtgtattactgtgcaa gataataagtgggctggtcgctacccgtcggctactggggccag ggaacc
Swe-110-I	IG	Humain	cctctcaattcaccctcagtaggtctagaatgagttgggtccgc caggctccagggaaggggctggaatgggtctcatctattactag tagtagtaattacatatactacgcagactcaatgaagggccgac tcaccatctccagagacaacgccaagaactcactgtatctgcaa atgaacagcctgagagccgaggacacggctgtgtattactgatc gaggaatcccggttcgtataactggaactacggctaatacaggt atacttatttatcgtactatgacactatggacgtctgg
FRAI2	IG	Humain	cttcgaagaccctgtccctcacctgcactgtctctggtggcccc atcaatagttactcctggagttggatccggcagccccccagggaa ggggctggagtggattgggaatatctattacagtgggaacacca actattacagtgggaacaccaactacaacccttccctcaagagt cgagtcaccatatcagtagacacgtccaagaaccagttctccct gaggctgagctctgtgaccgccgcggacacggccatttattact gtgcgagagacgcatactgtagtggtggtacctgctttgactgg tacttcgatgtctggggccgtggcaccctggtcaccgtctcctc ac
GREI1	IG	Humain	actgttgaagccttcggagaccctgtccctcacctgcgctgtct atggtgagtccttcaatggttactattggagctggatccgccag cccccagcttctggaaaggggctggagtggattggggaaatcga tcatagtggaagcaccaactacaacccgtccctcaagagtcgag tgaccatttcggtagacacgtcgaagaatcagttctccctgaag gtgagctccgtgaccgccgcagacacgggtgtatattactgtgc gagtggatcaacgccgactacgcgaaggtattactattacggta tggacgtctggggccaaggtaccctggtcaca
P781H	IG	Humain	tgcaggactgttgaagccttcggagaccctgtccctcacctgcg ctgtctatggtgagaccttcagtggttactactggacttggatc cgccagtccccagggaaggggctggaatggattggggaagtcaa taacagtggggaagtcaataacagtggaaccaccaactacaacc cgtccctcaagagtcgagtcaccatatcagtagacacgtcgaag aatcagttctccctgaggctgacctctgtgaccgccgcggacac ggctgtgtattactgtgcgggacgtttctattgttatggtggga actgtaataatgccaattactactactatggtatggacgtc tggggccaaggtaccctggtcacaa
ITAI1	IG	Humain	gaaatacaactggtggagtctgggggaggcttggtccagccggg ggggtccctgagactctcctgtgcagtctctggattcaccgtcc gtgacaacgatttcagctgggtccgccagtctcctgggaagggc ctggagtgggtctctattatttataaggacactggtgccacata ctacactgactccgtcaggggcagattcaccatctccagagaca attccaagaacacgctgtatcttcaaatgaacagtctgcgagtc gacgacacggctgtctatttctgtgcgagaactaggagagcagg agttttccacggctttgactactggggccagggagccctggtcc ccgtctcctcagggagtgcatccgccccaacccttttccccctc gtctcctgtgagaattccccgtcggatacgagcagcgtggccgt tggctgcctcgcacaggacttccttcccgactccatcacttt
P534	IG	Humain	actgttgaagccttcggagaccctgtccctcagttgcgcggtct atggtgggtcttttagtgattcatactggacctggatccgccag cccccagggaaggggctggagtggattggggaaattagtcatag tggaaacaccaactacaattacgacccgtccctcaagagtcgag tcaccatatcagtagacgcgtccaagaatcagttctccctgaaa ctgagctctgtgaccgccgcggacacggccgtctattactgtgc

			gagcgttaaagactacggtgagccgtctgacttctggggccagg gcaccctggtcaccgtc
ITADEL2V	IG	Humain	gaggtgcagctggtggagtctgggggaggcctggtcaagcctgg ggggtccctgagactctcctgtgcagcctctggattcaccttca gtagctatagcttgaattgggtccgccaggctccagggaagggg ctggagtgggtctcatccattagtagtagtcgttacatatacta ctcagactcagtgaagggccgattcactatctccagagacaacg ccaagaactcactgtatctgcaaatgaacagcctgagagccgag gacgcggctgtgtattactgtgcgagagatgctaacggtatgga cgtctggggccaagggaccacggtcaccgtctcctcagggagtg catccgccccaaccctttttccccctcgtctcctgtgagaattcc ccgtcggatacgagcagcgtggccgttggctgcctcgcacagga cttccttcccgactct
FRA-D3	IG	Humain	tcctgtgcagcctctggattcaccttcagtagctatagcatgaa ctgggtccgccaggctccagggaaggggctggagtgggtttcat acattagtagtagtaataccacatactacgcagactctgtgaag ggccgattcaccatctccagagacaatgccaagaactcactgta tctgcaaatgaacaccctgagagccgaggacacggctgtgtatt actgtgcgagagtagagggttcggggagttattttgactactgg ggccagggaaccctggtcaccgtctcctcaggtaaga
FRA-D2	IG	Humain	ctgggggtccctgagactctcctgtgcagcctctggattcaac ttcagtacctatagcatgaactgggtccgccaggctccagggaa ggggctggagtgggtctcatccattactagtagtagttacatat gctacgcagactcagtgaagggccgattcaccatctccagagac aacgccaagaactcactgtatctgcaaatgaacagcctgagagc cgaggacacggctgtatattactgtgcgagagatcttaacggca tggacgtctggggccaagggaccacggtcaccgtctcctca
P323H	IG	Humain	tgcagctggtggagtctgggggaggcttggtccagccgggggg tccctgagactctcctgtgcagcctctggattcacctttagtaa gtcttgcatgaggtgggtccgccaggctccagggaaggggcttg agtgggtggccaagaactcactgtatctacaaatgaacagcctg agagccgaggacacggctgtgtcttgctgtgcgagagatcgggg tttgtataacagtggctgtcgacgcttttcatatatggggccaa gggacaatggtcaccgtctcttca
AJ413992	IG	Humain	gaggtgcaacttgtggagtctgggggaggcttggtccatcctgg ggggtcacagacactctcctgtgcagtctctggattctccttta atatcgattggatcatctgggtccgccaggctccagggaaggggg ctggagtgggtggtcaacctataccaacatggatctgagaaata ctatgtggactctgtgaagggccgattcaccatctccagagaca acgccaataactcactgtatctgcaaatcaacacactgagagcc gagtacacggccgtgtatgacagtgcgagagatctcacgatttt tgggagtggttaccttgactactggggccagggaaccctggtca ccgtctcctc
AJ239361	IG	Humain	caggtgcagctggtggagtctgggggaggcgtggtccagcctgg gaggtccctgagactctcctgtgcagcctctggattcaccttca gtaggtatgctatgcactgggtccgccaggctccaggcaagggg ctggagtgggtggcagttatatcatatgatggaagcaataaata ctacgcagactccgtgaagggccgattcagcatctccagagaca ataccaagaacacgctgtatctgcaaatgaacagcctgagagct gaggacacggctgtatattactgtgtgagacccgggggtatag caacaactgcccaaagagtcgactagtggggccaggcaccctgg tcaccgtctcctca
AJ239389	IG	Humain	cagctggtgcagtctgggggaggcttggtccagcctgggggtc cctgagactctcctgtgcagcctctggattcacctttagtaatc attggatgagctgggtccggcaggctccagggaaggggctggag tgggtggccaacataaaccaagatggaagtgagaaacattttct

			ggactctgtgaagggccgactcagcatctccagagacaatggca agaagtcactgtatttgcaaatgaacagcctgagagtcgaggac acggctgtgtattactgtgcgagagtagccgataaaggatattg tagaggtgctagctgctatggtcgatcgggtgcatttgatatct gtggccaagggacaaaggtcaccgtctcttca
EB1	IG	Humain	gagatgcagctggtggagtctgggggaggcttggcaaagcctgc gtggtccccgagactctcctgtgcagcctctcaattcaccttca gtagctactacatgaactgtgtccgccaggctccagggaatggg ctggagttggtttgacaagttaatcctaatgggggtagcacata cctcatagactccggtaaggaccgattcaatacctccagagata acgccaagaacacacttcatctgcaaatgaacagcctgaaaacc gaggacacggccctcttttactgtaccagcctctaggggcccct tacgatttctggagatttgacattcgacccctggggccagggaa ccctggtcaccgtctcctc
Swe-360-II	IG	Humain	cctctggtttcaccttcagtgactactacatgaattgagtccgc caggctcccgggaaggggctggagtgggtacgttttactagaag taaagctaacggtgggacaacagaatagaccacgtctgtgaaag gcagattcacaatctcaagagatcattccaaaagcatcatctat ctgcaaataaacagcctgagagccgaggacacggctgtgtatta ctatgccagccgacccaaggaaatagttgtggtagctgctattc caggaactggttcgacccctgg
Swe-350-II	IG	Humain	ttccagatacaccttcaccaaatactttacacagtgggtgcgac agggccctggacaagggcatagtggttgggatgcatcaacccctt acaatgataatacacactacgcacagaagttccggggcagagtc accattaccagtgacaggtccgtgagcacagcctacatggagct gagcagcctgagatctgaagacatggtcgtgtattcctgtgtga gaggggaggcagcccaggaaactacggtatggacgtctgg
P103H	IG	Humain	aggtgcagctggtggagtctgggggaggcatggtacacccgggg gggtccctgagactatcctgtgcagtctttggatttacctttaa tagatgtggcatgaattgggtccgccaggctccagggaaggggc tggagtgggtctcgtagcacaaaaggagacttcgggaagggcca gttcaccatctccagagacaattccaagaacaccctgtatttgc aaacaaacagcctaagaggcgaggactcggccgtctattaatat gcgaaagagtccttggatgttataaggtcgttataaatgaataa tggggccagggaaccctggtcacccgtctcc
FD	IG	Humain	taggtgcagctggtggagtctgggggaggcctggtcaggcctgg ggggtccctgagactctcttgtgcagcctcttgattcaccttga atgaatgtgacatgaagtgggtccgccaggccccagggaagggg ctggagtgggtctcattcattacttatagtgccacttacataga ctatgcagattcagtgaagggccgattcaaggacccactctgtc tgcatgtccaagaagggccgatagaaggacccactctgtctgca tgtccaagaataagtgtatctacaaatgaatagcctgagagtcg aggacacggctatgtttcactgtgggaaaatcgcctaggatagt atactaattccaattgaccgtcgcccggcaggttcgacccctgg ggccagtgaaccctggtcaccgtctcctcagggagtg
N3881HB	IG	Humain	caggtccagctggtgcagtcttgggctgaggtgaggaagtctgg ggcctcagtgaaagtctcctgaagtttttctgggttcaccatca ccagctacggtatgcattgggtgcaacagtcccctggacaaggg cttgagtggatgggatggatcaaccctggcaatggtagcccaag ctatgccaagaagtttcagggcagattcaccatgaccaggcaca tgtccacaaccacatcctacacagacctgagcagcctgacatct gaggacatggctgtgtattaatatgtaagacgtcggactatcac gatattttgactggttgccggcggaggactactggggccaggga accctggtcaccgtctcctcaa

Annexe 6. Classes des Acides aminés IMGT

Cette table montre les 3 classes ‘d’hydropathie’ (hydrophobe, neutre, hydrophile), les 5 classes de 'volume' en Å ([60-90], [108-117], [138-154], [162-174], [189-228]) et les 7 classes de ‘caractéristiques physicochimiques’ (aliphatique, sulfidique, hydroxyle, acide, Aminique, basique, F, W, Y, G, P), définies par IMGT®.



L’abréviation des acides aminés : A (Ala) alanine, C (Cys) cystéine, D (Asp) aspartic acid, E (Glu) glutamic acid, F (Phe) phenylalanine, G (Gly) glycine, H (His) histine, I (Ileu) isoleucine, K (Lys) lysine, L (Leu) leucine, M (Met) méthionine, N (Asn) asparagine, P (Pro) proline, Q (Gln) glutamine, R (Arg) arginine, S (Ser) sérine, T (Thr) thréonine, V (Val) valine, W (Trp) tryptophan, Y (Tyr) tyrosine.

Annexe 7. Vocabulaire Contrôlé défini pour la base de données IMGT/CLL-DB

Intitulé	Valeur
Gender	M ; F
Ethnic origin	American indian; Australian aboriginal; Black ; Caucasoid; Hispanic; Mixed; Oriental; Pacific islander; Other
Country	Greece; Italy; Spain; USA; France; Macedonia; Armenia; Albania; Other
Current status	Alive; Lost to follow-up; dead; Unknown
CLL related death	Yes; No; Unknown
Rai	I, II, III, IV
Binet	A; B; C
Family history of CLL	Yes; No; Unknown
Family history of other B cell proliferations	Yes; No; Unknown
Need for treatment	Yes; No; Unknown; Not applicable
Response to treatment	Yes; No; Unknown; Not applicable
Coombs direct	Positive; Negative; Not applicable
Coombs indirect	Positive; Negative; Not applicable
ANA	Positive; Negative; Not applicable
RF	Positive; Negative; Not applicable
Sample status	fresh; frozen; cryopreserved; paraffin-embedded; other
Tissue type	blood; bone marrow, other
Cell type	Bcell; Other
Molecule type	cDNA; gDNA
Immunophenotype	Typical; Atypical;
Surface IGH chain expression	mu+delta; mu; delta; gamma; alpha ; not applicable
Surface IGL chain expression	kappa; Lambda; Kappa+Lambda; not applicable
FISH	Del11q; 12+; del13q; del17p; normal; complxex; del13q+del13q; del13q+del11; 12+del13q; del13q+del17p
Transformation	RICHTER; PLL; Other; No; Unknown
Clinical evolution	Progressive; Stable; Regression; Unknown
Clinical autoimmunity	Anemia; Thrombocytopenia; Other; No; Unknown; PAGET; AHA; Hashimoto; AIHA; Graves; Hyperthropemphigoid; Atopic dermatitis; Hypothyroidism
Infection	Hepatitis C; HIV; HIV, Other; Hepatitis C, HIV; hepatitis C, Other; No; HBV; HCV; Tuberculosis; Unknown; Negative
Tumors	Breast cancer; Breast cancer, Other; Other; No; Unknown; Intestinal cancer; gastric cancer; Thyroid cancer; Larynx cancer; Malnoma; Prostate cancer; Skin cancer

Annexe 8. Liste des partenaires du projet IMGT/CLL-DB

Kostas Stamatopoulos
Département d'hématologie et unité HCT, G. hôpital Papanicolaou, Thessalonique, Grèce.

Chrysoula Belessi
Département d'hématologie, hôpital général Nikea, Piraeus, Grèce.

Frédéric Davi
Département d'hématologie, Hôpital Pitié-Salpêtrière et Université Pierre et Marie Curie, Paris, France.

Nicholas Chiorazzi
The Feinsteins Institute for medical Research, North Shore – LIJ Health System, Manhasset, NY ; département de médecine, Hôpital universitaire de North Shore, Manhasset, NT, et département de biologie cellulaire et médecine, Albert Einstein School of Medecine, Bronx, NY, USA.

Paolo Ghia
Laboratoire et unité de lymphoïde maligne, Département d'oncologie, Université de Vita-Salute San Raffaele et Institut scientifique de San Raffaele, Milan, Italie.

Carol Moreno
Institut d'hématologie et oncologie, Département d'hématologie, hôpital Clinic, IDIBAPS, Université de Barcelone, Espagne.

Richard Rosenquist
Département de génétique et pathologie, Université d'Uppsala, Uppsala, Suède.

PUBLICATIONS

PUBLICATION 1

Published online 23 May 2008 *Nucleic Acids Research, 2008, Vol. 36, Web Server issue* ***W503–W508***
doi:10.1093/nar/gkn316

IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis

Xavier Brochet[1], Marie-Paule Lefranc[1,2,*] and Véronique Giudicelli[1]

[1]IMGT®, the international ImMunoGeneTics information system®, Laboratoire d'ImmunoGénétique Moléculaire LIGM, Université Montpellier 2, Institut de Génétique Humaine IGH, UPR CNRS 1142, 141 rue de la Cardonille, 34396 Montpellier cedex 5 and [2]Institut Universitaire de France, 103 Bd St Michel, 75005 Paris, France

Received January 30, 2008; Revised April 17, 2008; Accepted May 7, 2008

ABSTRACT

IMGT/V-QUEST is the highly customized and integrated system for the standardized analysis of the immunoglobulin (IG) and T cell receptor (TR) rearranged nucleotide sequences. IMGT/V-QUEST identifies the variable (V), diversity (D) and joining (J) genes and alleles by alignment with the germline IG and TR gene and allele sequences of the IMGT reference directory. New functionalities were added through a complete rewrite in Java. IMGT/V-QUEST analyses batches of sequences (up to 50) in a single run. IMGT/V-QUEST describes the V-REGION mutations and identifies the hot spot positions in the closest germline V gene. IMGT/V-QUEST can detect insertions and deletions in the submitted sequences by reference to the IMGT unique numbering. IMGT/V-QUEST integrates IMGT/JunctionAnalysis for a detailed analysis of the V-J and V-D-J junctions, and IMGT/Automat for a full V-J- and V-D-J-REGION annotation. IMGT/V-QUEST displays, in 'Detailed view', the results and alignments for each submitted sequence individually and, in 'Synthesis view', the alignments of the sequences that, in a given run, express the same V gene and allele. The 'Advanced parameters' allow to modify default parameters used by IMGT/V-QUEST and IMGT/Junction Analysis according to the users' interest. IMGT/V-QUEST is freely available for academic research at http://imgt.cines.fr

INTRODUCTION

IMGT®, the international ImMunoGeneTics information system® (http://imgt.cines.fr) (1), is the international reference in immunogenetics and immunoinformatics. Created in 1989 at the Laboratoire d'ImmunoGénétique Moléculaire (LIGM) (Université Montpellier 2 and CNRS), IMGT® provides a high quality integrated knowledge resource, specialized in the immunoglobulins (IG) or antibodies, T cell receptors (TR), major histocompatibility complex (MHC) of human and other vertebrates and related proteins of the immune system (RPI), which belong to the immunoglobulin superfamily (IgSF) and to the MHC superfamily (MhcSF). IMGT® includes databases, web resources and interactive tools. The accuracy and the consistency of the IMGT® data are based on IMGT-ONTOLOGY, the first ontology for immunogenetics and immunoinformatics (2,3). IMGT/V-QUEST, for 'V-QUEry and STandardization', is the first integrated IMGT® tool which has been online since 1997 (4). IMGT/V-QUEST analyses the IG and TR rearranged nucleotide sequences that result from the very complex mechanisms at the origin of antigen receptor diversity (10^{12} antibodies and 10^{12} TR per individual) and which include the rearrangements of the variable (V), diversity (D) and joining (J) genes, the N-diversity mechanism and, for IG, the somatic mutations [for review see (5,6)]. IMGT/V-QUEST identifies the V, D and J genes and alleles in rearranged V-J and V-D-J sequences by alignment with the germline IG and TR gene and allele sequences of the IMGT reference directory. It delimits the framework regions (FR-IMGT) and complementarity determining regions (CDR-IMGT) and provides a detailed and accurate characterization of the submitted sequences according to the IMGT Scientific chart rules, based on the IMGT-ONTOLOGY axioms and concepts of description, classification and numerotation (2,3). New functionalities were added to IMGT/V-QUEST through a complete rewrite in Java. Thus, IMGT/V-QUEST analyses batches of sequences (up to 50) in a single run. The analysis has been upgraded with the description of V-REGION mutations, with the identification of the hot spots positions in the closest germline V gene, and with the detection and accurate description of insertions and deletions in the submitted sequences by reference to the

*To whom correspondence should be addressed. Tel: +33 4 99 61 99 65; Fax: +33 4 99 61 99 01; Email: Marie-Paule.Lefranc@igh.cnrs.fr

IMGT unique numbering (7). IMGT/V-QUEST integrates IMGT/JunctionAnalysis (8) for a detailed analysis of the V-J and V-D-J junctions, and IMGT/Automat (9) for a full annotation of the V-J- and V-D-J-REGION. The interface has been customized to fit the users' needs: in addition to the standard 'Detailed view' which displays the results and alignments for each submitted sequence individually, a new results display 'Synthesis view' has been implemented to provide, for a given run, the alignments of the sequences that express the same V gene and allele and, per locus, the results of IMGT/JunctionAnalysis. The 'Advanced parameters' allow to modify default parameters used by IMGT/V-QUEST and IMGT/JunctionAnalysis algorithms, according to the users' interest. IMGT/V-QUEST is currently available for human and mouse rearranged sequences, and partly for 31 other species (nonhuman primates, rat, sheep, teleostei and chondrichthyes). IMGT/V-QUEST is freely available for academic research from the IMGT® Home page (http://imgt.cines.fr).

ALGORITHM AND IMPLEMENTATION

IMGT/V-QUEST was totally rewritten in Java language in order to fully unify the implementation of the different components. The identification of the closest V, D and J genes and alleles of a given receptor type (IG or TR) and of a given species is based on the same principles as previously described (4). IMGT/V-QUEST algorithm was developed using pairwise alignment and sequence comparison of experimental data, expertly annotated and standardized by IMGT® (10). Briefly, the identification of the closest V, D and J genes and alleles is based on global pairwise alignment (two for the V), without insertions nor deletions, of the user sequence with different subsets of the IMGT reference directory, followed by a similarity evaluation. This last step is preceded, for the V region, by the insertion of gaps according to the IMGT unique numbering (7).

The complete rewrite in Java has optimized the processing time and allowed the development and integration of new functionalities which improve the analysis accuracy with score, evaluation, warnings, etc. IMGT/V-QUEST can analyze batches of sequences (up to 50) in a single run. A new option allows to search for potential insertions and deletions in the user sequence. This search follows the identification of the closest V gene and allele. The program proceeds in two alignment steps [Smith and Waterman algorithm (11)] between the user sequence and the closest V genes and alleles, first looking for insertions, then for deletions. If insertions are detected, they are excluded from the user sequence as they disrupt the IMGT numbering but their positions are memorized so they can be displayed in the results page. If deletions are detected, gaps are entered in the user sequence to restore the IMGT numbering. After insertion and/or deletion detection, the steps of V gene and allele identification are performed again. IMGT/V-QUEST provides an accurate localization and description of the mutations in the user sequence by comparison with the closest germline V gene and allele. Nucleotide (nt) mutations are characterized as silent versus nonsilent, and transition versus transversion. Amino acid (AA) changes are qualified according to the IMGT AA classes (12) based on hydropathy, volume and physicochemical characteristics. The tool identifies the mutation hot spots patterns and positions in the closest germline V-REGION. IMGT/V-QUEST integrates IMGT/JunctionAnalysis (8) for the detailed analysis of the junction (performed even if cystein 104 and/or phenylalanine/tryptophan 118 are mutated) (7) and for the optional display of the eligible D genes and alleles, and IMGT/Automat (9) for the full annotation of the V-J-REGION or V-D-J-REGION in an IMGT/LIGM-DB-like format (13).

Regarding the input sequences, IMGT/V-QUEST accepts nucleotide sequences in the International Union of Biochemistry (IUB) code (14) and is able to deal with complementary reverse sequences. The tool trims the 'n' nucleotides at the 5′ and 3′ extremities of user sequences since they influence the alignments and the gene and allele identification.

IMGT/V-QUEST SEARCH

An IMGT/V-QUEST search consists of two easy steps: (i) the user selects the antigen receptor (IG or TR) and the species on the IMGT/V-QUEST Home page and (ii) on the next page, the user submits up to 50 nt sequences in FASTA format. By clicking on 'Start', the analysis is done automatically with the default parameters.

Prior to launching the search, users may customize the results display options in 'Selection for results display' (Figure 1). They can export the results in text and choose the 'Nb of nucleotides per line in alignments'. They can select the option 'A. Detailed view' for the display of the results of each analyzed sequence individually (with a choice of 14 different results displays), or the option 'B. Synthesis view' for the display of the alignments of sequences that express the same V gene and allele (with a choice of eight different results displays).

For sophisticated queries or for unusual sequences, the users can modify the default values in 'Advanced parameters'. The customizable values are:

(i) 'Selection of IMGT reference directory set' used for the V, D and J genes and alleles identification and alignments ('F + ORF', 'F + ORF + in-frame P', 'F + ORF including orphons' and 'F + ORF + in-frame P including orphons', where F is functional, ORF is open reading frame and P is pseudogene). This allows the user to work with only relevant gene sequences (e.g. orphon sequences are relevant for genomic but not for expressed repertoire studies). The selected set can also be chosen either 'With all alleles' or 'With allele *01 only'.
(ii) 'Search for insertions and deletions'. In that case, the number of submitted sequences in a single run is limited to 10.
(iii) 'Parameters for IMGT/JunctionAnalysis': Nb of D-GENEs allowed in the IGH, TRB and TRD junctions and Nb of accepted mutations in

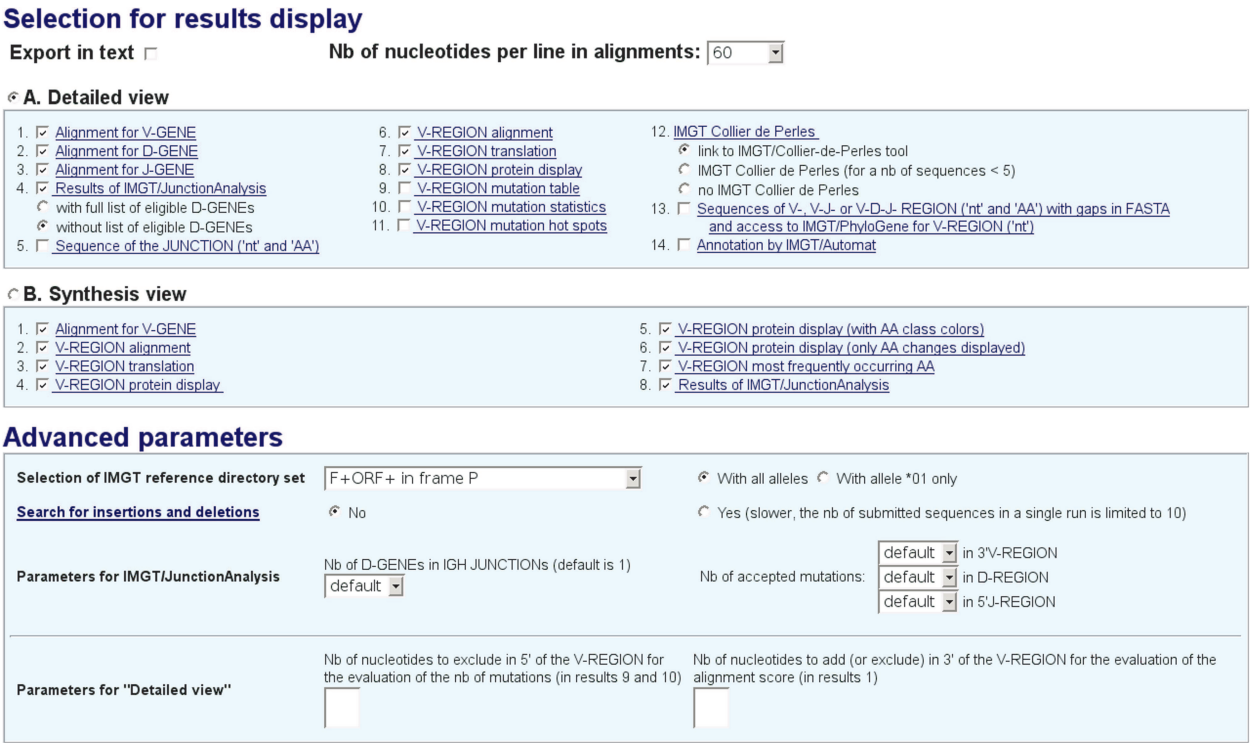


Figure 1. The customizable parameters for an IMGT/V-QUEST search. 'Selection for results display' allows the users to select the types of results to be displayed (14 for 'Detailed view', and 8 for 'Synthesis view'). The 'Advanced parameters' allow the users to modify the default values.

3′V-REGION, D-REGION and 5′J-REGION (default values are indicated per locus in the IMGT/V-QUEST Documentation).

(iv) 'Parameters for Detailed View': 'Nb of nucleotides to exclude in 5′ of the V-REGION for the evaluation of the nb of mutations' (to avoid, for example, to count primer specific nucleotides) and/or 'Nb of nucleotides to add (or exclude) in 3′ of the V-REGION for the evaluation of the alignment score' (e.g. in case of low or high exonuclease activity).

IMGT/V-QUEST OUTPUT

Detailed view

The top of the 'Detailed view' results page indicates the number of analyzed sequences with links to individual results. Each individual result comprises the user sequence displayed in FASTA format (a sequence submitted in antisense orientation is shown as complementary reverse sequence that is in V gene sense orientation), a 'Result summary' table (Figure 2A) followed, if all parameters were selected, by the 14 different results displays. The 'Result summary' provides a crucial feature that is the evaluation of the user sequence functionality performed by IMGT/V-QUEST: productive (if no stop codon and in-frame junction) or unproductive (if stop codons and/or out-of-frame junction). It also summarizes the main characteristics of the analyzed sequence which include the names of the closest 'V-GENE and allele' and 'J-GENE and allele' with the alignment score and the percentage of identity, the name of the closest 'D-GENE and allele' with the D-REGION reading frame, the three CDR-IMGT lengths (shown within brackets, e.g. [8.8.13]) which characterize a V domain and the AA JUNCTION sequence. IMGT/V-QUEST provides warnings that appear, as notes in red to alert the users, if potential insertions or deletions are suspected in the V (sequences with <85% of identity and/or with different CDR1-IMGT and/or CDR2-IMGT lengths compared with the closest germline V-REGION), or if other possibilities for the J gene and allele are identified. If the option 'Search for insertions and deletions' was selected, the detection and detailed description of insertions and/or deletions are shown in the 'Result summary' first row to capture the user attention. Moreover, insertions appear as capital letters in the FASTA sequence.

Below the 'Result summary', the alignments for the V-, D- and J-GENE (4,15,16) comprise the alignment score and the identity percentage with the five closest genes and alleles and, for the V, the length of the V-REGION taken into account for the score evaluation. The 'Results of IMGT/JunctionAnalysis' (8,15) include, if selected, the list of eligible D genes and alleles which match >4 nt with the junction, allowing users to visualize the result among other close solutions. 'Sequence of the JUNCTION ('nt' and 'AA')' is given by IMGT/V-QUEST if no results are obtained with IMGT/JunctionAnalysis. Different displays of the V region ('V-REGION alignment', 'V-REGION translation' and 'V-REGION protein display') and of the mutations affecting the V region are all based on the

A Result summary of 'Detailed view'

Example of a productive sequence

Result summary:	**Productive IGH rearranged sequence** (no stop codon and in frame junction)		
V-GENE and allele	IGHV5-51*01	score = 1327	identity = **95,83%** (276/288 nt)
J-GENE and allele	IGHJ3*02 (a)	score = 191	identity = 87,76% (43/49 nt)
D-GENE and allele by IMGT/JunctionAnalysis	IGHD1-26*01	D-REGION is in reading frame 3	
[CDR1-IMGT.CDR2-IMGT.CDR3-IMGT] lengths and AA JUNCTION	[8.8.13]	CARQGGSYPDAFDIW	

(a) Other possibilities: IGHJ6*02 (highest number of consecutive identical nucleotides)

Example of an unproductive sequence

Result summary:	**Unproductive IGH rearranged sequence** (out of frame junction)		
V-GENE and allele	IGHV5-51*01	score = 1408	identity = **98,96%** (285/288 nt)
J-GENE and allele	IGHJ4*02	score = 66	identity = 75,00% (18/24 nt)
D-GENE and allele by IMGT/JunctionAnalysis	IGHD6-6*01	D-REGION is in reading frame 3	
[CDR1-IMGT.CDR2-IMGT.CDR3-IMGT] lengths and AA JUNCTION	[8.8.X]	CANIPSLIAARRP##YW	

B Summary table of 'Synthesis view'

Sequence ID	V-GENE and allele	Functionality	V-REGION score	V-REGION identity % (nt)	J-GENE and allele	D-GENE and allele	D-REGION reading frame	CDR-IMGT lengths	AA JUNCTION	JUNCTION frame
seq1	IGHV3-73*01	Productive	1240	91,50% (269/294 nt)	IGHJ1*01 (a)	IGHD3-10*01	3	[8.10.10]	CVIRGDVYNRQW	in frame
seq2	IGHV5-51*01	Productive	1210	91,32% (263/288 nt)	IGHJ4*02	IGHD2-15*01	2	[8.8.17]	CVRGRGYCSAGSCYDFVYW	in frame
seq3	IGHV5-51*01	Unproductive	1408	98,96% (285/288 nt)	IGHJ4*02	IGHD6-6*01	3	[8.8.X]	CANIPSLIAARRP##YW	out of frame
seq4	IGHV5-51*01	Productive	1300	94,79% (273/288 nt)	IGHJ6*02	IGHD3-10*01	2	[8.8.18]	CARARGSGSYSYYYYGVDVW	in frame
seq5	IGHV5-51*01	Productive	1210	91,32% (263/288 nt)	IGHJ4*02	IGHD2-15*01	2	[8.8.17]	CVRGRGYCSAGSCYDFDYW	in frame
seq6	IGHV5-51*01	Unproductive (stop codons)	1426	99,65% (287/288 nt)	IGHJ4*02	IGHD6-6*01	1	[8.8.16]	CARRGKDSSSSFSEFDYW	in frame
seq7	IGHV5-51*01	Productive	1327	95,83% (276/288 nt)	IGHJ3*02	IGHD6-13*01	3	[8.8.14]	CARGSGPEVPDAFDIW	in frame
seq8	IGHV5-51*01	Productive	1327	95,83% (276/288 nt)	IGHJ3*02 (a)	IGHD1-26*01	3	[8.8.13]	CARQGGSYPDAFDIW	in frame
seq9	IGKV3-15*01	Productive	1291	96,06% (268/279 nt)	IGKJ1*01	-	-	[6.3.8]	CQQYHYWPTF	in frame
seq10	IGLV2-14*01	Productive	1408	98,96% (285/288 nt)	IGLJ2*01	-	-	[9.3.12]	CSSYTSSSTLGVVF	in frame
seq11	IGLV2-14*01	Productive	1228	92,01% (265/288 nt)	IGLJ3*02	-	-	[9.3.10]	CNSYTTDTIWVF	in frame

(a) Other possibilities may be found, please check the alignments for this sequence in "Detailed view"

Figure 2. IMGT/V-QUEST output. (**A**) 'Result summary' of 'Detailed view'. The note (a) represents a warning indicating that there is another possibility for the choice of the J-GENE, the criterion for that choice being between parentheses. In the case of an unproductive sequence due to an out-of-frame junction, the CDR3-IMGT length cannot be determined and is indicated by an 'X'. The symbol '#' indicates a frameshift. (**B**) 'Summary table' of 'Synthesis view'. The note (a) represents a warning indicating that there are other possibilities for the results.

IMGT unique numbering (7) and on the FR-IMGT and CDR-IMGT delimitation. The extensive and standardized characterization of the mutations performed by IMGT/V-QUEST answers the requirements of IG sequence studies, evolution and function analysis, clinical interpretations and antibody engineering. The 'V-REGION mutation table' lists the mutations (nt and AA) of the analyzed sequence compared to the closest V-REGION allele. They are described for the V-REGION and for each FR-IMGT and CDR-IMGT, with their positions, and for the AA changes according to the IMGT AA classes (12). For example, c16 > g, Q6 > E (++−) means that the nt mutation (c > g) leads to an AA change at codon 6 with the same hydropathy (+) and volume (+) but with different physicochemical properties (−) classes (12). It is the first time that such qualification of amino acid replacement is provided. The 'V-REGION mutation statistics' evaluates the number of silent and nonsilent mutations and the number of transitions and transversions of the analyzed nucleotide sequence, and the number of AA changes of its translated sequence. The 'V-REGION mutation hot spots' show the patterns and localization of hot spots in the closest germline V-REGION. The identified hot spots patterns are (a/t)a and (a/g)g(c/t)(a/t) and the complementary reverse motifs are t(a/t) and (a/t)(a/g)c(c/t) (see Lefranc,M-P. and Lefranc,G. Somatic hypermutations, in IMGT Education, http://imgt.cines.fr).

'IMGT Collier de Perles' provides a link to the IMGT/Collier-de-Perles tool (17) or directly includes this standardized graphical representation of the user IG or TR rearranged sequence. The 'Sequences of V-, V-J- or V-D-J-REGION ('nt' and 'AA') with gaps in FASTA and access to IMGT/PhyloGene for V-REGION ('nt')' provides the analyzed sequence with IMGT gaps, in FASTA format and on one line, and a link to IMGT/PhyloGene (18). The 'Annotation by IMGT/Automat' (9) uses the results of the analysis to provide a full automatic annotation of the user sequences for the V-J-REGION or V-D-J-REGION.

Synthesis view

The aim of 'Synthesis view', a novel IMGT/V-QUEST result, is to facilitate the comparison of sequences that express the same V gene and allele: it allows to compare the localization of the mutations and the composition of their junctions. The 'Synthesis view' comprises a 'Summary table' (Figure 2B) and eight different displays (if all were selected). The 'Summary table' shows, for each sequence, the name of the closest V gene and allele, the evaluation of the sequence functionality, the V score and

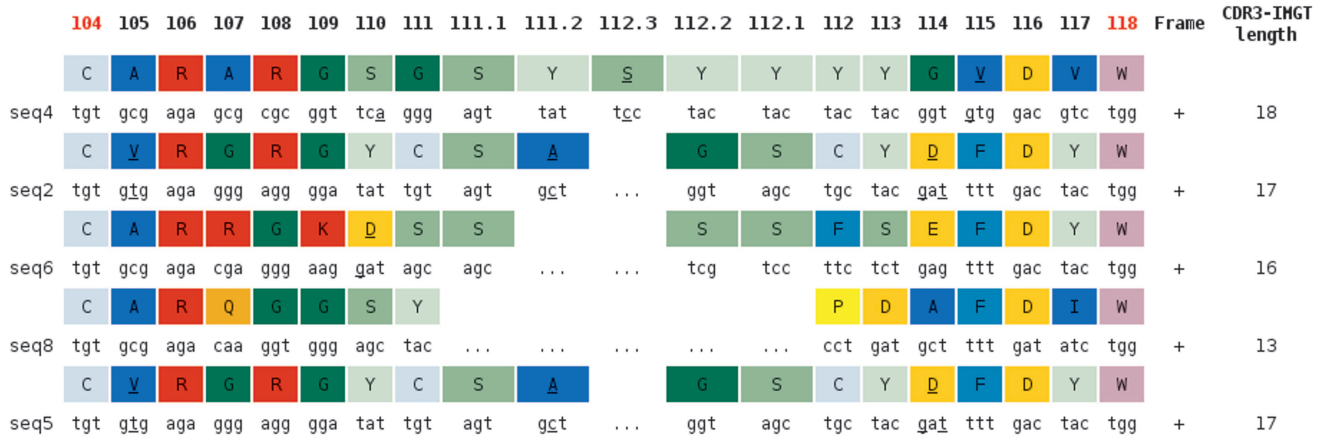


Figure 3. IMGT/JunctionAnalysis translation of 'Synthesis view'. Results are given per locus, here five IGH sequences are shown. Amino acids of the JUNCTIONs are colored according to the 11 IMGT AA physicochemical classes (12). Underlined nt and AA are mutated compared with the closest germline 3′V-REGION, D-REGION and 5′J-REGION of each JUNCTION.

percentage of identity, the name of the closest J and D genes and alleles, the D-REGION reading frame, the three CDR-IMGT lengths, the AA JUNCTION and the JUNCTION frame. Warnings appear to alert users on potential insertions or deletions in the V or on other possibilities for the J gene and allele. In such cases, it is strongly recommended to check the individual results of these sequences in 'Detailed view'.

The originality of 'Synthesis view' is also to provide alignments of sequences which, in a given run, are assigned to the same V gene and allele. The 'Alignment for V-GENE', 'V-REGION alignment' and 'V-REGION translation' are based on the same characteristics as those of 'Detailed view'. In addition, the hot spots positions are underlined in the germline V-REGION (for an easy comparison with the mutation localizations) and the name of the closest J gene allele is indicated at the 3′ end of each sequence. The 'V-REGION protein display' shows amino acid sequences aligned with the closest V-REGION allele. This protein display is also provided with AA colors according to the IMGT AA classes (12) or with only the AA changes displayed. The 'V-REGION most frequently occurring AA per position and per FR-IMGT and CDR-IMGT' table is given for each alignment to highlight the position of conserved AA in sequence batches. The 'Results of IMGT/JunctionAnalysis' are displayed per locus (IGH, IGK and IGL for the IG heavy, kappa and lambda sequences, TRA, TRB, TRG and TRD for the TR alpha, beta, gamma and delta sequences) (Figure 3).

CONCLUSION

IMGT/V-QUEST is a highly customized and integrated system for the IG and TR standardized V-J and V-D-J sequence analysis. The addition of new features (AA physicochemical characteristics, mutations and insertions/deletions description, 'Synthesis view') and 'Advanced parameters' provides a wide spectrum of new types of analysis across species. The integration of tools has allowed to answer specific needs: IMGT/JunctionAnalysis for the accurate and extensive description of the V-J and V-D-J junction, IMGT/Automat for the full and detailed user sequence annotation and IMGT/Collier-de-Perles for bridging the gap between sequences and 3D structures. By its wide range of novel customizable and integrated functionalities and by its high level of standardization, IMGT/V-QUEST is unique among the other software programs (19–22). The information provided by IMGT/V-QUEST is of much value for the comparative analysis of the IG and TR sequences (including those of the newly sequenced vertebrate genomes), for statistical analysis of the junctions (23) and of the repertoires, for antibody engineering and therapeutical antibodies (24) (single chain Fragment variable (scFv), phage displays, combinatorial libraries, chimeric, humanized and human antibodies). IMGT/V-QUEST, with an average of 30 000 requests per month, is widely used by the clinicians to analyze the expressed repertoire in normal and pathological immune responses (autoimmune diseases, leukemias, lymphomas, myelomas, infectious diseases, etc.), in diagnostics of clonalities, detection and follow-up of minimal residual diseases.

Given the necessity to have an international and publicly available benchmarch of high quality for clinical data comparison, IMGT® standards based on IMGT-ONTOLOGY have recently been approved by the World Health Organization–International Union of Immunological Societies (WHO–IUIS) subcommittee for IG and TR (25,26). Standardized criteria for the analysis of IG and TR rearranged V-J and V-D-J sequences and 3D structures include IMGT reference directory (available free for academics on the Web site), IMGT gene nomenclature (5,6,10) approved by the HUGO Nomenclature Committee, IMGT label description, IMGT unique numbering for V-DOMAIN (7) and CDR-IMGT delimitations. IMGT/V-QUEST has been recommended by the European Research Initiative on chronic lymphocytic leukemia CLL (ERIC) for the analysis of the IGHV gene mutational status in CLL (27). IMGT/V-QUEST output results evaluation by biologists and feedback from

researchers and clinicians are part of the new developments to answer biological and medical needs in an international collaboration (28–30).

AVAILABILITY AND CITATION

Authors who use IMGT/V-QUEST are encouraged to cite this article and to quote the IMGT® Home page (http://imgt.cines.fr). IMGT/V-QUEST is freely available for academic research.

ACKNOWLEDGEMENTS

We are grateful to Vincent Nègre and to the IMGT team for its expertise and constant motivation. IMGT® received funding from Centre National de la Recherche Scientifique CNRS, Ministère de l'Enseignement Supérieur et de la Recherche MESR (University Montpellier 2), Région Languedoc-Roussillon, Agence Nationale de la Recherche ANR (ANR-06-BYOS-0005-01) and European Community (ImmunoGrid, FP6-2004-IST-4). Funding to pay the Open Access publication charges for this article was provided by CNRS.

Conflict of interest statement. None declared.

REFERENCES

1. Lefranc,M.-P., Giudicelli,V., Kaas,Q., Duprat,E., Jabado-Michaloud,J., Scaviner,D., Ginestoux,C., Clément,O., Chaume,D. and Lefranc,G. (2005) IMGT, the international ImMunoGeneTics information system®. *Nucleic Acids Res.*, **33**, D593–D597.
2. Giudicelli,V. and Lefranc,M.-P. (1999) Ontology for immunogenetics: IMGT-ONTOLOGY. *Bioinformatics*, **15**, 1047–1054.
3. Duroux,P., Kaas,Q., Brochet,X., Lane,J., Ginestoux,C., Lefranc,M.-P. and Giudicelli,V. (2008) IMGT-Kaleidoscope, the formal IMGT-ONTOLOGY paradigm. *Biochimie*, **90**, 570–583.
4. Giudicelli,V., Chaume,D. and Lefranc,M.-P. (2004) IMGT/V-QUEST, an integrated software program for immunoglobulin and T cell receptor V–J and V–D–J rearrangement analysis. *Nucleic Acids Res.*, **32**, W435–W440.
5. Lefranc,M.-P. and Lefranc,G. (2001) *The Immunoglobulin FactsBook,* Academic Press, London, pp.1–458.
6. Lefranc,M.-P. and Lefranc,G. (2001) *The T Cell Receptor FactsBook,* Academic Press, London, pp.1–398.
7. Lefranc,M.-P., Pommié,C., Ruiz,M., Giudicelli,V., Foulquier,E., Truong,L., Thouvenin-Contet,V. and Lefranc,G. (2003) IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Dev. Comp. Immunol.*, **27**, 55–77.
8. Yousfi Monod,M., Giudicelli,V., Chaume,D. and Lefranc,M.-P. (2004) IMGT/JunctionAnalysis: the first tool for the analysis of the immunoglobulin and T cell receptor complex V-J and V-D-J JUNCTIONs. *Bioinformatics*, **20**, i379–i385.
9. Giudicelli,V., Protat,C. and Lefranc,M.-P. (2003) The IMGT strategy for the automatic annotation of IG and TR cDNA sequences: IMGT/Automat. In Christophe,C., Lenhof,H.-P. and Sagot,M.-F. (eds), *Proceedings of the European Conference on Computational Biology (ECCB 2003)*, INRIA (DISC/Spid), Paris, DKB-31, pp. 103–104.
10. Giudicelli,V., Chaume,D. and Lefranc,M.-P. (2005) IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucleic Acids Res.*, **33**, D256–D261.
11. Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
12. Pommié,C., Levadoux,S., Sabatier,R. and Lefranc,M.-P. (2004) IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties. *J. Mol. Recognit.*, **17**, 17–32.
13. Giudicelli,V., Duroux,P., Ginestoux,C., Folch,G., Jabado-Michaloud,J., Chaume,D. and Lefranc,M.-P. (2006) IMGT/LIGM-DB, the IMGT comprehensive database of immunoglobulin and T cell receptor nucleotide sequences. *Nucleic Acids Res.*, **34**, D781–D784.
14. Cornish-Bowden,A. (1985) Nomenclature for incompletely specified bases in nucleic acid sequences; recommendations 1984. *Nucleic Acids Res.*, **10**, 3021–3030.
15. Lefranc,M.-P. (2004) IMGT, the international ImMunoGenetics information system® http://imgt.cines.fr. In Lo,B.K.C. (ed.), *Antibody engineering methods and protocols. Methods in Molecular Biology* , Vol. 248, 2nd edn. Humana Press, Totowa, NJ, USA, pp. 27–49.
16. Giudicelli,V. and Lefranc,M.-P. (2005) Interactive IMGT on-line tools for the analysis of immunoglobulin and T cell receptor repertoires. In Veskler,B.A. (ed.), *New Research on Immunology*. Nova Science, New York, pp. 77–105.
17. Kaas,Q. and Lefranc,M.-P. (2007) IMGT Colliers de Perles: standardized sequence-structure representations of the IgSF and MhcSF superfamily domains. *Curr. Bioinformatics*, **2**, 21–30.
18. Elemento,O. and Lefranc,M.-P. (2003) IMGT/PhyloGene: an on-line tool for comparative analysis of immunoglobulin and T cell receptor genes. *Dev. Comp. Immunol.*, **27**, 763–779.
19. Souto-Carneiro,M.M., Longo,N.S., Russ,D.E., Sun,H.W. and Lipsky,P.E. (2004) Characterization of the human Ig heavy chain antigen binding complementarity determining region 3 using a newly developed software algorithm, JOINSOLVER. *J. Immunol.*, **172**, 6790–6802.
20. Ohm-Laursen,L., Nielsen,M., Larsen,S.R. and Barington,T. (2006) No evidence for the use of DIR, D-D fusions, chromosome 15 open reading frames or VH replacement in the peripheral repertoire was found on application of an improved algorithm, JointML, to 6329 human immunoglobulin H rearrangements. *Immunology*, **119**, 265–277.
21. Volpe,J.M., Cowell,L.G. and Kepler,T.B. (2006) SoDA: implementation of a 3D alignment algorithm for inference of antigen receptor recombinations. *Bioinformatics*, **22**, 438–444.
22. Gaëta,B.A., Malming,H.R., Jackson,K.J., Bain,M.E., Wilson,P. and Collins,A.M. (2007) iHMMune-align: hidden Markov model-based alignment and identification of germline genes in rearranged immunoglobulin gene sequences. *Bioinformatics*, **23**, 1580–1587.
23. Bleakley,K., Giudicelli,V., Wu,Y., Lefranc,M.-P. and Biau,G. (2006) IMGT standardization for statistical analyses of T cell receptor junctions: the TRAV-TRAJ example. *In Silico Biol.*, **6**, 573–588.
24. Magdelaine-Beuzelin,C., Kaas,Q., Wehbi,V., Ohresser,M., Jefferis,R., Lefranc,M.-P. and Watier,H. (2007) Structure–function relationships of the variable domains of monoclonal antibodies approved for cancer treatment. *Crit. Rev. Oncol. Hematol.*, **64**, 210–225.
25. Lefranc,M.-P. (2007) WHO-IUIS Nomenclature Subcommittee for Immunoglobulins and T cell receptors report. *Immunogenetics*, **59**, 899–902.
26. Lefranc,M.-P. (2008) WHO-IUIS Nomenclature Subcommittee for Immunoglobulins and T cell receptors report August 2007, 13th International Congress of Immunology, Rio de Janeiro, Brazil. *Dev. Comp. Immunol.*, **32**, 461–463.
27. Ghia,P., Stamatopoulos,K., Belessi,C., Moreno,C., Stilgenbauer,S., Stevenson,F., David,F. and Rosenquist,R. (2007) ERIC recommendations on IGHV gene mutational status analysis in chronic lymphocytic leukemia. *Leukemia*, **21**, 1–3.
28. Belessi,C.J., Davi,F.B., Stamatopoulos,K.E., Degano,M., Andreou,T.M., Moreno,C., Merle-Béral,H., Crespo,M., Laoutaris,N.P., Montserrat,E. *et al.* (2006) IGHV gene insertions and deletions in chronic lymphocytic leukemia: "CLL-biased" deletions in a subset of cases with stereotyped receptors. *Eur. J. Immunol.*, **36**, 1963–1974.
29. Murray,F., Darzentas,N., Hadzidimitriou,A., Tobin,G., Boudjogra,M., Scielzo,C., Laoutaris,N., Karlsson,K., Baran-Marzsak,F., Tsaftaris,A. *et al.* (2008) Stereotyped patterns of somatic hypermutation in subsets of patients with chronic lymphocytic leukemia: implications for the role of antigen selection in leukemogenesis. *Blood*, **111**, 1524–1533.
30. Davi,F., Rosenquist,R., Ghia,P., Belessi,C. and Stamatopoulos,K. (2008) Determination of IGHV gene mutational status in chronic lymphocytic leukemia: bioinformatics advances meet clinical needs. *Leukemia*, **22**, 212–214.

PUBLICATION 2

Available online at www.sciencedirect.com
ScienceDirect
Biochimie 90 (2008) 570–583
BIOCHIMIE
www.elsevier.com/locate/biochi

Research paper

IMGT-Kaleidoscope, the formal IMGT-ONTOLOGY paradigm

Patrice Duroux [a], Quentin Kaas [a], Xavier Brochet [a], Jérôme Lane [a], Chantal Ginestoux [a], Marie-Paule Lefranc [a,b,*], Véronique Giudicelli [a]

[a] *IMGT®, the international ImMunoGeneTics information system®, Université Montpellier 2, Laboratoire d'ImmunoGénétique Moléculaire LIGM, UPR CNRS 1142, Institut de Génétique Humaine IGH, 141 rue de la Cardonille, 34396 Montpellier Cedex 5, France*
[b] *Institut Universitaire de France, 103 Bd Saint-Michel, 75005 Paris, France*

Received 8 June 2007; accepted 4 September 2007
Available online 11 September 2007

Abstract

IMGT®, the international ImMunoGeneTics information system® (http://imgt.cines.fr), is the reference in immunogenetics and immunoinformatics. IMGT standardizes and manages the complex immunogenetic data which include the immunoglobulins (IG) or antibodies, the T cell receptors (TR), the major histocompatibility complex (MHC) and the related proteins of the immune system (RPI) which belong to the immunoglobulin superfamily (IgSF) and the MHC superfamily (MhcSF). The accuracy and consistency of IMGT data and the coherence between the different IMGT components (databases, tools and Web resources) are based on IMGT-ONTOLOGY, the first ontology for immunogenetics and immunoinformatics. IMGT-ONTOLOGY manages the immunogenetics knowledge through diverse facets relying on seven axioms, ''IDENTIFICATION'', ''DESCRIPTION'', ''CLASSIFICATION'', ''NUMEROTATION'', ''LOCALIZATION'', ''ORIENTATION'' and ''OBTENTION'', that postulate that objects, processes and relations have to be identified, described, classified, numerotated, localized, orientated, and that the way they are obtained has to be determined. These axioms constitute the Formal IMGT-ONTOLOGY, also designated as IMGT-Kaleidoscope. Through the example of the IG molecular synthesis, the concepts generated from the ''IDENTIFICATION'', ''DESCRIPTION'', ''CLASSIFICATION'' and ''NUMEROTATION'' axioms are detailed with their main instances and semantic relations. The axioms have been essential for the conceptualization of the molecular immunogenetics knowledge and can be used to generate concepts for multi scale approaches at the molecule, cell, tissue, organ, organism or population level, emphasizing the generalization of the application domain. In that way the Formal IMGT-ONTOLOGY represents a paradigm for the elaboration of ontologies in system biology.

Keywords: IMGT; Ontology; System biology; Immunogenetics; Immunoinformatics

1. Introduction

Immunogenetics, the science that studies the genetics of the immune responses, has shown a considerable expansion in biomedical fields since the last decades. It has highlighted the complex mechanisms by which B cells and T cells are at the origin of the extreme diversity of antigen receptors that comprise the immunoglobulins (IG) or antibodies and the T cell receptors (TR) (10^{12} different immunoglobulins and 10^{12} different T cell receptors per individual, in humans) [1,2]. These mechanisms include in particular DNA rearrangements [3] and, for the IG, somatic hypermutations [1,2]. In addition, there is a considerable polymorphism of the major histocompatibility complex (MHC), human leucocyte antigens (HLA) in humans. These particularities of the adaptive immune system of the vertebrates, and a better knowledge of the innate immune response found in any species, allow the immune system to be an excellent model for system biology. The huge amount of immunological experimental data

* Corresponding author. IMGT®, the international ImMunoGeneTics information system®, Université Montpellier 2, Laboratoire d'ImmunoGénétique Moléculaire LIGM, UPR CNRS 1142, Institut de Génétique Humaine IGH, 141 rue de la Cardonille, 34396 Montpellier Cedex 5, France. Tel.: +33 4 99 61 99 65; fax: +33 4 99 61 99 01.
E-mail address: marie-paule.lefranc@igh.cnrs.fr (M.-P. Lefranc).

0300-9084/$ - see front matter © 2007 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.biochi.2007.09.003

continues to grow exponentially and necessitates to be managed and analysed computationally. This is the goal of immunoinformatics, a new emerging science that implements the bioinformatics methodologies to answer these needs. At the same time, standardized representation of genomics, genetics, proteomics and three-dimensional (3D) structures data is required to organize immunogenetics knowledge towards system biology and for the modelling and a better understanding of the immune system.

IMGT®, the international ImMunoGeneTics information system® (http://imgt.cines.fr) [4], is the international reference in immunogenetics and immunoinformatics. Created in 1989 at the Laboratoire d'ImmunoGénétique Moléculaire (LIGM) by Marie-Paule Lefranc (Université Montpellier 2 and CNRS) in Montpellier, France, IMGT provides a high-quality integrated knowledge resource, specialized in the IG, TR, MHC of human and other vertebrates, and related proteins of the immune system (RPI), which belong to the immunoglobulin superfamily (IgSF) and to the MHC superfamily (MhcSF) of any species. The IMGT information system consists of databases (three of sequences, one of genes and one of 3D structures) and interactive tools for sequence, genome and 3D structure analysis, which interact together according to genomic, genetic and structural approaches [5] (Fig. 1).

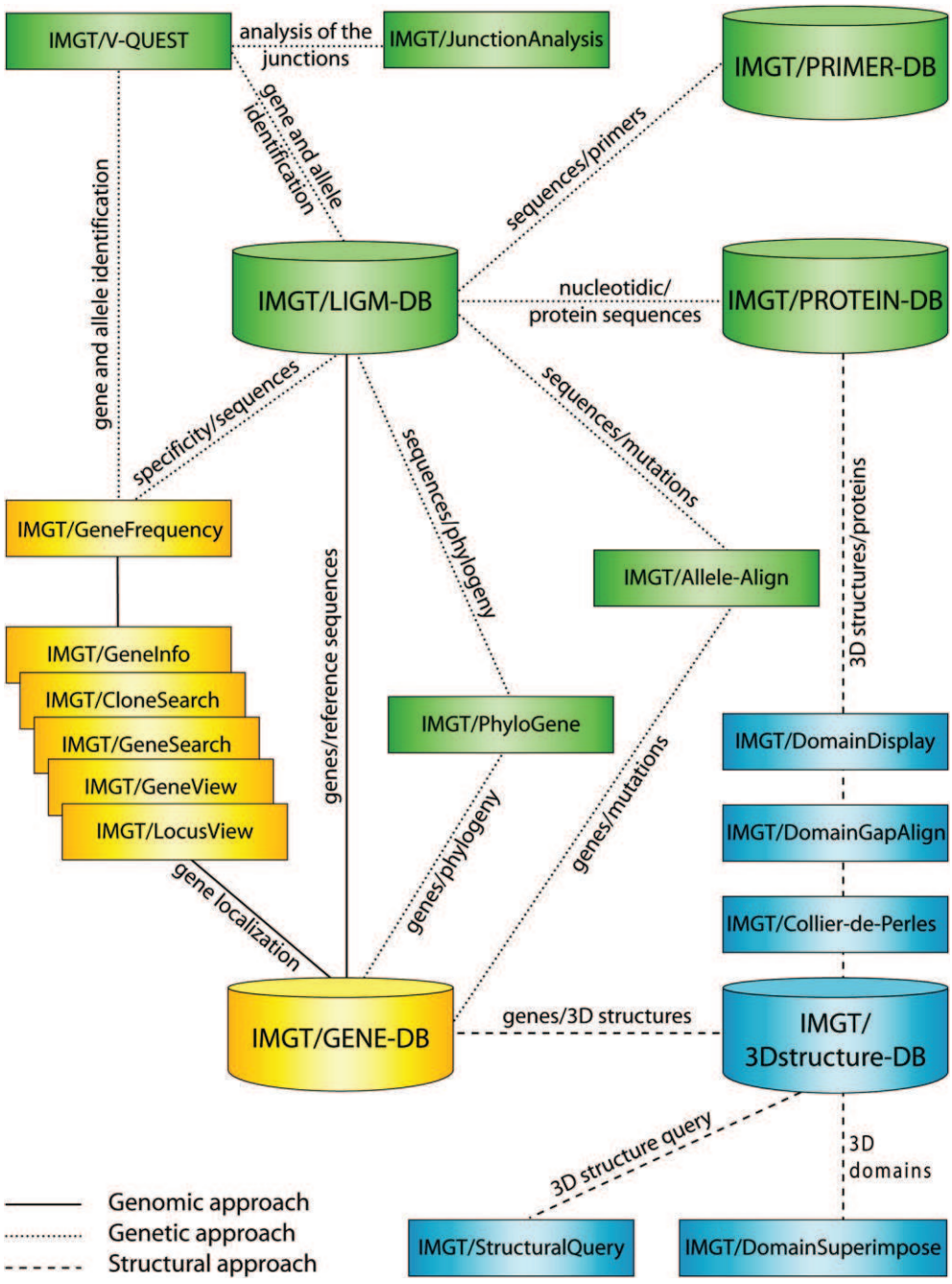


Fig. 1. IMGT® databases and tools with their interactions according to the genomic, genetic and/or structural approaches.

Moreover, IMGT provides Web resources comprising more than 10,000 HTML pages of synthesis (IMGT Repertoire), knowledge (IMGT Scientific chart, IMGT Education, IMGT Index) and external links (IMGT Bloc-notes and IMGT other accesses) [4].

The accuracy and the consistency of the IMGT data, as well as the coherence between the different IMGT components (databases, tools and Web resources), are based on IMGT-ONTOLOGY, the first ontology for immunogenetics and immunoinformatics [6]. IMGT-ONTOLOGY provides a semantic specification of the terms to be used in immunogenetics and immunoinformatics and manages the related knowledge, thus allowing the standardization for immunogenetics data from genome, proteome, genetics and 3D structures [7–9]. IMGT-ONTOLOGY results from a deep expertise in the domain and an extensive effort of conceptualization. The first standardization step was the identification of the IG and TR nucleotide sequences and the second step their description which led to the creation of IMGT/LIGM-DB [10], the first on-line IMGT database. The resulting controlled vocabulary comprises a thesaurus of keywords for the sequence identification and a set of labels for the description of the constitutive motifs. The third standardization step was the classification of the IG and TR genes which gave rise to the IMGT gene nomenclature for IG and TR of human and other vertebrates [1,2], approved by the Human Genome Organisation (HUGO) Nomenclature Committee HGNC in 1999 [11] and currently used in the generalist genome databases. The fourth standardization step was the setting up of the principles for the unique numbering of antigen receptor sequences and structures [12–16].

The standardization rules, defined in the IMGT Scientific chart [4], are based on the concepts of identification, description, classification and numerotation which characterize IMGT-ONTOLOGY [6] and which, interestingly, were defined before the term ''ontology'' became commonly used in biology and bioinformatics. IMGT-ONTOLOGY manages the immunogenetics knowledge through diverse facets that rely on the axioms of the Formal IMGT-ONTOLOGY or IMGT-Kaleidoscope. Four of these axioms, ''IDENTIFICATION'', ''DESCRIPTION'', ''CLASSIFICATION'' and ''NUMEROTATION'' are presented in this paper, with the concepts that have been essential for the conceptualization of the molecular immunogenetics knowledge. As the same axioms can be used to generate concepts for multi-scale level approaches, the Formal IMGT-ONTOLOGY represents a paradigm for system biology ontologies, which need to identify, to describe, to classify and to numerotate objects, processes and relations at the molecule, cell, tissue, organ, organism or population levels.

2. Methods

2.1. Terminology

An ontology is a formal representation of a knowledge domain [6,17–19]. IMGT-ONTOLOGY manages the immunogenetics knowledge through diverse facets relying on seven axioms, ''IDENTIFICATION'', ''CLASSIFICATION'', ''DESCRIPTION'', ''LOCALIZATION'', ''NUMEROTATION'', ''ORIENTATION'' and ''OBTENTION''. These axioms postulate that objects, processes and relations have to be identified, described, classified, numerotated, localized, orientated, and that the way they are obtained has to be determined (Fig. 2). The axioms constitute the Formal IMGT-ONTOLOGY, also designated as IMGT-Kaleidoscope.

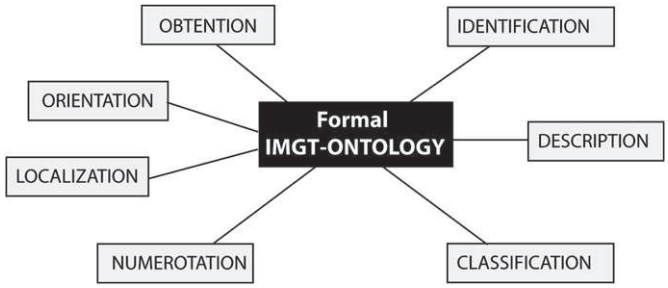


Fig. 2. The axioms of the Formal IMGT-ONTOLOGY or IMGT-Kaleidoscope.

Each axiom gives rise to a set of concepts. Concepts are general in the reality [6,20–23]. Concept instances correspond to all possible examples of representation of a concept at a given granularity. A concept may be exemplified by one or several concept instances. New concept instances may be defined with the advancement of science. Concepts are linked by relations, the simplest being ''is_a'' which represents the edge between concepts at different levels of granularity and organizes the main hierarchy of IMGT-ONTOLOGY. Properties are semantic characteristics of a concept or of a concept instance: they may be simple attributes as a name alias, or they may be specific relations between concepts and instances across the main hierarchy. These relations are fundamental since they reveal strong semantic constraints and dependencies on which relies the coherence within or between IMGT components.

2.2. An example of knowledge at the molecular level: the immunoglobulin synthesis

The immunoglobulin synthesis, an example of knowledge at the molecular level, will be used to define the concepts generated by four of the axioms of the Formal IMGT-ONTOLOGY or IMGT-Kaleidoscope. The concepts of identification (IDENTIFICATION axiom) identify the nucleotide and protein sequences and the 3D structures according to a structured terminology, the concepts of description (DESCRIPTION axiom) describe the composition of the sequences and structures with standardized labels, the concepts of classification (CLASSIFICATION axiom) classify the genes and alleles with a standardized nomenclature, and the concepts of numerotation (NUMEROTATION axiom) numerotate the nucleotide and amino acid numbering within sequences and structures.

An IG or antibody is composed of two identical heavy chains associated with two identical light chains, kappa or lambda. In humans, heavy chain genes (locus IGH), light chain kappa genes (locus IGK) and light chain lambda genes (locus IGL) are located on the chromosomes 14 (14q32.3), 2 (2p11.2) and 22 (22q11.2), respectively. The synthesis of an immunoglobulin

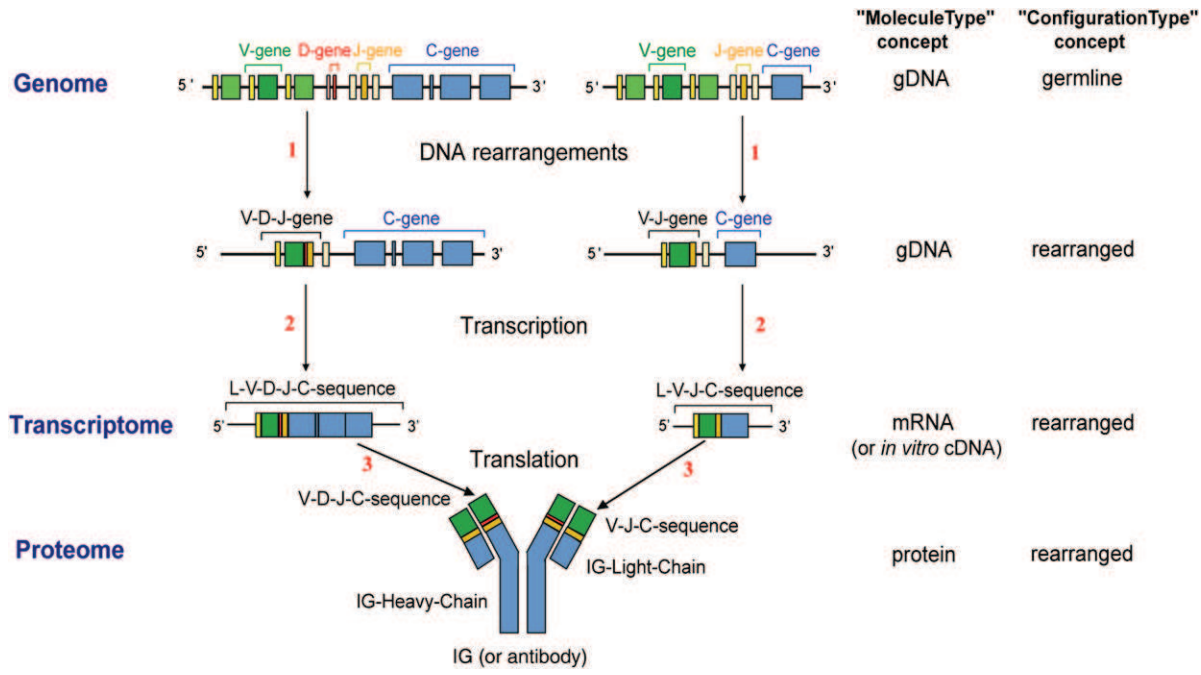


Fig. 3. An example of knowledge at the molecular level: the synthesis of an IG or antibody in humans. A human being may potentially synthesize 10^{12} different antibodies [1]. 1: DNA rearrangements (is_rearranged_into), 2: Transcription (is_transcribed_into), 3: Translation (is_translated_into). The configuration of C-gene is undefined.

requires rearrangements of the IGH, IGK and IGL genes during the differentiation of the B lymphocytes.

In the human genome (genomic DNA or gDNA), four types of genes code the IG (and TR): variable (V), diversity (D), joining (J) and constant (C). The configuration of the V-gene, D-gene and J-gene is identified as ''germline'' (Fig. 3), the configuration of the C-gene is ''undefined''. During the differentiation of the B lymphocytes in the bone marrow, the genomic DNA is rearranged first in the IGH locus, and then in the IGK and IGL loci. The rearrangements in the IGH locus lead to the junction of a D-gene and a J-gene to form a D-J-gene, and then to the junction of a V-gene to the D-J-gene to form a V-D-J-gene. The rearrangements in the IGK or IGL loci lead to the junction of a V-gene and a J-gene to form a V-J-gene. The configuration of these genes is identified as ''rearranged''. After transcription and maturation of the pre-messenger by splicing, the messenger RNA (mRNA) L-V-D-J-C-sequence and L-V-J-C-sequence (L for leader) are obtained and then translated into the heavy chain (IG-Heavy-Chain) and the light chain (IG-Light-Chain) of an IG (or antibody) (Fig. 3).

The variable domains VH and VL are coded by the V-D-J-REGION and the V-J-REGION (Fig. 4). Each domain includes four framework regions (FR) (in pale grey in Fig. 4) and three hypervariable loops or complementarity determining regions (CDR). The CDR, and more particularly the CDR3 that result from the junction of the V-D-J genes (in the VH domain) and V-J genes (in the VL domain), are involved in the antigen recognition. The VH and VL amino acids in contact with the antigen constitute the paratope. The part of the antigen recognized by the antibody is the epitope. The number of potential V-D-J and V-J rearrangements depends on the number of functional V, D and J genes in the genome. Additional mechanisms (N diversity at the V-D-J and V-J junctions and somatic hypermutations) allow to reach 10^{12} different antibodies per individual [1] (IMGT®, http://imgt.cines.fr).

2.3. Implementation

The main hierarchy of the IMGT-ONTOLOGY concepts has previously been described [6]. IMGT-ONTOLOGY

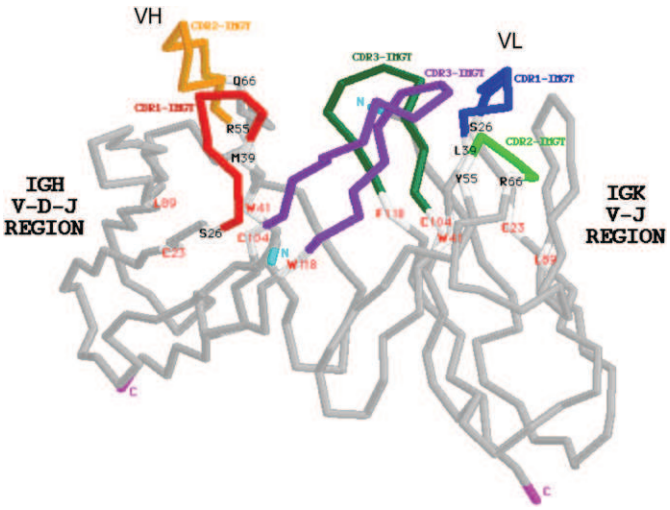


Fig. 4. The variable domains VH and VL of the heavy and light chains of an IG or antibody. VH CDR1-IMGT is in red, CDR2-IMGT in orange and CDR3-IMGT in purple. VL CDR1-IMGT is in blue, CDR2-IMGT in green and CDR3-IMGT in dark green.

concepts are available for the biologists and IMGT users in natural language in the IMGT Scientific chart [4], and have been formalized for programming purpose in IMGT-ML [24,25] which is an XML Schema (http://www.w3.org/TR/xmlschema-0/). In order to formalize the semantic relations between concepts and instances that are essential for high-quality data processing and coherence control, IMGT-ONTOLOGY is currently designed with Protégé [26] and OBO-Edit (http://oboedit.org/), that are frequently used ontology editors for biological ontologies. Protégé and OBO-Edit ontologies can be exported into RDF (http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/) and OWL (http://www.w3.org/2004/OWL/) which allow interoperability with other ontologies.

3. Results

3.1. The necessity of identification: the IDENTIFICATION axiom

The IDENTIFICATION axiom of the Formal IMGT-ONTOLOGY or IMGT-Kaleidoscope postulates that molecules, cells, tissues, organs, organisms or populations, their processes and relations, have to be identified. The IDENTIFICATION axiom has generated the concepts of identification which provide the terms and rules to identify an entity, its processes and its relations. In molecular biology, the concepts of identification allow to identify the molecules, their processes and their relations at the genome, transcriptome and proteome levels.

3.1.1. Identification of an organism: the ''Taxon'' concept

The ''Taxon'' concept allows to identify the type of taxon in which an object, process or relation is found. The ''Taxon'' concept manages a hierarchy of concepts at various levels of granularity. The corresponding hierarchical taxonomy is that provided by the National Center for Biotechnology Information NCBI (http://www.ncbi.nlm.nih.gov/) up to the rank of species (''Species'' concept) and subspecies (''Subspecies'' concept) in order to establish complete interoperability with generalist databases. Since IG, TR and MHC genes are only present in jawed vertebrates (gnathostoma), only vertebrate species were originally represented in IMGT-ONTOLOGY. However, with the extension of IMGT-ONTOLOGY to the IgSF and MhcSF, invertebrate species are incorporated whenever necessary. The ''EthnicGroup'', ''Breed'' and ''Strain'' concepts have been added to IMGT-ONTOLOGY to allow the identification of data specific to ethnic groups for humans (http://www.ebi.ac.uk/imgt/hla/help/ethnic_help.html), breeds for domestic animals or strains for laboratory [27] and wild animals.

3.1.2. Identification of an entity: the ''EntityType'' concept

The ''EntityType'' concept identifies the type of entity. An entity can be a molecule, a cell, a tissue, an organ, an organism or a population. If the object is a molecule, the ''EntityType'' concept is designated as ''Molecule_EntityType'', which is defined by the ''MoleculeType'', ''GeneType'' and ''ConfigurationType'' concepts of identification and has properties identified in the ''Functionality'' and ''StructureType'' concepts (Fig. 5).

3.1.2.1. The ''MoleculeType'' concept. The ''MoleculeType'', concept identifies the type of molecule based on the type of the constitutive elements and on the concepts of obtention (not detailed here). The four main instances of the ''MoleculeType'' concept are 'gDNA' (genomic DNA, a nucleotide sequence made of A, T, C, G, obtained from a genome), 'mRNA' (messenger RNA or transcript, a nucleotide sequence made of A, U, C, G, obtained by transcription of a genomic DNA), 'cDNA' (complementary DNA, a nucleotide sequence made of A, T, C, G, obtained *in vitro* by reverse transcription of the messenger RNA) and 'protein' (a sequence made of amino acids, obtained by translation of a transcript). Thus, the instances of the ''MoleculeType'' concept allow to identify a sequence: nucleotide sequence that can be either genomic ('gDNA') or a transcript ('mRNA', 'cDNA'), and amino acid sequence ('protein').

3.1.2.2. The ''GeneType'' concept. The ''GeneType'' concept identifies the type of gene and comprises five instances (Fig. 5). The first instance, 'conventional', refers to any (coding or not coding) gene other than IG or TR genes. The other four instances are specific to immunogenetics: 'variable' (V), 'diversity' (D) and 'joining' (J) gene types that rearrange at the DNA level and code the variable domains of IG and TR, and 'constant' (C) gene type that codes the constant region of IG and TR [1,2].

3.1.2.3. The ''ConfigurationType'' concept. The ''ConfigurationType'' concept identifies the type of gene configuration and comprises three instances (Fig. 5). The instance 'undefined' identifies the configuration of the conventional and of the constant (C) genes. The instances 'germline' and 'rearranged' identify the status of the V, D and J genes, before and after DNA rearrangements, respectively [1,2].

3.1.2.4. The ''Molecule_EntityType'' concept. The ''Molecule_EntityType'' concept, defined by the ''MoleculeType'', ''GeneType'' and ''ConfigurationType'' concepts, includes 19 instances. Three instances, 'gene', 'nt-sequence' and 'AA-sequence', respectively identify the gDNA, mRNA and protein (''MoleculeType'') of a conventional gene (''GeneType'') in undefined configuration (''ConfigurationType''). The nt-sequence instance is also valid for cDNA. Sixteen instances allow to identify the IG and TR. Ten of them are represented in Fig. 3: six for the gDNA ('V-gene', 'D-gene', 'J-gene', 'C-gene', 'V-D-J-gene' and 'V-J-gene'), two for the mRNA, 'L-V-D-J-C-sequence' and 'L-V-J-C-sequence', also valid for cDNA, and two for the protein, 'V-D-J-C-sequence' and 'V-J-C-sequence'. For example, the instance 'V-gene' identifies a gDNA (''MoleculeType'') containing a gene V (''GeneType''), in germline configuration (''ConfigurationType''). The instance 'L-V-J-C-sequence' identifies a sequence of mRNA or cDNA (''MoleculeType'') corresponding to V, J and C genes (''GeneType''), in rearranged configuration

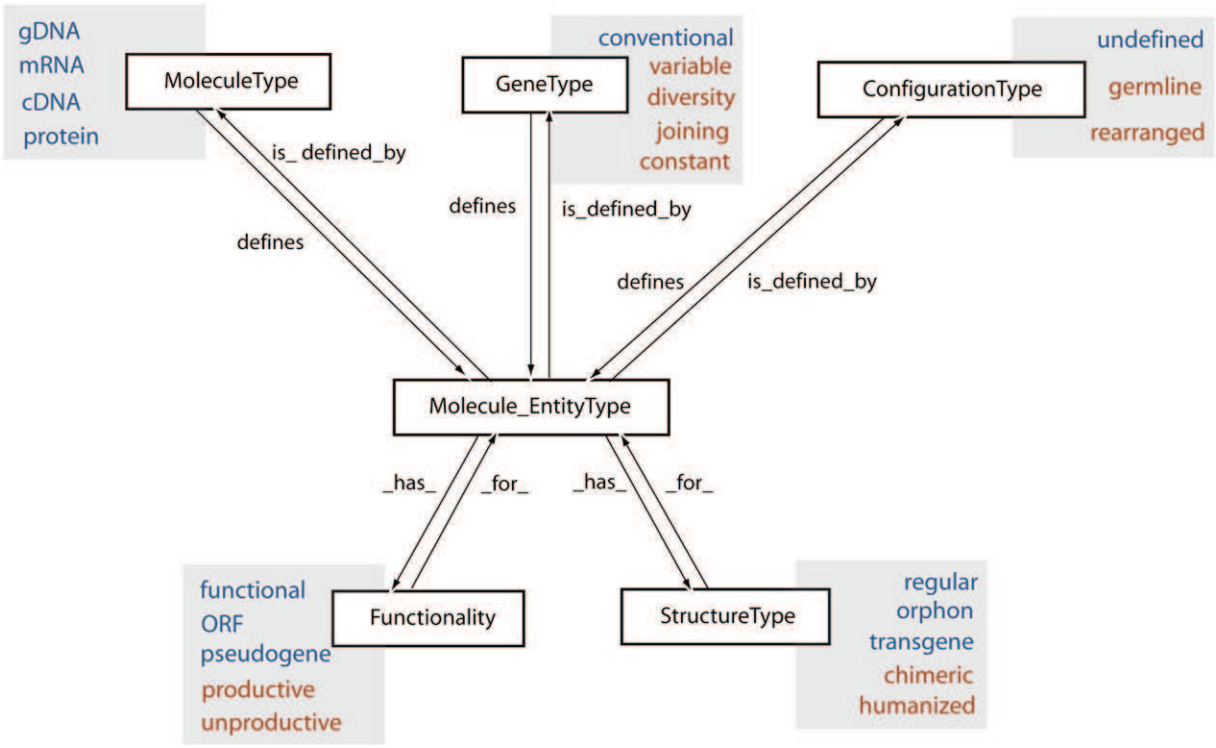


Fig. 5. The ‘‘Molecule_EntityType’’ concept. The ‘‘Molecule_EntityType’’ concept is defined by the ‘‘MoleculeType’’, ‘‘GeneType’’ and ‘‘ConfigurationType’’ concepts of identification and has properties identified in the ‘‘Functionality’’ and ‘‘StructureType’’ concepts (IDENTIFICATION axiom). Arrows indicate reciprocal relations ‘‘is_defined_by’’ and ‘‘defines’’, ‘‘_has_’’ and ‘‘_for_’’. Concept instances which are general are in blue, those which are specific of the IG and TR are in red. The ‘‘Molecule_EntityType’’ concept has 19 instances (listed in Section 3.1.2.4). Only a few examples of the ‘‘StructureType’’ concept instances are shown.

(‘‘ConfigurationType’’) (Fig. 3). The last six instances correspond to partial rearrangement (‘D-J-gene’) or to sterile transcripts (‘L-V-sequence’, ‘D-sequence’, ‘J-sequence’, ‘J-C-sequence’ and ‘C-sequence’).

3.1.2.5. The ‘‘Functionality’’ concept. The ‘‘Functionality’’ concept identifies the type of functionality for the ‘‘Molecule_EntityType’’ concept (Fig. 5). It includes five instances, divided into two categories, according to the configuration type. Three instances, ‘functional’, ‘ORF’ (open reading frame) and ‘pseudogene’ identify the functionality of a ‘‘Molecule_EntityType’’ instance in undefined or germline configuration. They allow to identify the functionality of conventional genes, that of C genes, and that of V, D and J genes before their rearrangement in the genome, and by extension the functionality of their transcripts and proteins. The two instances ‘productive’ and ‘unproductive’ identify the functionality of ‘‘Molecule_EntityType’’ instances in rearranged configuration. They allow to identify the functionality of IG and TR entities after their rearrangement in the genome, that of fusion genes resulting from translocations, and that of hybrid genes obtained by biotechnology molecular engineering, and by extension the functionality of their transcripts and proteins.

3.1.2.6. The ‘‘StructureType’’ concept. The ‘‘StructureType’’ concept identifies the structure for the ‘‘Molecule_EntityType’’ concept. This concept allows to identify structures with a classical organization (‘regular’), from those which have been modified either naturally *in vivo* (‘orphon’, ‘processed orphon’, ‘unprocessed orphon’, ‘unspliced’, ‘partially spliced’, etc.), or artificially *in vitro* (‘chimeric’, ‘humanized’, transgene, etc.).

3.1.3. Identification of a receptor: the ‘‘ReceptorType’’ concept

The ‘‘ReceptorType’’ concept identifies the type of receptor. A receptor can be a molecule, a cell, a tissue, an organ, an organism or a population. If the object is a molecule, the ‘‘ReceptorType’’ concept is designated as ‘‘Molecule_ReceptorType’’ which is defined by the ‘‘ChainType’’ concept of identification and has properties identified in the ‘‘StructureType’’, ‘‘Specificity’’ and ‘‘Function’’ concepts (Fig. 6). The ‘‘ChainType’’ concept is itself defined by the ‘‘Molecule_EntityType’’ and the ‘‘DomainType’’ concepts of identification and by concepts of classification (see CLASSIFICATION axiom). These latter are organized in a hierarchy which confers different levels of granularity to the ‘‘Molecule_ReceptorType’’ and ‘‘ChainType’’ concepts.

3.1.3.1. The ‘‘Molecule_ReceptorType’’ concept. The ‘‘Molecule_ReceptorType’’ concept identifies the type of protein receptor, defined by its chain composition. Thus, IG is an instance of the ‘‘Molecule_ReceptorType’’ concept, defined as comprising 4 chains, two heavy chains and two light chains,

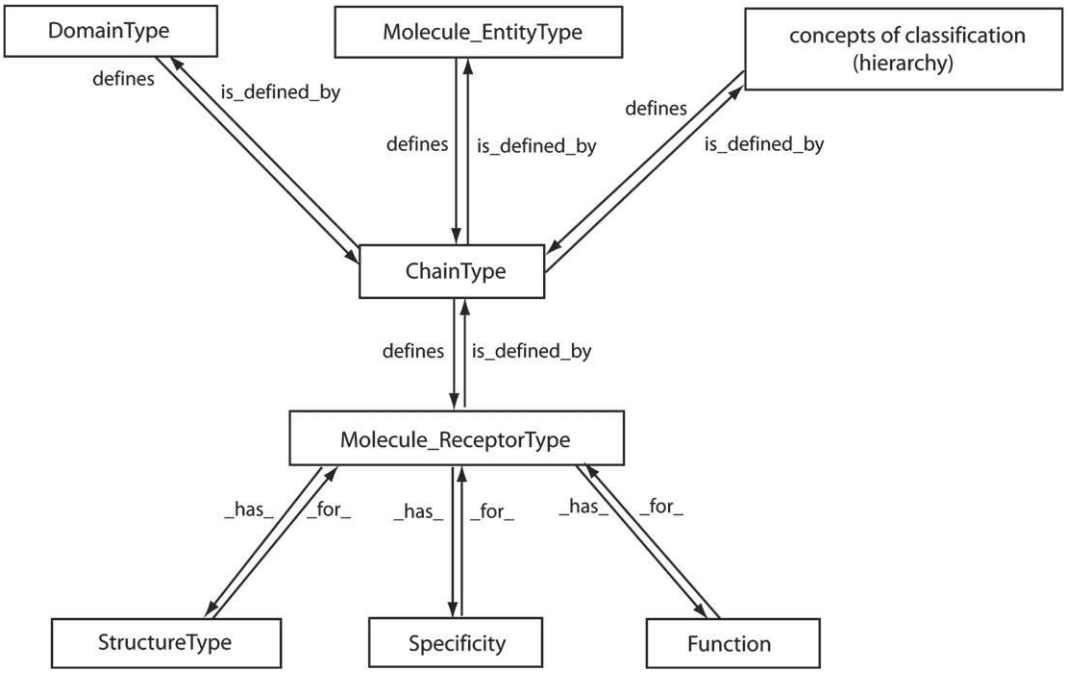


Fig. 6. The ''Molecule_ReceptorType'' concept. The ''Molecule_ReceptorType'' concept, defined by the ''ChainType'' concept of identification, has properties identified in the ''StructureType'', ''Specificity'' and ''Function'' concepts (IDENTIFICATION axiom). The ''ChainType'' concept is itself defined by the ''Molecule_EntityType'' and ''DomainType'' concepts and by concepts of classification (hierarchy). Arrows indicate reciprocal relations ''is_defined_by'' and ''defines'', ''_has_'' and ''_for_''. These concepts have different levels of granularity, up to six for ''Molecule_ReceptorType'' and ''ChainType''.

identical two by two and covalently linked (Fig. 7). A receptor may comprise one chain (monomer) or several associated chains (multimer).

3.1.3.2. The ''ChainType'' concept. The ''ChainType'' concept identifies the type of chain (Fig. 6). It is one of the most important concepts of identification for the standardization of genome, transcriptome and proteome data in system biology. Indeed, being able to identify a type of chain means that it is possible to identify the transcript and the encoding gene(s). The ''ChainType'' concept contains a hierarchy of concepts which identify the chain type at different levels of granularity. The finest level of granularity, the ''GeneLevelChainType'' concept, identifies the type of chain by reference to the gene(s) which code(s) the chain. It represents the main concept for a very precise identification because it establishes a relationship with the ''Gene'' concept which belongs to the concepts of classification (reciprocal relations ''is_coded_by'' and ''codes''). The number of instances of the ''GeneLevelChainType'' concept depends on the number of functional genes and ORF per haploid genome in a given species (in the case of the IG and TR, it is the number of functional and ORF constant genes which is taken into account). If only the functional genes are considered, the instances of this concept correspond to the isotypes.

3.1.3.3. The ''DomainType'' concept. A chain can be defined by its constitutive structural units (''DomainType'' concept) (Fig. 6). A domain is a chain subunit characterized by its three-dimensional (3D) structure, and by extension its amino acid sequence and the nucleotide sequence which encodes it. This concept may theoretically comprise many instances, but so far only the instances which have been carefully characterized by LIGM have been entered in IMGT-ONTOLOGY. The ''DomainType'' concept has currently three instances, V type domain (variable domains of the IG and TR and V-like domains of other IgSF proteins), C type domain (constant domains of the IG and TR and C-like domains of other IgSF proteins) and G type domain (groove domains of the MHC and G-like domains of other MhcSF proteins) [14–16].

3.1.3.4. The ''Specificity'' and ''Function'' concepts. The ''Specificity'' concept identifies the specificity of the ''Molecule_ReceptorType'' (Fig. 6), and by extension the specificity of the chains and domains and of the corresponding transcripts. Instances of the ''Specificity'' concept identify the antigen recognized by an antigen receptor (IG or TR). The ''Specificity'' concept is particularly important because of the unlimited number of antigens and of the complexity of the antigen/antigen receptor interactions. The conceptualization of knowledge associated with this concept is in the course of modelling. The instances of the ''Specificity'' concept (several hundreds at the present time) will be connected on the one hand, with the ''Epitope'' concept which identifies the part of the antigen recognized by the antigen receptor and on the other hand, with the ''Paratope'' concept which identifies the part of the antigen receptor (IG or TR) which recognizes and binds to the antigen. The ''Function'' concept identifies the function of the ''Molecule_ReceptorType'' (Fig. 6), and by extension the function of the chains and domains and of the corresponding

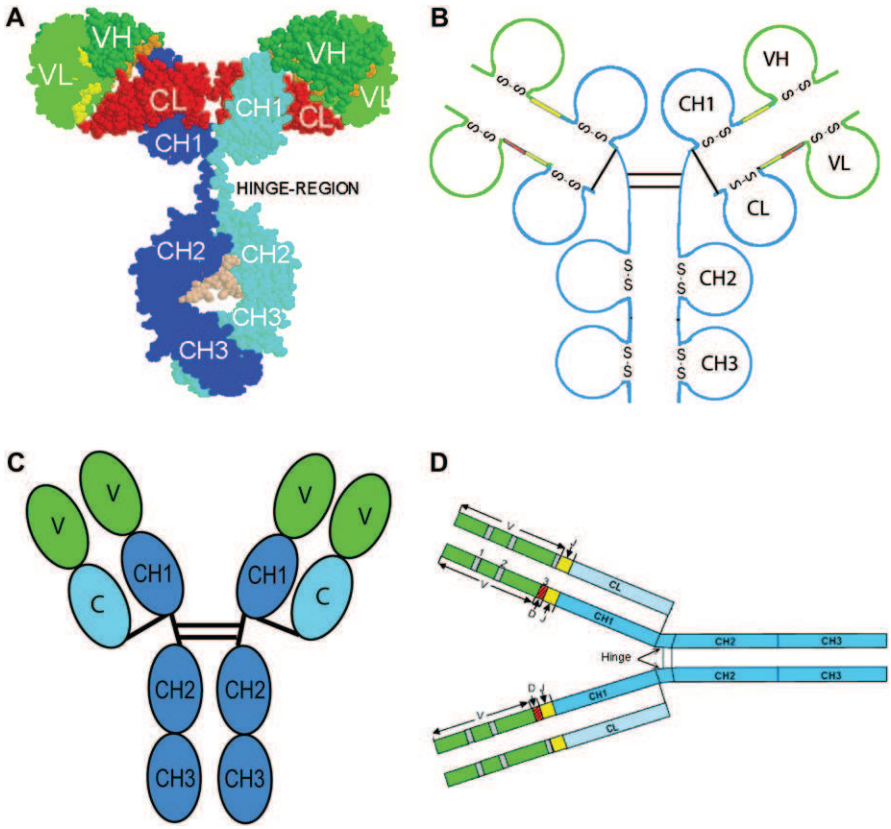


Fig. 7. Identification of an IG or antibody as an instance of the ''Molecule_ReceptorType'' concept made of four chains, two IG-Heavy-Chain and two IG-Light-Chain (''ChainType'' concept). The four representations, although different, allow to identify an IG as a receptor of four chains, themselves organized in domains (''DomainType'' concept). VH and VL are V type domains, coded by the V-D-J region and V-J region, respectively. CL, CH1, CH2 and CH3 are C type domains. (A) 3D structure, (B) organization in Ig-like domains, (C) organization in modules, (D) regions coded by the V, D, J and C gene types. The C gene type codes the constant region (CL for the IG-Light-Chain and CH1, hinge, CH2 and CH3 for the IG-Heavy-Chain). This representation, schematized as a Y shape, is frequently used to represent an IG.

transcripts. Instances of the ''Function'' concept identify the dual function of the antigen receptors [2]. Their identification and definition are still in development.

3.2. The necessity of description: the DESCRIPTION axiom

The DESCRIPTION axiom of the Formal IMGT-ONTOLOGY or IMGT-Kaleidoscope postulates that molecules, cells, tissues, organs, organisms or populations, their processes and their relations, have to be described.

3.2.1. Description of an entity: the ''EntityPrototype'' concept

The ''EntityPrototype'' concept, generated from the DESCRIPTION axiom, provides the description of the ''EntityType'' concept (IDENTIFICATION axiom). Each instance of the ''EntityPrototype'' concept is linked to an instance of the ''EntityType'' concept by the reciprocal relations ''describes'' and ''is_described_by''. The ''EntityPrototype'' concept allows the description of the entity organization and of its constitutive motifs. The ''Core'' concept allows to describe the parts of the entities which need to be described in all instances of the ''EntityPrototype'' concept. These two concepts of description, ''Molecule_EntityPrototype'' and ''Core'', which have been particularly highlighted by IMGT, are described below as examples.

3.2.2. The ''Molecule_EntityPrototype'' concept

In molecular biology, the DESCRIPTION axiom has generated the concepts of description which provide the terms and the rules to describe motifs in the nucleotide and protein sequences and in 3D structures. These concepts gave rise to a standardized terminology and to a precise definition of the annotation rules. The ontology for sequences and 3D structures has been the focus of IMGT for many years. The instances of the concepts of description correspond to IMGT labels. More than 550 labels were defined (270 for the nucleotide sequences (http://imgt.cines.fr/cgi-bin/IMGTlect.jv?query=7) [10] and 285 for the 3D structures [28] (http://imgt.cines.fr/textes/IMGTScientificChart/SequenceDescription/IMGT3Dlabeldef.html). Interestingly, 64 IMGT labels defined for nucleotide sequences are used and cross-referenced in the recently created Sequence Ontology (SO) (http://song.sourceforge.net/) [29] to describe specific IG and TR gene organization (http://imgt.cines.fr/textes/IMGTindex/ontology.html).

The "Molecule_EntityPrototype" concept allows the description of the entity (gene, transcrit and protein) organization and of their constitutive motifs. This concept is fundamental in IMGT-ONTOLOGY because it allows the representation of the knowledge related to the complex mechanisms of IG and TR gene rearrangements (Fig. 8). The relation "is_rearranged_into" is specific to the synthesis of the IG and TR. The relations "is_transcribed_into" and "is_translated_into" are general for molecular biology. These three relations allow the organization of the various instances of the "Molecule_EntityPrototype" concept during the synthesis of the IG and the TR, and in a more general way for the expression of any protein. They allow in addition, by more specific relations, to take into account the alternative transcripts, the protein isoforms and the post-translational modifications.

Each of the 19 instances of the concept "Molecule_EntityPrototype" can be described with its constitutive motifs which belong to the other concepts of description. Thus Fig. 9 shows as examples the graphical representation of the V-GENE and V-D-J-GENE instances with their constitutive motifs.

A set of ten relations are necessary and sufficient to compare the localization of the motifs of an instance of the concept "Molecule_EntityPrototype" (Table 1). These relations are part of the concepts of localization (LOCALIZATION axiom) (IMGT Index, http://imgt.cines.fr).

3.2.3. The "Core" concept

The "Core" concept allows to describe the coding region of genes and contains five instances which are 'REGION' (for conventional gene type), 'V-REGION', 'D-REGION', 'J-REGION' and 'C-REGION' (for V, D, J and C gene types, respectively). These instances are particularly important since they can be described in all the instances of the "Molecule_EntityPrototype" concept. They allow to describe the chains of the antigen receptors in spite of the complexity of their structure and to link sequences, structures and functions. Moreover, these are the instances of the "Core" concept which allowed the definition and standardized description of the IG and TR alleles (concepts of classification), now approved at the international level [1,2].

3.3. The necessity of classification: the CLASSIFICATION axiom

The CLASSIFICATION axiom of the Formal IMGT-ONTOLOGY or IMGT-Kaleidoscope postulates that molecules, cells, tissues, organs, organisms or populations, their processes and their relations, have to be classified. In molecular biology, the concepts of classification generated from the CLASSIFICATION axiom allow to classify and name the genes and their alleles. The genes which code the IG and TR belong to highly polymorphic multigenic families. A major contribution of IMGT-ONTOLOGY was to set the principles of their classification and to propose a standardized nomenclature [1,2] (Fig. 10). The IMGT gene nomenclature has been approved at the international level by the Human Genome Organisation (HUGO) Nomenclature Committee (HGNC), in 1999 [11]. The IMGT IG and TR gene names are the official reference for the genome projects and, as such, have been integrated in the Genome Database (GDB), in LocusLink and in Entrez Gene at NCBI [30]. The IG and TR genes [1,2] are managed in the IMGT/GENE-DB database [31].

3.3.1. The "Group" and "Subgroup" concepts

The "Group" concept classifies a set of genes which belong to the same multigene family, within the same species

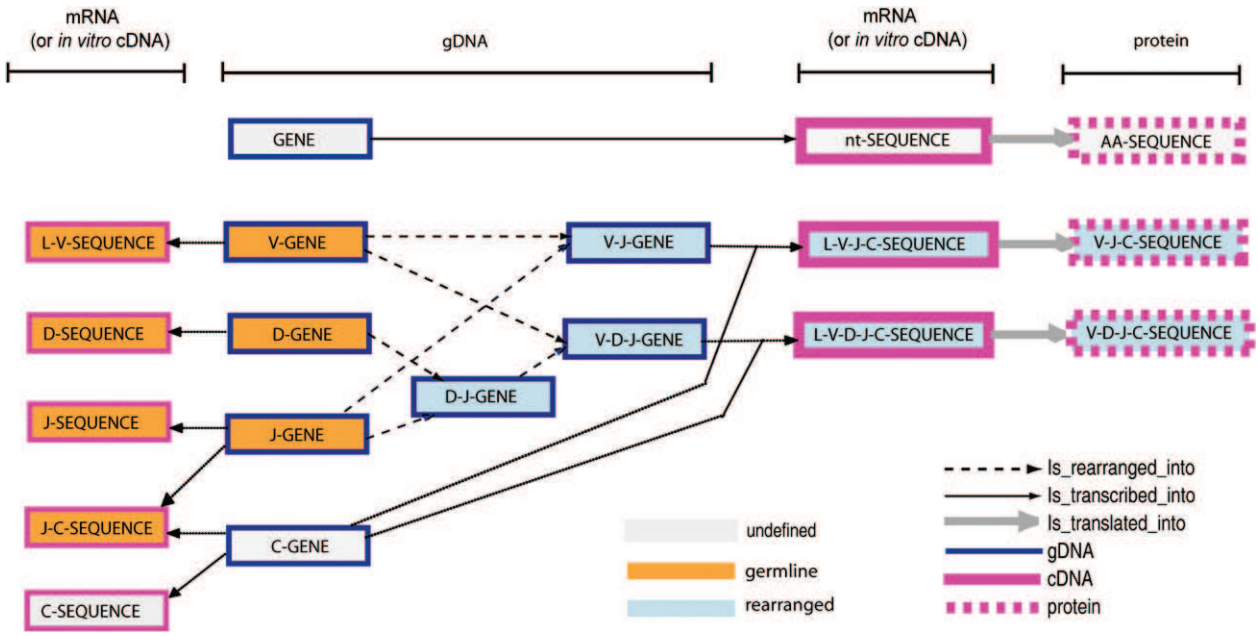


Fig. 8. Instances of the "Molecule_EntityPrototype" concept (DESCRIPTION axiom). The three instances "GENE", "nt-SEQUENCE" and "AA-SEQUENCE" correspond to conventional genes while the 16 other instances are specific of the IG and TR. The concept instances for mRNA are also valid for *in vitro* cDNA. The first column corresponds to 'sterile transcript' instances.

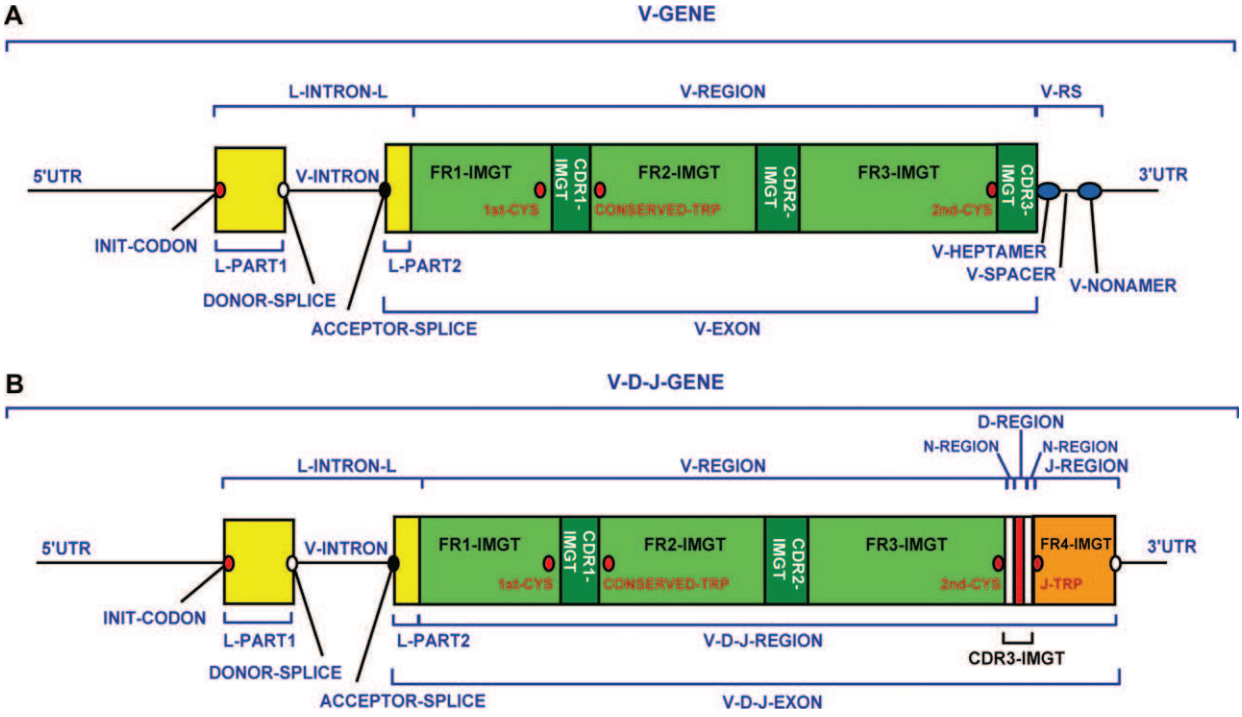


Fig. 9. Graphical representation of two instances of the ''Molecule_EntityPrototype'' concept (DESCRIPTION axiom). (A) V-GENE. (B) V-D-J-GENE. Twenty-five labels and ten relations are necessary and sufficient for a complete description of these instances.

or between different species. For the IG and TR, the set of genes is identified by an instance of the ''GeneType'' concept (V, D, J or C). The ''Subgroup'' concept classifies a subset of genes which belong to the same group, and which, in a given species, share at least 75% of identity at the nucleotide sequence level (and in the germline configuration for the V, D, and J genes).

3.3.2. The ''Gene'' and ''Allele'' concepts

The ''Gene'' concept classifies a unit of DNA sequence that can be potentially transcribed and/or translated (this definition includes the regulatory elements in 5′ and 3′, and the introns, if present). The instances of the ''Gene'' concept are gene names. In IMGT-ONTOLOGY, a gene name is composed of the name of the species (instance of the Taxon ''Species'' concept) and of the international HGNC/IMGT gene symbol, for example, *Homo sapiens* IGLV1−2. By extension, orphons and pseudogenes are also instances of the ''Gene'' concept. The ''Allele'' concept classifies a polymorphic variant of a gene. The instances of the ''Allele'' concept are allele names. Alleles identified by the mutations of the nucleotide sequence are classified by reference to allele *01.

Table 1
Relations for sequence description (LOCALIZATION axiom)

Relation	Reciprocal relation
''adjacent_at_its_5_prime_to''	''adjacent_at_its_3_prime_to''
''included_with_same_5_prime_in'',	''includes_with_same_5_prime'',
''included_with_same_3_prime_in'',	''includes_with_same_3_prime'',
''overlaps_at_its_5_prime_with''	''overlaps_at_its_3_prime_with''
''included_in''	''includes''

Full description of mutations and allele name designations are currently recorded for the core sequences (V-REGION, D-REGION, J-REGION, C-REGION). They are reported in Alignment tables, in IMGT Repertoire http://imgt.cines.fr and in IMGT/GENE-DB [16].

3.4. The necessity of numbering: the NUMEROTATION axiom

The NUMEROTATION axiom of the Formal IMGT-ONTOLOGY or IMGT-Kaleidoscope postulates that molecules, cells, tissues, organs, organisms or populations, their processes and their relations, have to be numerotated. So far, these concepts have essentially been defined at the molecular level. The NUMEROTATION axiom and the concepts of numerotation determine the principles of a unique numbering for a domain (sequences and 3D structures) [14−16] (Fig. 11). The ''IMGT_unique_numbering'' concept has three concept instances: ''IMGT_unique_numbering_for_V_Type_domain'', ''IMGT_unique_numbering_for_C_type_domain'', ''IMGT_unique_numbering_for_G_type_domain'' [14−16].

The ''IMGT_unique_numbering'' concept determines the ''FR-IMGT_length'', ''CDR-IMGT_length'', ''Strand_length'', and ''Helix_length'' concepts [14−16]. The ''IMGT_unique_numbering'' concept is illustrated by the ''IMGT_Collier_de_Perles'' concept which allows graphical representation in two dimensions (2D) of the amino acid sequences of V, C or G type domains [32,33] and comprises three concept instances (Fig. 12). This concept is largely recognized at the international level and the expression ''IMGT Collier de Perles'' is now used in scientific publications.

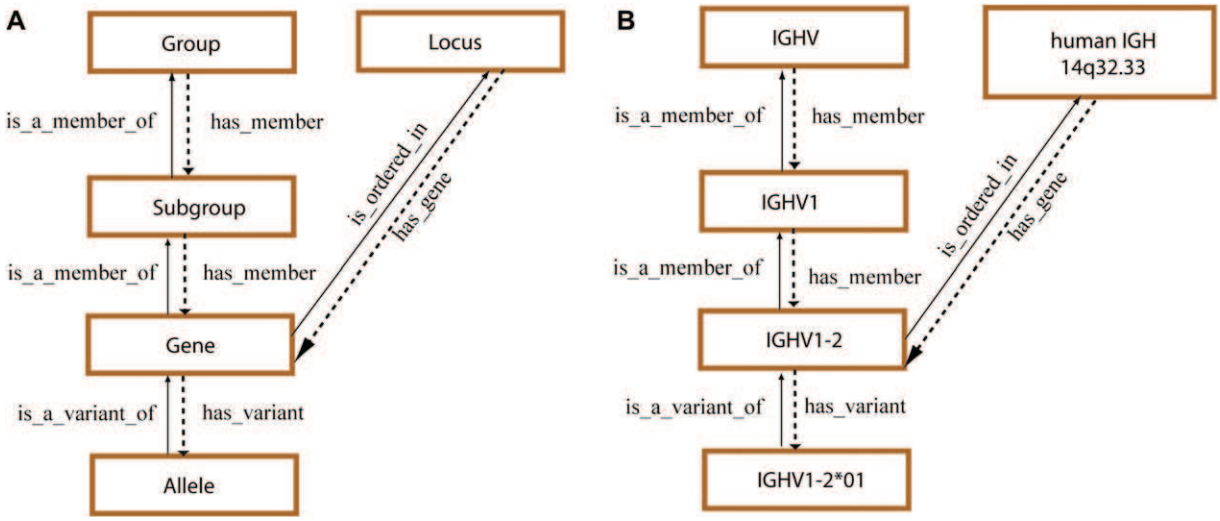


Fig. 10. Concepts of classification for gene and allele nomenclature (CLASSIFICATION axiom). (A) Hierarchy of the concepts of classification and their relations. (B) Examples of concept instances for each concept of classification. The concepts instances are associated to an instance of the "Taxon" concept, and more precisely for the "Gene" and "Allele" concepts to an instance of the "Species" concept (here, *Homo sapiens*). The "Locus" concept is a concept of localization (LOCALIZATION axiom).

The "IMGT_Collier_de_Perles" concept is particularly used in antibody engineering for the humanization of murine antibodies in which it is necessary to precisely delimit the murine CDR-IMGT to be grafted, in order to preserve the antibody specificity. The concepts of numerotation are also at the origin of the standardization of the allele description and, more generally of the mutation description (IMGT Scientific chart, http://imgt.cines.fr).

4. Conclusion

The inherent difficulties due to the complexity and diversity of immunogenetics knowledge gave rise to a conceptualization in IMGT-ONTOLOGY which has been developed on an original and unprecedented approach. The axioms of the Formal IMGT-ONTOLOGY or IMGT-Kaleidoscope postulate that the approach to manage biological data and to represent

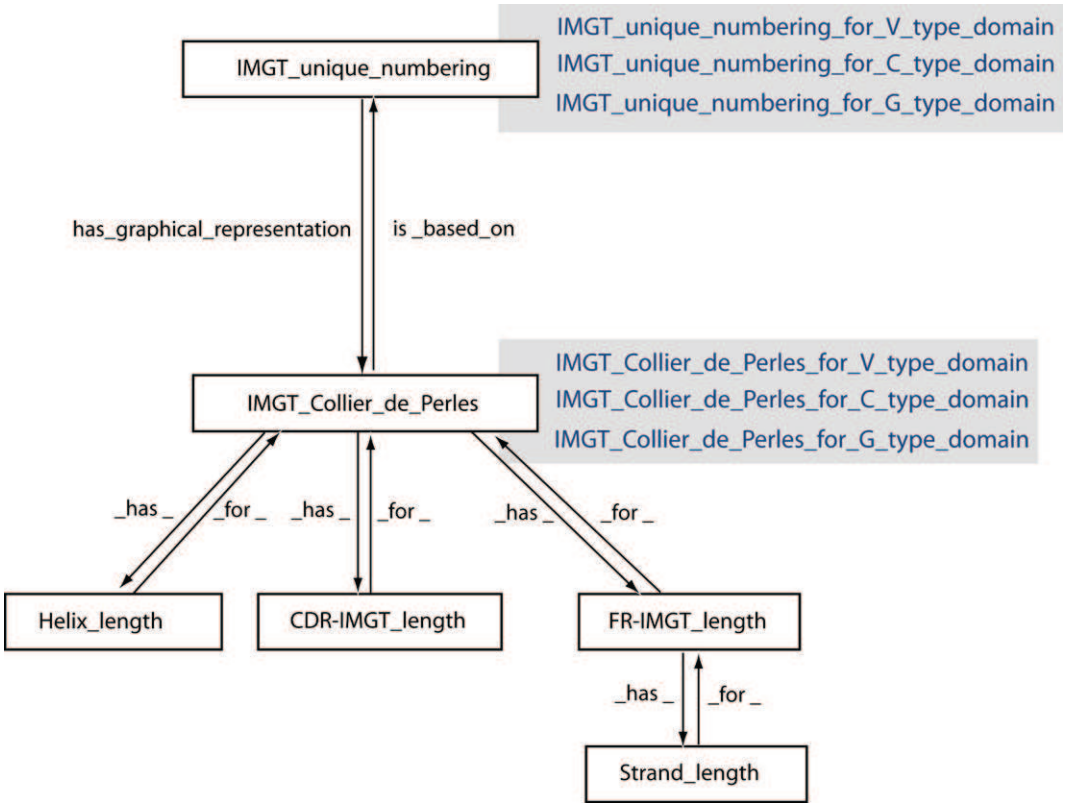


Fig. 11. The "IMGT_unique_numbering" and "IMGT_Collier_de_Perles" concepts and relations with other concepts of numerotation (NUMEROTATION Axiom). Concept instances are written in blue. Arrows indicate reciprocal relations "has_graphical_representation" and "is_based_on", "_has_" and "_for_".

knowledge in biology comprises various facets. The IMGT-ONTOLOGY concepts generated from these axioms have allowed the representation, at the molecular level, of knowledge related to the genome, transcriptome, proteome, genetics and 3D structures. This multi-faceted approach has great potential for multi-scale system biology. Indeed, the IDENTIFICATION, DESCRIPTION, CLASSIFICATION and NUMEROTATION axioms are valid, not only for molecules, but also for cells, tissues, organs, organisms or populations. In addition, the LOCALIZATION, ORIENTATION and OBTENTION axioms (in development) will allow the integration of the time and space concepts and the follow-up of the components and their changes of states and properties, as well as the definition and characterization of processes, functions and activities. Thus, IMGT-ONTOLOGY represents, by its 7 axioms and the concepts generated from them, a paradigm for the elaboration of ontologies in system biology which requires to identify, to describe, to classify, to numerotate, to localize, to orientate and to determine

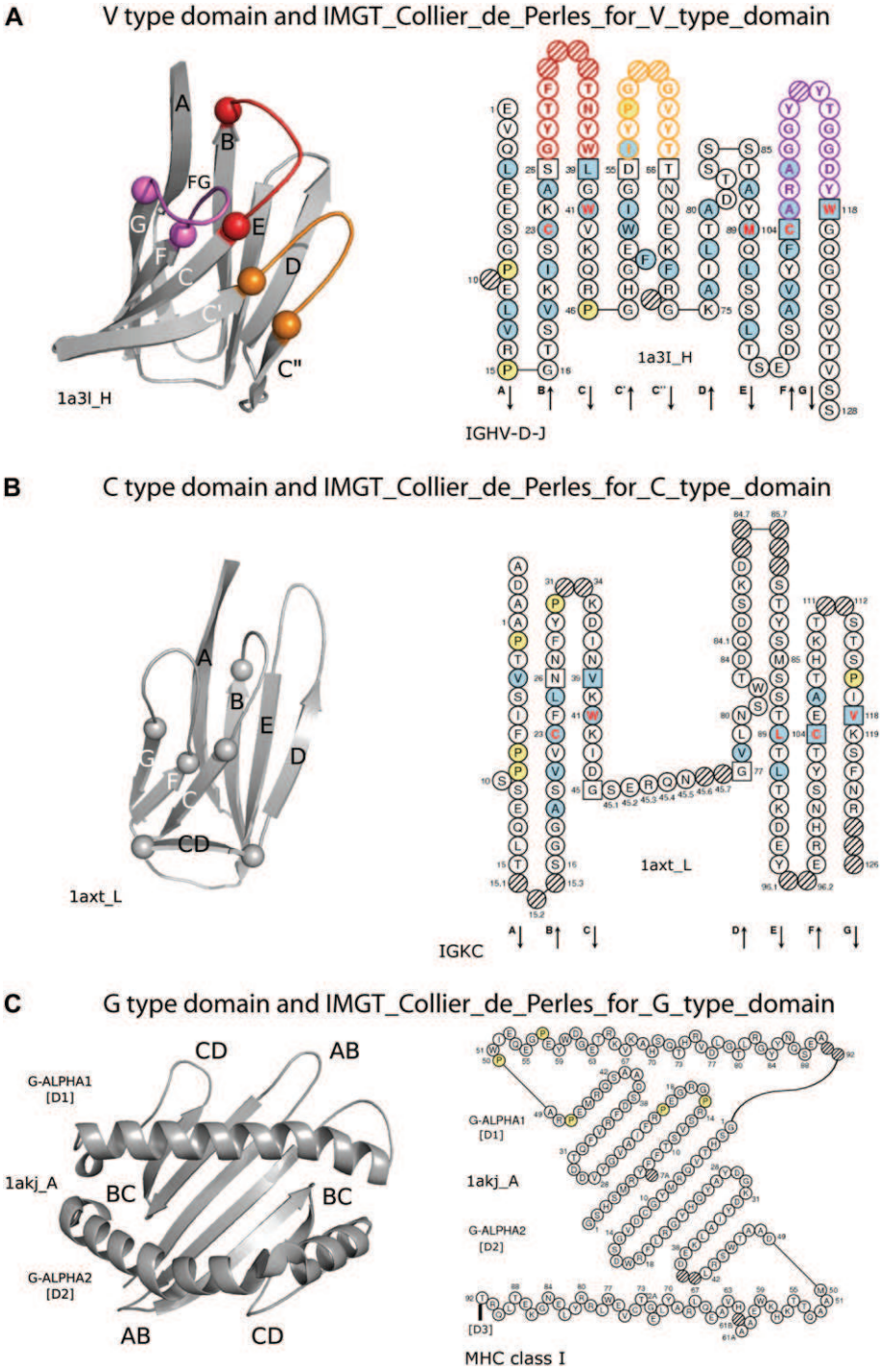


Fig. 12. "DomainType" and "IMGT_Collier_de_Perles" concept instances.

the obtaining and evolution of biological knowledge from molecule to population, in time and space.

The concepts of IMGT-ONTOLOGY are available, for the users of IMGT and the biologists in general, in natural language in IMGT Scientific chart (http://imgt.cines.fr), and have been formalized for programming purpose in IMGT-ML (XML Schema). IMGT-ONTOLOGY is being implemented in Protégé and OBO-Edit to facilitate the export in formats such as OWL, and to link, whenever possible, the concepts of IMGT-ONTOLOGY to those of other ontologies in biology such as the Gene Ontology (GO) [34], and in immunology, such as the Immunome Epitope database and Analysis Resource (IEDB) [35] and other Open Biomedical Ontologies (OBO) (http://obo.sourceforge.net).

The concepts of IMGT-ONTOLOGY are currently used for the exchange and the sharing of knowledge in very diverse fields of research at the molecular level: (i) fundamental and medical research (repertoire analysis of the IG antibody sites and of the TR recognition sites in normal and pathological situations such as autoimmune diseases, infectious diseases, AIDS, leukaemias, lymphomas, myelomas), (ii) veterinary research (IG and TR repertoires in farm and wild life species), (iii) genome diversity and genome evolution studies of the adaptive immune responses, (iv) structural evolution of the IgSF and MhcSF proteins, (v) biotechnology related to antibody engineering (scFv, phage displays, combinatorial libraries, chimaeric, humanized and human antibodies), (vi) diagnostics (clonalities, detection and follow-up of residual diseases) and (vii) therapeutical approaches (grafts, immunotherapy, vaccinology). IMGT-ONTOLOGY represents a key component in the elaboration and setting up of standards of the European ImmunoGrid project (http://www.immunogrid.org/) whose aim is to define the essential concepts for modelling of the immune system.

Acknowledgements

We are grateful to Gérard Lefranc for helpful discussion and to the IMGT® team for its constant motivation and its expertise. IMGT® was funded by the Centre National de la Recherche Scientifique (CNRS), the BIOMED1 (BIOCT 930038), Biotechnology BIOTECH2 (BIO4CT960037) and 5th PCRDT Quality of Life and Management of LivingResources (QLG2-2000-01287) programmes of the European Union. IMGT received subventions from Association pour la Recherche sur le Cancer (ARC), Génopole-Montpellier-Languedoc-Roussillon and the Réseau National des Génopoles (RNG). IMGT has been a National Bioinformatics RIO Platform since 2001 (CNRS, INSERM, CEA, INRA). IMGT is currently supported by the CNRS, the Ministère de l'Education Nationale de l'Enseignement Supérieur et de la Recherche (MENESR), Université Montpellier 2 Plan Pluri-Formation, Région Languedoc-Roussillon, BIOSTIC-LR2004, ACI-IMPBIO (IMP82-2004), GIS-AGENAE, Agence Nationale de la Recherche ANR (BIOSYS06_135457) and the European Union ImmunoGrid project (IST-2004-0280069).

References

[1] M.-P. Lefranc, G. Lefranc, The Immunoglobulin FactsBook, Academic Press, London UK, 2001, 458 pp.
[2] M.-P. Lefranc, G. Lefranc, The T Cell Receptor FactsBook, Academic Press, London UK, 2001, 398 pp.
[3] H. Sakano, K. Huppi, G. Heinrich, S. Tonegawa, Sequences at the somatic recombination sites of immunoglobulin light-chain genes, Nature 280 (1979) 288–294.
[4] M.-P. Lefranc, V. Giudicelli, Q. Kaas, E. Duprat, J. Jabado-Michaloud, D. Scaviner, C. Ginestoux, O. Clément, D. Chaume, G. Lefranc, IMGT, the international ImMunoGeneTics information system®, Nucleic Acids Res. 33 (2005) D593–D597.
[5] M.-P. Lefranc, O. Clément, Q. Kaas, E. Duprat, P. Chastellan, I. Coelho, K. Combres, C. Ginestoux, V. Giudicelli, D. Chaume, G. Lefranc, IMGT-Choreography for immunogenetics and immunoinformatics, In Silico Biol. 5 (2005) 45–60.
[6] V. Giudicelli, M.-P. Lefranc, Ontology for Immunogenetics: IMGT-ONTOLOGY, Bioinformatics 15 (1999) 1047–1054.
[7] M.-P. Lefranc, V. Giudicelli, C. Ginestoux, N. Bosc, G. Folch, D. Guiraudou, J. Jabado-Michaloud, S. Magris, D. Scaviner, V. Thouvenin, K. Combres, D. Girod, S. Jeanjean, C. Protat, M. Yousfi Monod, E. Duprat, Q. Kaas, C. Pommié, D. Chaume, G. Lefranc, IMGT-ONTOLOGY for Immunogenetics and Immunoinformatics http://imgt.cines.fr, In Silico Biol. 4 (2004) 17–29.
[8] V. Giudicelli, D. Chaume, J. Jabado-Michaloud, M.-P. Lefranc, Immunogenetics sequence annotation: the strategy of IMGT based on IMGT-ONTOLOGY, Stud. Health Technol. Inform. 116 (2005) 3–8.
[9] Q. Kaas, M.-P. Lefranc, T cell receptor/peptide/MHC molecular characterization and standardized pMHC contact sites in IMGT/3Dstructure-DB, In Silico. Biol. 5 (2005) 505–528.
[10] V. Giudicelli, C. Ginestoux, G. Folch, J. Jabado-Michaloud, D. Chaume, M.-P. Lefranc, IMGT/LIGM-DB, the IMGT® comprehensive database of immunoglobulin and T cell receptor nucleotide sequences, Nucleic Acids Res. 34 (2006) D781–D784.
[11] H.M. Wain, E.A. Bruford, R.C. Lovering, M.J. Lush, M.W. Wright, S. Povey, Guidelines for human gene nomenclature, Genomics 79 (2002) 464–470.
[12] M.-P. Lefranc, Unique database numbering system for immunogenetic analysis, Immunol. Today 18 (1997) 509.
[13] M.-P. Lefranc, The IMGT unique numbering for Immunoglobulins, T cell receptors and Ig-like domains, Immunologist 7 (1999) 132–136.
[14] M.-P. Lefranc, C. Pommié, M. Ruiz, V. Giudicelli, E. Foulquier, L. Truong, V. Thouvenin-Contet, G. Lefranc, IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains, Dev. Comp. Immunol. 27 (2003) 55–77.
[15] M.-P. Lefranc, C. Pommié, Q. Kaas, E. Duprat, N. Bosc, D. Guiraudou, C. Jean, M. Ruiz, I. Da Piedade, M. Rouard, E. Foulquier, V. Thouvenin, G. Lefranc, IMGT unique numbering for immunoglobulin and T cell receptor constant domains and Ig superfamily C-like domains, Dev. Comp. Immunol. 29 (2005) 185–203.
[16] M.-P. Lefranc, E. Duprat, Q. Kaas, M. Tranne, A. Thiriot, G. Lefranc, IMGT unique numbering for MHC groove G-DOMAIN and MHC superfamily (MhcSF) G-LIKE-DOMAIN, Dev. Comp. Immunol. 29 (2005) 917–938.
[17] T.R. Gruber, A translation approach to portable ontologies, Knowledge Acquisit. 5 (1993) 199–220.
[18] N. Guarino, P. Giaretta, Ontologies and knowledge bases: towards a terminological clarification, in: N. Mars (Ed.), Towards Very Large Knowledge Bases, IOS Press, Amsterdam, 1995, pp. 29–45.
[19] N. Guarino, Understanding, building and using ontologies, Int. J. Human-Comput. Stud. 46 (1997) 293–310.
[20] Noy N.F., McGuinness D.L., Ontology development 101: A guide to creating your first ontology, Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, March 2001.

[21] B. Smith, Ontology, in: L. Floridi (Ed.), Blackwell Guide to the Philosophy of Computing and Information, Blackwell, Oxford, 2003, pp. 155–166.
[22] B. Smith, W. Ceusters, B. Klagges, J. Köhler, A. Kumar, J. Lomax, C. Mungall, F. Neuhaus, A. Rector, C. Rosse, Relations in biomedical ontologies, Genome Biol. 6 (2005) R46.
[23] L.-N. Soldatova, A. Clare, A. Sparkes, R.D. King, An ontology for a Robot Scientist, Bioinformatics 22 (2006) e464–e471.
[24] D. Chaume, V. Giudicelli, M.-P. Lefranc, IMGT-ML a XML language for IMGT-ONTOLOGY and IMGT/LIGM-DB data, in: Proceedings of NETTAB 2001, Network Tools and Applications in Biology, Genoa, Italy, May 17–18, 2001, pp. 71–75.
[25] D. Chaume, K. Combres, V. Giudicelli, M.-P. Lefranc, Retrieving factual data and documents using IMGT-ML in the IMGT information system®, in: Proceedings of NETTAB 2005, Workflows management: new abilities for the biological information overflow, Naples, Italy, Oct 5–7, 2005, pp. 47–51.
[26] N.F. Noy, M. Crubezy, R.W. Fergerson, H. Knublauch, S.W. Tu, J. Vendetti, et al., Protege-2000: an open-source ontology-development and knowledge-acquisition environment, AMIA Annu. Symp. Proc. (2003) 953.
[27] J.T. Eppig, C.J. Bult, J.A. Kadin, J.E. Richardson, J.A. Blake, the members of the Mouse Genome Database Group 2005, The Mouse Genome Database (MGD): from genes to mice—a community resource for mouse biology, Nucleic Acids Res. 33 (2005) D471–D475.
[28] Q. Kaas, M. Ruiz, M.-P. Lefranc, IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data, Nucleic Acids Res. 32 (2004) D208–D210.
[29] K. Eilbeck, S.E. Lewis, Sequence Ontology annotation guide, Comp. Funct. Genomics 5 (2004) 642–647.
[30] D. Maglott, J. Ostell, K.D. Pruitt, T. Tatusova, Entrez Gene: gene-centered information at NCBI, Nucleic Acids Res. 35 (2007) D26–D31.
[31] V. Giudicelli, D. Chaume, M.-P. Lefranc, IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes, Nucleic Acids Res. 33 (2005) D256–D261.
[32] E. Duprat, Q. Kaas, V. Garelle, G. Lefranc, M.-P. Lefranc, IMGT standardization for alleles and mutations of the V-LIKE-DOMAINs and C-LIKE-DOMAINs of the immunoglobulin superfamily, in: S.G. Pandalai (Ed.), Recent Research Developments in Human Genetics, 2, Research Signpost, Trivandrum, India, 2004, pp. 111–136.
[33] Q. Kaas, M.-P. Lefranc, IMGT Colliers de Perles: standardized sequence-structure representations of the IgSF and MhcSF superfamily domains, Curr. Bioinformat. 2 (2007) 21–30.
[34] The Gene Ontology Consortium, The Gene Ontology (GO) project in 2006, Nucleic Acids Res. 34 (2006) D322–D326.
[35] B. Peters, J. Sidney, P. Bourne, H.H. Bui, S. Buus, G. Doh, W. Fleri, M. Kronenberg, R. Kubo, O. Lund, D. Nemazee, J.V. Ponomarenko, M. Sathiamurthy, S. Schoenberger, S. Stewart, P. Surko, S. Way, S. Wilson, A. Sette, The immune epitope database and analysis resource: from vision to blueprint, PLoS Biol. 3 (2005) e91.

PUBLICATION 3

Nucleic Acids Research Advance Access published October 31, 2008

Nucleic Acids Research, 2008, 1–7
doi:10.1093/nar/gkn838

IMGT®, the international ImMunoGeneTics information system®

Marie-Paule Lefranc*, Véronique Giudicelli, Chantal Ginestoux, Joumana Jabado-Michaloud, Géraldine Folch, Fatena Bellahcene, Yan Wu, Elodie Gemrot, Xavier Brochet, Jérôme Lane, Laetitia Regnier, François Ehrenmann, Gérard Lefranc and Patrice Duroux

IMGT®, the international ImMunoGeneTics information system®, Université Montpellier 2, Laboratoire d'ImmunoGénétique Moléculaire LIGM, UPR CNRS 1142, Institut de Génétique Humaine IGH, 141 rue de la Cardonille, 34396 Montpellier Cedex 5, France

Received September 13, 2008; Accepted October 14, 2008

ABSTRACT

IMGT®, the international ImMunoGeneTics information system® (http://www.imgt.org), was created in 1989 by Marie-Paule Lefranc, Laboratoire d'ImmunoGénétique Moléculaire LIGM (Université Montpellier 2 and CNRS) at Montpellier, France, in order to standardize and manage the complexity of immunogenetics data. The building of a unique ontology, IMGT-ONTOLOGY, has made IMGT® the global reference in immunogenetics and immunoinformatics. IMGT® is a high-quality integrated knowledge resource specialized in the immunoglobulins or antibodies, T cell receptors, major histocompatibility complex, of human and other vertebrate species, proteins of the IgSF and MhcSF, and related proteins of the immune systems of any species. IMGT® provides a common access to standardized data from genome, proteome, genetics and 3D structures. IMGT® consists of five databases (IMGT/LIGM-DB, IMGT/GENE-DB, IMGT/3Dstructure-DB, etc.), fifteen interactive online tools for sequence, genome and 3D structure analysis, and more than 10 000 HTML pages of synthesis and knowledge. IMGT® is used in medical research (autoimmune diseases, infectious diseases, AIDS, leukemias, lymphomas and myelomas), veterinary research, biotechnology related to antibody engineering (phage displays, combinatorial libraries, chimeric, humanized and human antibodies), diagnostics (clonalities, detection and follow-up of residual diseases) and therapeutical approaches (graft, immunotherapy, vaccinology). IMGT is freely available at http://www.imgt.org.

INTRODUCTION

The number of genomics, genetics, 3D and functional data published in the immunogenetics field is growing exponentially and involves fundamental, clinical, veterinary, and pharmaceutical research. The number of potential protein forms of the antigen receptors, immunoglobulins (IG) and T cell receptors (TR) is almost unlimited. The potential repertoire of each individual is estimated to comprise about 10^{12} different IG (or antibodies) and TR, and the limiting factor is only the number of B and T cells that an organism is genetically programmed to produce. This huge diversity is inherent to the particularly complex and unique molecular synthesis and genetics of the antigen receptor chains. This includes biological mechanisms such as DNA molecular rearrangements in multiple loci (three for IG and four for TR in humans) located on different chromosomes (four in humans), nucleotide deletions and insertions at the rearrangement junctions (or N-diversity), and somatic hypermutations in the IG loci (1,2).

IMGT®, the international ImMunoGeneTics information system® (http://www.imgt.org) (3), was created in 1989 by Marie-Paule Lefranc, Laboratoire d'ImmunoGénétique Moléculaire LIGM (Université Montpellier 2 and CNRS) at Montpellier, France, in order to standardize and manage the complexity of immunogenetics data. IMGT® has reached that goal through the building of a unique ontology, IMGT-ONTOLOGY (4), the first ontology in immunogenetics and immunoinformatics. IMGT-ONTOLOGY has allowed the setting up of the official nomenclature of the IG and TR genes and alleles (5,6), the definition of IMGT standardized labels, and the IMGT unique numbering that bridges the gap between sequences and 3D structures for the variable (V) and constant (C) domains of the IG and TR (7–10) and for the groove (G) domains of the major histocompatibility

*To whom correspondence should be addressed. Tel: +33 4 99 61 99 65; Fax: +33 4 99 61 99 01; Email: marie-paule.lefranc@igh.cnrs.fr

complex (MHC) (11). IMGT® is recognized as the global reference that provides the standards in immunogenetics and immunoinformatics. IMGT® is a high-quality integrated knowledge resource, specialized in the IG, TR, MHC of human and other vertebrates, the proteins that belong to the immunoglobulin superfamily (IgSF) and to the MHC superfamily (MhcSF), and the related proteins of the immune systems (RPI) of any species. IMGT® provides a common access to standardized data from genome, proteome, genetics and 3D structures.

The IMGT® information system consists of databases, tools and Web resources (3). IMGT® databases include one genome database, several sequence databases and one 3D structure database. Fifteen IMGT® interactive online tools are provided for genome, sequence and 3D structure analysis. IMGT® Web resources comprise more than 10 000 HTML pages of synthesis and knowledge (IMGT Scientific chart, IMGT Repertoire, The IMGT Medical page, The IMGT Veterinary page, The IMGT Biotechnology page, IMGT Education, IMGT Lexique, IMGT Aide-Mémoire, Tutorials, IMGT Index), external links (IMGT Bloc-notes, The IMGT Immunoinformatics page) and IMGT other accesses (SRS, MRS). Despite the heterogeneity of these different components, all data in IMGT® are expertly annotated. The accuracy, the consistency and the integration of the IMGT® data, as well as the coherence between the different IMGT® components (databases, tools and Web resources) are based on the IMGT-ONTOLOGY axioms and concepts (4,12).

IMGT-ONTOLOGY

Formal IMGT-ONTOLOGY

The Formal IMGT-ONTOLOGY, also designated as IMGT Kaleidoscope (12), comprises seven axioms: IDENTIFICATION, DESCRIPTION, CLASSIFICATION, NUMEROTATION, ORIENTATION, LOCALIZATION and OBTENTION that postulate that objects, processes and relations have to be identified, described, classified, numerotated, localized, orientated, and that the way they are obtained has to be determined. IMGT-ONTOLOGY concepts derived from these axioms are available, for the biologists and IMGT® users, in the IMGT Scientific chart, and have been formalized, for the computing scientists, in IMGT-ML which is an XML Schema (http://www.w3.org/TR/xmlschema-0/). In order to formalize the semantic relations between concepts and instances that are essential for high-quality data processing and coherence control, IMGT-ONTOLOGY is currently designed with Protégé (13) and OBO-Edit (http://oboedit.org/).

IMGT Scientific chart

The IMGT Scientific chart is constituted by controlled vocabulary and annotation rules for data and knowledge management of the IG, TR, MHC, IgSF, MhcSF and RPI. All IMGT® data are expertly annotated according to the IMGT Scientific chart rules.

Keywords and labels. IMGT standardized keywords (concepts of identification) are assigned to all entries in the IMGT® databases. More than 500 IMGT standardized labels (concepts of description) were necessary to describe all structural and functional subregions that compose IG and TR (221 labels for sequences and 285 for 3D structures). Interestingly, 64 IMGT specific labels defined for nucleotide sequences have been entered in the newly created Sequence Ontology (SO) (14).

Gene and allele nomenclature. All the human IMGT standardized gene names (5,6) (concepts of classification) were approved by the Human Genome Organisation (HUGO) Nomenclature Committee (HGNC) in 1999 (15), and entered in IMGT/GENE-DB (16), and in Entrez Gene NCBI (17), and more recently on the Ensembl server (18) at the European Bioinformatics Institute (EBI) in 2006, and in the Vega (19) database at the Wellcome Trust Sanger Institute in 2008. All the mouse IMGT® gene and allele names and the corresponding IMGT reference sequences were provided to HGNC and to the Mouse Genome Informatics Mouse Genome Database (20) in July 2002 and were presented by IMGT® at the 19th International Mouse Genome Conference, IMGC 2005, in Strasbourg, France, and entered in IMGT/GENE-DB. IMGT reference sequences have been defined for each allele of each gene based on one or, whenever possible, several of the following criteria: germline sequence, first sequence published, longest sequence, mapped sequence.

IMGT unique numbering. The IMGT unique numbering (concepts of numerotation) (7–11) is, with its 2D graphical representation or IMGT Collier de Perles (21,22), the flagship of IMGT® that allows to bridge the gap between sequences, genes and 3D structures in the IMGT® databases, tools and Web resources (23). Structural and functional domains of the IG and TR chains comprise the V-DOMAIN (9-strand β-sandwich) which corresponds to the V-J-REGION or V-D-J-REGION and is encoded by two or three genes (1,2), and the constant domain or C-DOMAIN (7-strand β-sandwich). The IMGT unique numbering initially defined for the IG and TR domains has been extended to the V-LIKE-DOMAIN and C-LIKE-DOMAIN of IgSF proteins other than IG and TR (9,10,22). The IMGT unique numbering for the MHC G-DOMAIN (four β-strand and one α-helix) has been extended to the G-LIKE-DOMAIN of MhcSF proteins other than MHC (11,22).

IMGT Choreography

In order to extract knowledge from IMGT® standardized immunogenetics data, three main IMGT® biological approaches have been developed: genomic, genetic and structural approaches (Figure 1). The IMGT® genomic approach is gene-centered and mainly orientated towards the study of the genes within their loci and on the chromosomes. The IMGT® genetic approach refers to the study of the genes in relation with their sequence polymorphisms and mutations, their expression, their specificity and their evolution. The IMGT® structural

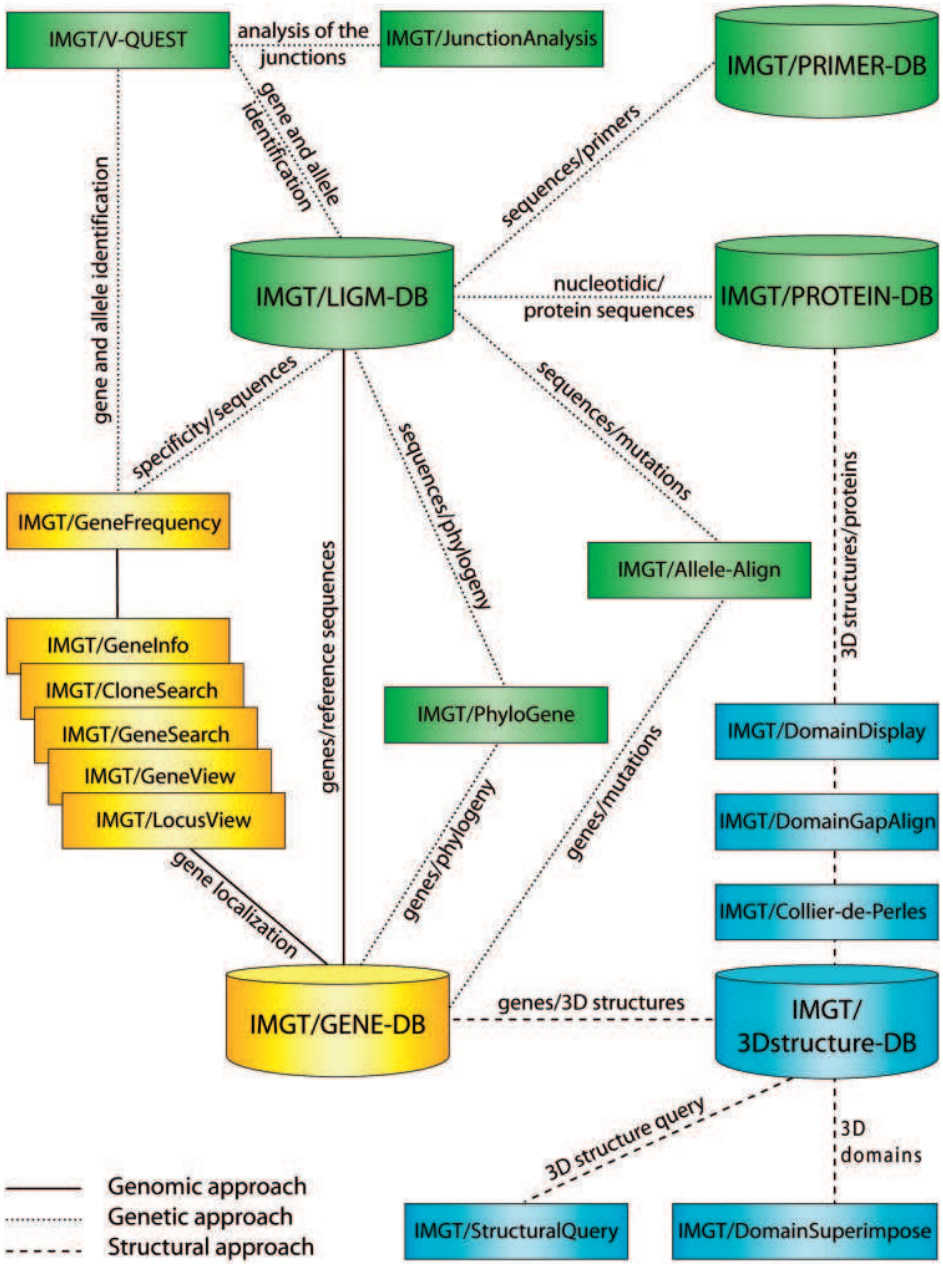


Figure 1. IMGT®, the international ImMunoGeneTics information system® (http://www.imgt.org). Genomic, genetic and structural components (databases and tools) are in yellow, green and blue, respectively. The IMGT Repertoire and other Web resources are not shown. Interactions in the genomic, genetic and structural approaches are represented with continuous, dotted and broken lines, respectively.

approach refers to the study of the 2D and 3D structures of the IG, TR, MHC, IgSF, MhcSF and RPI, and to the antigen- or ligand-binding characteristics in relationship with the protein functions, polymorphisms and evolution. For each approach, IMGT® provides databases, tools and Web resources (Figure 1 and Table 1). IMGT-Choreography (33), based on the Web service architecture paradigm, has been developed with the goal to enable significant biological and clinical requests involving every part of the IMGT® information system.

IMGT® DATABASES

Gene database

IMGT/GENE-DB (16) is the comprehensive IMGT® genome database. IMGT/GENE-DB is the official repository of all the IG and TR genes and alleles approved by the World Health Organization (WHO)/International Union of Immunological Societies (IUIS) Nomenclature Subcommittee for IG and TR (34,35). In September 2008, IMGT/GENE-DB contained 1911 IG and TR genes from human, mouse and rat and 2909 alleles. Reciprocal links

Table 1. IMGT® databases, tools and Web resources for genomic, genetic and structural approaches

Approaches	Databases	Tools	Web resources[a]
Genomic	IMGT/GENE-DB (16)	IMGT/GeneView IMGT/LocusView IMGT/CloneSearch IMGT/GeneSearch IMGT/GeneInfo (28) IMGT/GeneFrequency	IMGT Repertoire 'Locus and genes' section: –chromosomal localizations (1,2) –locus representations (1,2) –locus description –gene tables, etc. –potential germline repertoires –lists of genes –correspondence between nomenclatures (1,2)
Genetic	IMGT/LIGM-DB (24) IMGT/PRIMER-DB (25) IMGT/MHC-DB (26)	IMGT/V-QUEST (29) IMGT/JunctionAnalysis (30) IMGT/Allele-Align IMGT/PhyloGene (31) IMGT/DomainDisplay	IMGT Repertoire 'Proteins and alleles' section: –alignments of alleles –protein displays –tables of alleles etc.
Structural	IMGT/3Dstructure-DB (27)	IMGT/DomainGapAlign IMGT/Collier-de-Perles IMGT/DomainSuperimpose IMGT/StructuralQuery (27)	IMGT Repertoire '2D and 3D structures' section: –IMGT Colliers de Perles (2D representations on one layer or two layers) (21,22) –IMGT classes for amino acid characteristics (32) –IMGT Colliers de Perles reference profiles (32) –3D representations

[a]Only Web resources examples from the IMGT Repertoire section are shown.

exist between IMGT/GENE-DB and the HGNC database (36) and Entrez Gene (17). IMGT-GENE-DB allows a query per gene and allele name. IMGT/GENE-DB interacts dynamically with IMGT/LIGM-DB (24) to download and display human, mouse and rat gene-related sequence data. This is the first example of an interaction between IMGT® databases using the concepts of classification.

Sequence databases

IMGT/LIGM-DB. IMGT/LIGM-DB (24) is the comprehensive IMGT® database of IG and TR nucleotide sequences from human and other vertebrate species, with translation for fully annotated sequences, created in 1989 by LIGM, Montpellier, France, on the Web since July 1995. IMGT/LIGM-DB is the first and the largest IMGT® database. In September 2008, IMGT/LIGM-DB contained 126 667 nucleotide sequences of IG and TR from 223 vertebrate species. The unique source of data for IMGT/LIGM-DB is EMBL-Bank (18) which shares data with the other two generalist databases GenBank (37) and DDBJ (38). IMGT/LIGM-DB sequence data are identified by the EMBL/GenBank/DDBJ accession number. Based on expert analysis, specific detailed annotations are added to IMGT flat files. The Web interface allows searches according to immunogenetic specific criteria and is easy to use without any knowledge in a computing language. Selection is displayed at the top of the resulting sequences page, so the users can check their own queries. Users have the possibility to modify their request or consult the results with a choice of nine possibilities. The IMGT/LIGM-DB annotations (gene and allele name assignment, labels) allow data retrieval not only from IMGT/LIGM-DB, but also from other IMGT® databases. For example, the IMGT/GENE-DB entries provide the IMGT/LIGM-DB accession numbers of the IG and TR cDNA sequences that contain a given V, D, J or C gene. The automatic annotation of rearranged human and mouse cDNA sequences in IMGT/LIGM-DB is performed by IMGT/Automat (39), an internal Java tool that implements IMGT/V-QUEST (29) and IMGT/JunctionAnalysis (30). IMGT/LIGM-DB data are also distributed by anonymous FTP servers at CINES (ftp://ftp.cines.fr/IMGT/) and EBI (ftp://ftp.ebi.ac.uk/pub/databases/imgt/) and from several Sequence Retrieval System (SRS) sites. IMGT/LIGM-DB can be searched by BLAST or FASTA on different servers (EBI, IGH, Institut Pasteur Paris).

Other IMGT sequence databases. IMGT/PRIMER-DB (25) is the IMGT® oligonucleotide database on the Web since February 2002. In September 2008, IMGT/PRIMER-DB contained 1864 entries. The database manages standardized information on oligonucleotides (or Primers) and combinations of primers (Sets and Couples) for IG and TR. These primers are useful for combinatorial library constructions, scFv, phage display or microarray technologies. IMGT/PROTEIN-DB, in developement, will contain the translations of the IMGT/LIGM-DB and IMGT/GENE-DB sequences. IMGT/MHC-DB hosted at EBI comprises IMGT/HLA for human MHC (or HLA) and IMGT/MHC-NHP, for MHC of non-human primates (26).

Structure database

IMGT/3Dstructure-DB is the IMGT® 3D structure database, created by LIGM, on the Web since November 2001 (27). IMGT/3Dstructure-DB comprises IG, TR, MHC, IgSF, MhcSF and RPI with known 3D structures. In September 2008, IMGT/3Dstructure-DB contained 1461 atomic coordinate files. These coordinate files extracted from the Protein Data Bank (PDB) (http://www.rcsb.org/pdb/) (40) are renumbered according to the standardized IMGT unique numbering (9–11). The IMGT/3Dstructure-DB cards provide chain details with IMGT

annotations (receptor, chain and domain description with IMGT labels, assignment of IMGT gene and allele names, domain delimitations and amino acid positions according to the IMGT unique numbering, and IMGT Colliers de Perles on one layer and two layers), contact analysis, downloadable renumbered IMGT/3Dstructure-DB flat files, visualization tools (Jmol and QuickPDB), and external links. IMGT Residue@Position cards provide detailed information on the inter- and intra-domain contacts at each residue position, based on the IMGT unique numbering. The contacts are described per domain (intra- and inter-domain contacts) and annotated in terms of IMGT® labels (chain and domain), positions (IMGT unique numbering), backbone or side-chain implication.

IMGT® TOOLS

Gene tools

The IMGT® gene tools (genomic approach) manage the locus organization and gene location and provide the display of physical maps for the human and mouse IG, TR and MHC loci. They allow to view genes in a locus (IMGT/GeneView, IMGT/LocusView), to search for clones (IMGT/CloneSearch), or to search for genes in a locus (IMGT/GeneSearch, IMGT/GeneInfo) based on IMGT® gene names, functionality or localization on the chromosome. IMGT/GeneFrequency provides a graphical representation of the numbers of cDNA and gDNA IMGT/LIGM-DB sequences containing rearranged IG and TR genes.

Sequence tools

The IMGT® sequence analysis tools (genetic approach) comprise IMGT/V-QUEST (29) for the identification of the V, D and J genes and of their mutations, IMGT/JunctionAnalysis (30) for the analysis of the V-J and V-D-J junctions that confer the antigen receptor specificity, IMGT/Allele-Align for the detection of polymorphisms, IMGT/Phylogene (31) for gene evolution analyses, and IMGT/DomainDisplay for the display of amino acid sequences from the IMGT domain directory. IMGT/V-QUEST (V-QUEry and STandardization) (29) is an integrated software for IG and TR. This tool, which is easy to use, analyses an input of up to fifty IG or TR germline or rearranged variable nucleotide sequences. IMGT/V-QUEST results comprise, for rearranged sequences, the identification of the V, D and J genes and alleles, nucleotide alignments by comparison with the IMGT reference directory, the delimitations of the framework regions (FR-IMGT) and complementarity determining regions (CDR-IMGT) based on the IMGT unique numbering, the protein translation of the input sequences, the result of IMGT/JunctionAnalysis, the description of the mutations and amino acid changes of the V-REGION and the IMGT Collier de Perles representation of the V-DOMAIN.

Structure tools

The IMGT® structure tools bridge the gap between sequences and 3D structures: IMGT/DomainGapAlign analyses amino acid sequences per domain, IMGT/Collier-de-Perles allows to make your own IMGT Collier de Perles, and IMGT/DomainSuperimpose allows to superimpose two domain 3D structures from IMGT/3Dstructure-DB. IMGT/StructuralQuery (27) allows to retrieve the IMGT/3Dstructure-DB entries containing a V-DOMAIN, based on specific structural characteristics of the intramolecular interactions: phi and psi angles, accessible surface area, amino acid type, distance in angstrom between amino acids, and CDR-IMGT lengths.

IMGT® WEB RESOURCES

IMGT Repertoire

The IMGT® Web resources for genomic, genetic and structural approaches are compiled in the sections of the IMGT Repertoire and provide a synthetic view of data managed in the databases and tools.

Genomics Web resources. The IMGT® genomics resources are compiled in the 'Locus and genes' section which includes 'Chromosomal localizations', 'Locus representations', 'Locus description', 'Gene exon/intron organization', 'Gene exon/intron splicing sites', 'Gene tables', 'Potential germline repertoires', lists of IG and TR genes and links between IMGT®, HGNC, Entrez Gene and OMIM, and correspondence between nomenclatures (1,2). The IMGT Repertoire 'Probes and RFLP' section provides data on gene insertion/deletion.

Genetics Web resources. The IMGT® genetics resources are compiled in the 'Proteins and alleles' section which includes 'Alignments of alleles', 'Tables of alleles', 'Allotypes', 'Isotypes', 'Protein displays', etc.

Structural Web resources. The IMGT® structural resources are compiled in the '2D and 3D structures' section which includes IMGT Colliers de Perles (21,22), FR-IMGT and CDR-IMGT lengths, amino acid chemical characteristics profiles (32). To appropriately analyse the amino acid resemblances and differences between IG, TR, MHC and RPI chains, eleven IMGT classes were defined for the amino acid 'chemical characteristics' properties and used to set up IMGT Colliers de Perles references profiles. IMGT Colliers de Perles reference profiles allow to easily compare amino acid properties at each position whatever the domain, the chain, the receptor or the species. The visualization of 3D representations of IG and TR variable domains allows rapid correlation between protein sequences and 3D data.

Other Web resources

In addition to the IMGT Scientific chart and IMGT Repertoire, other major components of the IMGT® Web resources comprise The IMGT Medical page, The IMGT Veterinary page, The IMGT Biotechnology page,

IMGT Education, IMGT Lexique, IMGT Aide-Mémoire, Tutorials, IMGT Index, and external links (IMGT Blocnotes, The IMGT Immunoinformatics page, Interesting links) and IMGT other accesses (SRS,MRS).

CONCLUSION

Since July 1995, IMGT® has been available on the Web at the IMGT Home page http://www.imgt.org (Montpellier, France). IMGT® has an exceptional response with more than 150 000 requests a month. The information is of much value to clinicians and biological scientists in general. IMGT® databases, tools, and Web resources are extensively queried and used by scientists from both academic and industrial laboratories, who are equally distributed between the United States, Europe and the remaining world. IMGT® is used in very diverse domains: (i) fundamental and medical research (repertoire analysis of the IG antibody recognition sites and of the TR recognition sites in normal and pathological situations such as autoimmune diseases, infectious diseases, AIDS, leukemias, lymphomas and myelomas); (ii) veterinary research (IG and TR repertoires in farm and wild life species); (iii) genome diversity and genome evolution studies of the adaptive immune responses; (iv) structural evolution of the IgSF and MhcSF proteins; (v) biotechnology related to antibody engineering (single chain Fragment variable (scFv), phage displays, combinatorial libraries, chimeric, humanized and human antibodies); (vi) diagnostics (clonalities, detection and follow-up of residual diseases) and (vii) therapeutical approaches (grafts, immunotherapy and vaccinology). The creation of dynamic interactions between the IMGT® databases and tools, using Web services and IMGT-ML, and the design of IMGT-Choreography, represent novel and major developments of IMGT®, the international reference in immunogenetics and immunoinformatics.

CITING IMGT

Users are requested to cite this article and quote the IMGT home page URL, http://www.imgt.org.

ACKNOWLEDGEMENTS

We thank all the IMGT® users from academic and industrial laboratories and the clinicians and scientists from the European Research Initiative on CLL who help promoting standardization. IMGT® has received the National Bioinformatics Platform RIO label since the RIO creation in 2001 (CNRS, INSERM, CEA, INRA) and the National Bioinformatics Platform IBiSA label since the IBiSA creation in 2007. IMGT® is an Institutional Academic Member of the International Medical Informatics Association. IMGT® is a registered mark of the Centre National de la Recherche Scientifique.

FUNDING

The BIOMED1 (BIOCT930038), Biotechnology BIOTECH2 (BIO4CT960037) and 5th PCRDT Quality of Life and Management of Living Resources programmes (QLG2-2000-01287) programmes of the European Union (EU). IMGT® is currently supported by the CNRS, the Ministère de l'Enseignement Supérieur et de la Recherche (MESR) (Université Montpellier 2 Plan Pluri-Formation, Institut Universitaire de France), Réseau National des Génopoles, the Région Languedoc-Roussillon, the Agence Nationale de la recherche ANR (BIOSYS06_135457, FLAVORES), and the EU ImmunoGrid (IST-028069). Funding for open access charge: CNRS.

Conflict of interest statement. None declared.

REFERENCES

1. Lefranc,M.-P. and Lefranc,G. (2001) *The Immunoglobulin FactsBook*, Academic Press, London, UK, pp. 1–458.
2. Lefranc,M.-P. and Lefranc,G. (2001) *The T cell receptor FactsBook*, Academic Press, London, UK, pp. 1–398.
3. Lefranc,M.-P., Giudicelli,V., Kaas,Q., Duprat,E., Jabado-Michaloud,J., Scaviner,D., Ginestoux,C., Clément,O., Chaume,D. and Lefranc,G. (2005) IMGT, the international ImMunoGeneTics information system. *Nucleic Acids Res.*, **33**, D593–D597.
4. Giudicelli,V. and Lefranc,M.-P. (1999) Ontology for Immunogenetics: the IMGT-ONTOLOGY. *Bioinformatics*, **12**, 1047–1054.
5. Lefranc,M.-P. (2000) Nomenclature of the human immunoglobulin genes. In Coligan,J.E., Bierer,B.E., Margulies,D.E., Shevach,E.M. and Strober,W. (eds), *Current Protocols in Immunology*, John Wiley & Sons, Inc., NJ, Hoboken, pp. A.1P.1–A.1P.37.
6. Lefranc,M.-P. (2000) Nomenclature of the human T cell receptor genes. In Coligan,J.E., Bierer,B.E., Margulies,D.E., Shevach,E.M. and Strober,W. (eds), *Current Protocols in Immunology*, John Wiley & Sons, Inc., NJ, Hoboken, pp. A.1O.1–A.1O.23.
7. Lefranc,M.-P. (1997) Unique database numbering system for immunogenetic analysis. *Immunol. Today*, **18**, 509.
8. Lefranc,M.-P. (1999) The IMGT unique numbering for Immunoglobulins, T cell receptors and Ig-like domains. *The Immunologist*, **7**, 132–136.
9. Lefranc,M.-P., Pommié,C., Ruiz,M., Giudicelli,V., Foulquier,E., Truong,L., Thouvenin-Contet,V. and Lefranc,G. (2003) IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Dev. Comp. Immunol.*, **27**, 55–77.
10. Lefranc,M.-P., Pommié,C., Kaas,Q., Duprat,E., Bosc,N., Guiraudou,D., Jean,C., Ruiz,M., Da Piedade,I., Rouard,M. *et al.* (2005) IMGT unique numbering for immunoglobulin and T cell receptor constant domains and Ig superfamily C-like domains. *Dev. Comp. Immunol.*, **29**, 185–203.
11. Lefranc,M.-P., Duprat,E., Kaas,Q., Tranne,M., Thiriot,A. and Lefranc,G. (2005) IMGT unique numbering for MHC groove G-DOMAIN and MHC superfamily (MhcSF) G-LIKE-DOMAIN. *Dev. Comp. Immunol.*, **29**, 917–938.
12. Duroux,P., Kaas,Q., Brochet,X., Lane,J., Ginestoux,C., Lefranc,M.-P. and Giudicelli,V. (2008) IMGT-Kaleidoscope, the Formal IMGT-ONTOLOGY paradigm. *Biochimie*, **90**, 570–583.
13. Noy,N.F., Crubezy,M., Fergerson,R.W., Knublauch,H., Tu,S.W., Vendetti,J. and Musen,M.A. (2003) Protege-2000: an open-source ontology-development and knowledge-acquisition environment. *AMIA Annu. Symp. Proc.*, **2003**, 953.
14. Eilbeck,K and Lewis,S.E (2004) Sequence Ontology Annotation Guide. *Com. Funct. Genomics*, **5**, 642–647.
15. Wain,H.M., Bruford,E.A., Lovering,R.C., Lush,M.J., Wright,M.W. and Povey,S. (2002) Guidelines for human gene nomenclature. *Genomics*, **79**, 464–470.

16. Giudicelli,V., Chaume,D. and Lefranc,M.-P. (2005) IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucleic Acids Res*., **33**, D256–D261.
17. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2007) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res*., **35**, D26–D31.
18. Cochrane,G., Akhtar,R., Aldebert,P., Althorpe,N., Baldwin,A., Bates,K., Bhattacharyya,S., Bonfield,J., Bower,L., Browne,P. *et al.* (2008) Priorities for nucleotide trace, sequence and annotation data capture at the Ensembl Trace Archive and the EMBL Nucleotide Sequence Database. *Nucleic Acids Res*., **36**, D5–D12.
19. Wilming,L.G., Gilbert,J.G.R., Howe,K., Trevanion,S., Hubbard,T. and Harrow,J.L. (2008) The vertebrate genome annotation (Vega) database. *Nucleic Acids Res*., **36**, D753–D760.
20. Bult,C.J., Eppig,J.T., Kadin,J.A., Richardson,J.E. and Blake,J.A. and the Mouse Genome Database Group (2008) The Mouse Genome Database (MGD): mouse biology and model systems. *Nucleic Acids Res*., **36**, D724–D728.
21. Ruiz,M. and Lefranc,M.-P. (2002) IMGT gene identification and Colliers de Perles of human immunoglobulins with known 3D structures. *Immunogenetics*, **53**, 857–883.
22. Kaas,Q., Ehrenmann,F. and Lefranc,M.-P. (2007) IG, TR, MHC, IgSf and MhcSF: what do we learn from the IMGT Colliers de Perles? Brief. *Funct. Genomic Proteomic*, **6**, 253–264.
23. Lefranc,M.-P., Giudicelli,V., Regnier,L. and Duroux,P. (2008) IMGT, a system and an ontology that bridge biological and computational spheres in bioinformatics. *Brief. Bioinform*., **9**, 263–275.
24. Giudicelli,V., Duroux,P., Ginestoux,C., Folch,G., Jabado-Michaloud,J., Chaume,D. and Lefranc,M.-P. (2006) IMGT/LIGM-DB, the IMGT® comprehensive database of immunoglobulin and T cell receptor nucleotide sequences. *Nucleic Acids Res*., **34**, D781–D784.
25. Folch,G., Bertrand,J., Lemaitre,M. and Lefranc,M.-P. (2004) IMGT/PRIMER-DB. The Molecular Biology Database Collection: 2004 update. *Nucleic Acids Res*., **32**, 3–22.
26. Robinson,J., Waller,M.J., Parham,P., de Groot,N., Bontrop,R., Kennedy,L.J., Stoehr,P. and Marsh,S.G. (2003) IMGT/HLA and IMGT/MHC sequence databases for the study of the major histocompatibility complex. *Nucleic Acids Res*., **31**, 311–314.
27. Kaas,Q., Ruiz,M. and Lefranc,M.-P. (2004) IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data. *Nucleic Acids Res*., **32**, D208–D210.
28. Baum,T.P., Hierle,V., Pasqual,N., Bellahcene,F., Chaume,D., Lefranc,M.-P., Jouvin-Marche,E., Marche,P.N. and Demongeot,J. (2006) IMGT/GeneInfo: T cell receptor gamma TRG and delta TRD genes in database give access to all TR potential V(D)J recombinations. *BMC Bioinformatics*, **7**, 224.
29. Brochet,X., Lefranc,M.-P. and Giudicelli,V. (2008) IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic Acids Res*, **36**, W503–W508.
30. Yousfi Monod,M., Giudicelli,V., Chaume,D. and Lefranc,M.-P. (2004) IMGT/JunctionAnalysis: the first tool for the analysis of the immunoglobulin and T cell receptor complex V-J and V-D-J JUNCTIONs. *Bioinformatics*, **20**, i379–i385.
31. Elemento,O. and Lefranc,M.-P. (2003) IMGT/PhyloGene: an on-line tool for comparative analysis of immunoglobulin and T cell receptor genes. *Dev. Comp. Immunol*., **27**, 763–779.
32. Pommié,C., Sabatier,S., Lefranc,G. and Lefranc,M.-P. (2004) IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties. *J. Mol. Recognit*., **17**, 17–32.
33. Lefranc,M.-P., Clément,O., Kaas,Q., Duprat,E., Chastellan,P., Coelho,I., Combres,K., Ginestoux,C., Giudicelli,V., Chaume,D. *et al.* (2004) IMGT-Choreography for Immunogenetics and Immunoinformatics. *In Silico Biol*., **5**, 45–60.
34. Lefranc,M.-P. (2008) WHO-IUIS Nomenclature Subcommittee for Immunoglobulins and T cell receptors report Immunoglobulins and T cell receptors report August 2007, 13th International Congress of Immunology, Rio de Janeiro, Brazil. *Dev. Comp. Immunol*., **32**, 461–463.
35. Lefranc,M.-P. (2007) WHO-IUIS Nomenclature Subcommittee for Immunoglobulins and T cell receptors report. *Immunogenetics*, **59**, 899–902.
36. Bruford,E.A., Lush,M.J., Wright,M.W., Sneddon,T.P., Povey,S. and Birney,E. (2008) The HGNC Database in 2008: a resource for the human genome. *Nucleic Acids Res*., **36**, D445–D448.
37. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2008) GenBank. *Nucleic Acids Res*., **36**, D25–D30.
38. Sugawara,H., Ogasawara,O., Okubo,K., Gojobori,T. and Tateno,Y. (2008) DDBJ with new system and face. *Nucleic Acids Res*., **36**, D22–D24.
39. Giudicelli,V., Chaume,D., Jabado-Michaloud,J. and Lefranc,M.-P. (2005) Immunogenetics sequence annotation: the strategy of IMGT based on IMGT-ONTOLOGY. *Stud. Health Technol. Inform*., **116**, 3–8.
40. Henrick,K., Feng,Z., Bluhm,W.F., Dimitropoulos,D., Doreleijers,J.F., Dutta,S., Flippen-Anderson,J.L., Ionides,J., Kamada,C., Krissinel,E. *et al.* (2008) Remediation of the protein data bank archive. *Nucleic Acids Res*., **36**, D426–D433.

RESUME en français

La Leucémie Lymphoïde Chronique (LLC) est la plus commune des leucémies chez l'adulte en occident. Bien que sa cause reste inconnue, des progrès importants ont été réalisés, grâce aux analyses des gènes des immunoglobulines (IG) exprimées par les cellules leucémiques. IMGT® est le système d'information international en immunogénétique spécialisé dans les IG et les récepteurs T (TR). Mes objectifs étaient de concevoir et d'intégrer au sein d'IMGT® une nouvelle composante dans le domaine médical pour corréler les données produites dans le cadre de la LLC aux caractéristiques génétiques des IG afin d'en extraire de nouvelles connaissances. J'ai conçu et développé un système d'information en accord avec les standards IMGT®, basé sur IMGT-ONTOLOGY, dans le cadre d'une collaboration avec des cliniciens spécialistes de la LLC: il est constitué de la base de données IMGT/CLL-DB dédiée à la gestion, d'une part de données associées aux patients, et d'autre part des résultats issus de l'analyse détaillée des IG exprimées à la surface des lymphocytes B. Dans ce but, j'ai intégré à la base de données le logiciel IMGT/V-QUEST, spécialisé dans l'analyse des séquences réarrangées des IG dont j'ai amélioré les performances. J'ai intégré de nouvelles fonctionnalités en accord avec les règles d'identification, de classification et de description d'IMGT®. IMGT/V-QUEST identifie les gènes V, D et J, décrit leurs principales caractéristiques structurales, propose une analyse de leur jonction et caractérise de façon standardisée les mutations. Le système d'information validé par le 'European Research Initiative on CLL' est utilisé en routine pour établir le pronostic de patients atteints de LLC, en tenant compte du taux de mutations des gènes IGHV. Ce système d'information constitué d'une base de données et d'un outil d'analyse peut être appliqué à l'étude du répertoire dans d'autres pathologies du système immunitaire (maladies autoimmunes, SIDA...).

TITRE en anglais

CONCEPTION AND INTEGRATION OF AN INFORMATION SYSTEM DEDICATED TO THE ANALYSES AND MANAGEMENT OF THE REARRANGED SEQUENCES OF ANTIGEN RECEPTORS AT IMGT: APPLICATION TO THE CHRONIC LYMPHOCYTIC LEUKEMIA

RESUME en anglais

The Chronic Lymphocytic Leukemia (CLL) is the most common leukemia for adult in occident. Although its etiology remains unknown, considerable progress have been achieved, thanks to the analyses of the genes of immunoglobulins (IG) expressed by the leukemic cells. IMGT® is the international information system in ImMunoGeneTics specialized in the IG and the T receptors (TR). My project was to conceive and to integrate for IMGT® a new component in the medical domain to correlate the data produced within the framework of the CLL, in the genetic characteristics of the IG to extract from it new knowledge. I conceived and developed an information system in agreement with the standards IMGT®, based on IMGT-ONTOLOGY, in collaboration with clinicians who are specialists of the CLL: it is constituted by the database IMGT/CLL-DB, dedicated to the management, on one hand of data associated to the patients and on the other hand of the results coming from the detailed analysis of the IG expressed on the surface of lymphocytes B. In order to achieve this, I integrated the IMGT/V-QUEST software into the database, this tool is dedicated to the analysis of rearranged sequences of IG and has been considerably improved. I integrated new features in agreement with the rules of identification, classification and description of IMGT®. IMGT/V-QUEST identifies the V, D and J genes, describes their main structural characteristics, proposes an analysis of their junction and characterizes in a standardized way the mutations. The information system validated by the 'European Research Initiative on CLL', is used there in routine to establish the prognostic and in the continuation of the clinical research on the CLL. This information system composed of a database and of a tool of analysis forms a coherent model for the integration of clinical and fundamental data which can be applied to other pathologies of the immune system (diseases autoimmune, AIDS...).

DISCIPLINE

Bioinformatique

MOTS-CLES

Immunogénétique, Immunoinformatique, Base de données, Immunoglobuline, récepteur T, Leucémie lymphoïde chronique, analyse de séquence.

Laboratoire d'ImmunoGénétique Moléculaire, Institut de Génétique Humaine, UPR CNRS 1142, 141 rue de la Cardonille, 34396 Cedex 5