جامعة محمد الخامس بالرباط
Université Mohammed V de Rabat

Polycopié du Cours de Génomique humaine & ImmunoInformatique
Master de Génomique et de Bioinformatique
Faculté des Sciences de Rabat

Pr. Khalid SADKI
Département des Sciences fondamentales
Faculté de Médecine Dentaire de Rabat

Années universitaires : 2022_2023

Plan du cours:

Définitions:

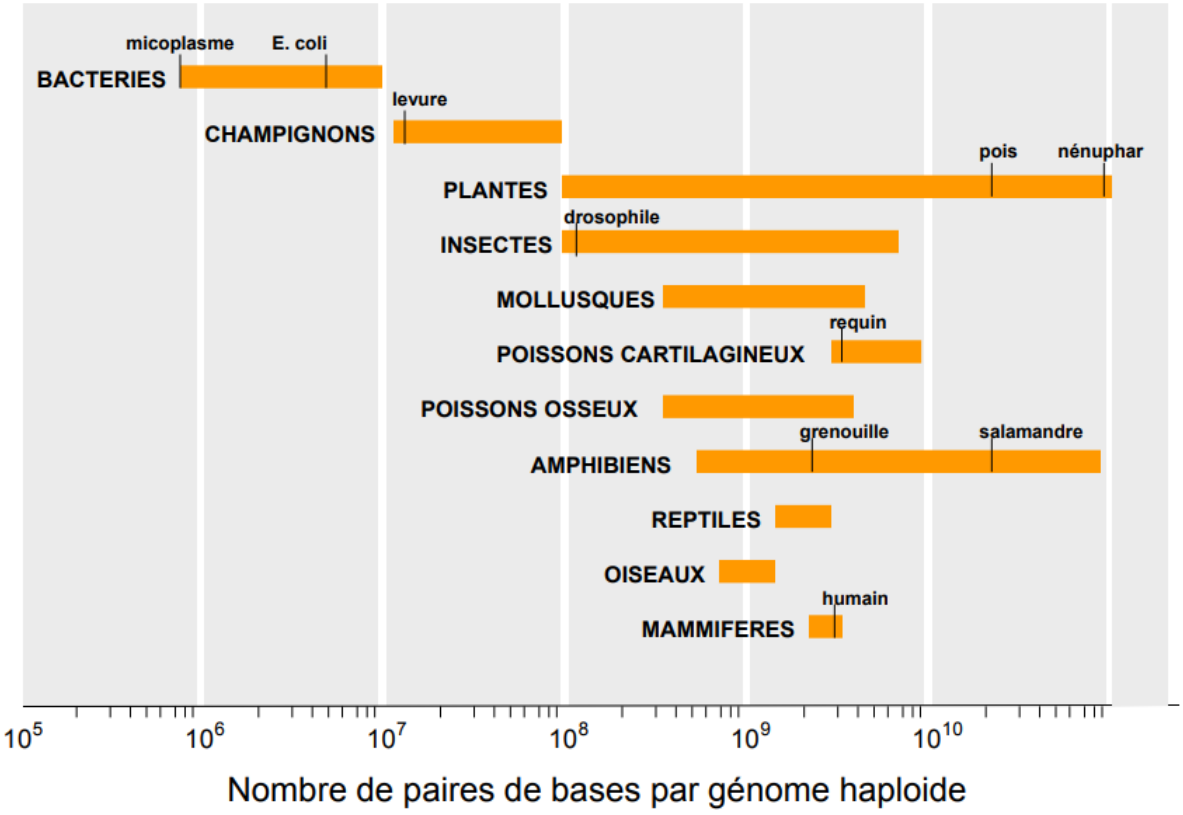
Le **Génome** est un Ensemble d'information héréditaire d'un organisme, présente en totalité dans chaque cellule :

- Le mot « génome » est la combinaison des mots « gène » et « chromosome ».
- Génome : Ensemble d'infor génétique d'un organisme contenu dans chacune de ses cellules sous la forme de chromosomes.
  - En Gnr, le support matériel du génome est : l'ADN,
  - sauf chez certains virus où il s'agit d'ARN.
- Gène : Fragment d'ADN contenant toutes les informations nécessaires pour produire un ARN ou, le plus souvent, une protéine.
  - Un gène correspond à une instruction à effectuer par la cellule.
- Chromosome : Élément constitutif du génome, composé d'une longue molécule d'ADN.

Le génome humain est constitué de 46 chromosomes **(23 paires).**

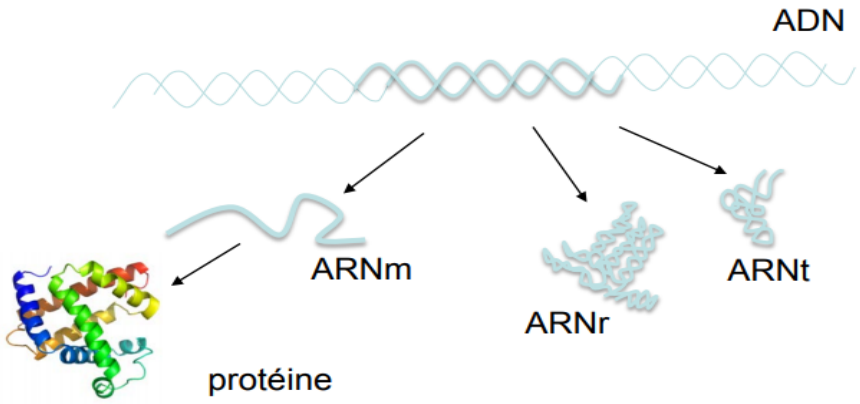
Taille des génomes :

- - Quelques dizaines de milliers de bases pour le génome d'un virus.
- - Quelques millions de bases pour une bactérie
- - 3 milliards de bases pour le génome humain
- -16 milliards de bases pour le génome du blé



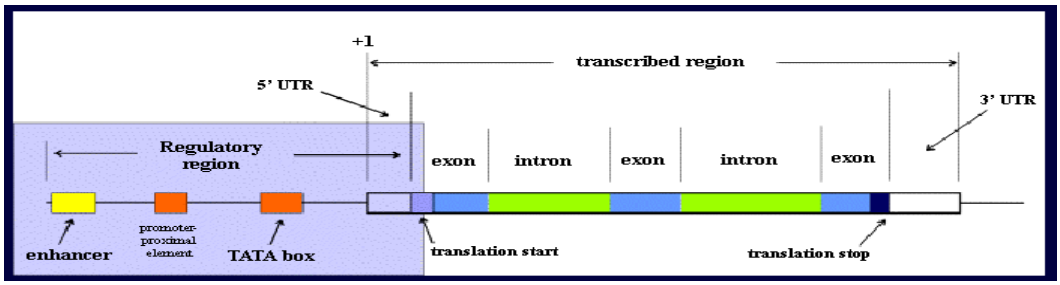
Un gène :

- Un gène est une séquence d'acide désoxyribonucléique (ADN) qui spécifie la synthèse d'une chaîne de polypeptide ou d'un ARN fonctionnel
- Un gène peut donc coder pour un ARN messager ou pour un ARN nonmessager (ARNr, ARNt, …)



- **Génétique:**
  - comment les caractéristiques des organismes vivants (traits génétiques) se transmettent d'une génération à l'autre,
  - **Gregor Mendel (1822-1884)**, étudia les plants de pois et a établi de nombreuses règles de l'hérédité.
  - implique l'étude d'un nombre spécifique et limité de gènes, ou de parties de gènes, qui ont une fonction connue.
  - Dans la recherche biomédicale, les scientifiques tentent de comprendre comment les gènes guident le développement de l'organisme, provoquent des maladies ou affectent la réponse aux médicaments.

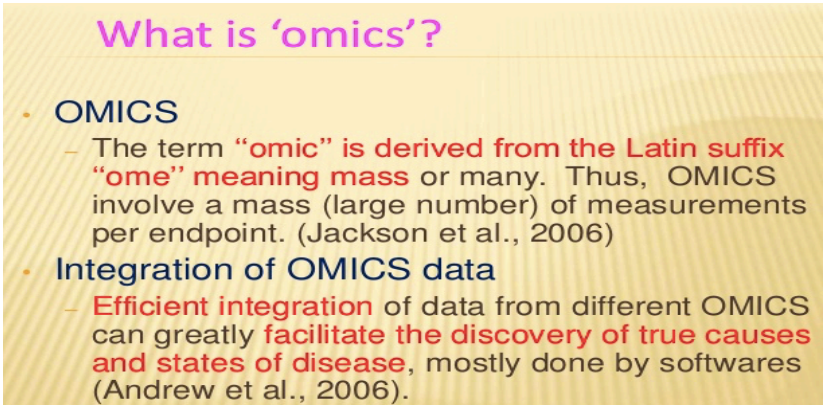
Structure d'un Gène :

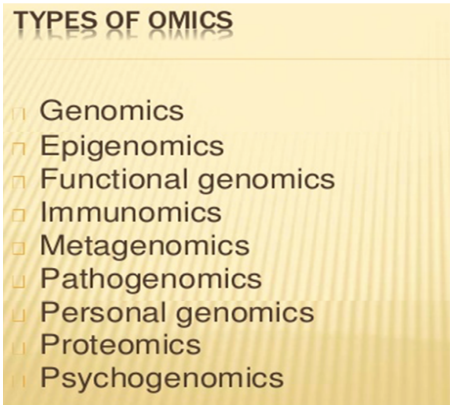


- Basic Definitions:
- Transcription start: Base 1 of the primary transcript
- Transcription stop: Where the RNA polymerase falls off the DNA
- Translation start: The AUG (Methionine) codon
- Translation stop: The termination codon (TAA, TGA, TAG)
- Polyadenylation site: 3' end of the mRNA
- 5' untranslated region: Between base 1 and the AUG, often called the leader
- 3' untranslated region: Between the termination codon and polyadenylation site
- 5' flanking region: From -1 continuing upstream in a 5' direction
- 3' flanking region: Usually from polyadenylation site continuing downstream in 3' direction
- Regulatory region: Refers to enhancer and promoter when these have yet to be characterized

- **Génomique** :

- C'est l'étude de l'ensemble des gènes d'un organisme **: le génome**. En utilisant des techniques de calcul et de mathématiques à haute performance connues sous le nom de bioinformatique,
– les chercheurs en génomique analysent d'énormes quantités de données de séquences d'ADN pour trouver des variations qui affectent la santé, la maladie ou la réponse aux médicaments.
– Chez l'homme, cela signifie rechercher environ 3 milliards d'unités d'ADN dans **23 000 gènes !!.**
– La génomique est un domaine beaucoup plus récent que la génétique et n'est devenue possible qu'au cours des deux dernières décennies grâce aux progrès techniques du séquençage de l'ADN et de la biologie computationnelle.





Le séquençage du génome humain

- 1943-1953: ADN support de l'information génétique
- • 1977: Techniques modernes de séquençage de l'ADN (Sanger)
- • 1981: Séquençage du génome mitochondrial humain
- • 1982: Premières banques de données de séquences (GenBank, EMBL)
- • 1990: Début du projet génome humain (cartographie)
- • 1995: Premier génome complet d'un organisme cellulaire (H. influenzae) • 1999: Chromosome 22 humain
- • 2001: Première ébauche du génome humain
- • 2003: Séquençage du génome humain achevé
- • 2007: Première séquence complète du génome d'un individu

Le Projet de Génome humain :

Le projet du génome humain (HGP) = un effort international de 13 ans (1990-14 avril 2003) pour séquencer les 3 milliards de bp d'ADN humain.
Projet de 300 millions de dollars => U.S. DOE et NIH.

International Human Genome Sequencing Consortium (IHGSC) = groupe de chercheurs financés par des fonds publics

≈ 200 laboratoires aux États-Unis ont soutenu ces efforts + > 18 pays différents à travers le monde ont contribué au HGP.

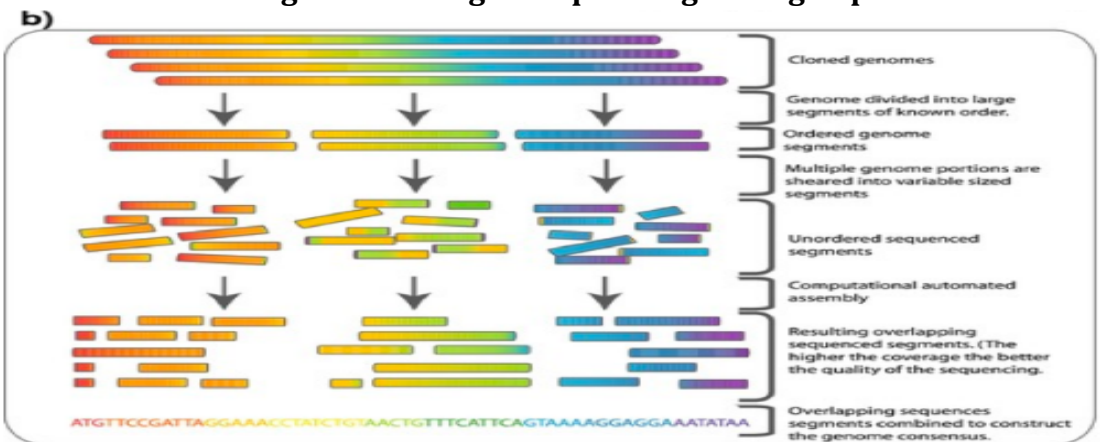
Il y avait deux équipes en compétition pour le séquençage :
- Public (coopération internationale, 1990-14 avril 2003)
- Privé (Celera Genomics, 1998-2000)

Deux Philosophies opposées : - HGP Bermuda Agreement (1996) => toutes les informations du projet seraient mises gratuitement à la disposition de tous dans les 24h.
- Privé
=>accès réservé aux clients payants !

En février 2001, une ébauche de la séquence du génome humain a été publiée simultanément par les deux groupes public-privé dans des articles séparés (Lander et al (IHGSC)., Feb 2001 Nature; Venter et al., Feb 2001 Science) :
http://www.genome.gov/sequencingcosts/ hmp://www.yourgenome.org/
www.sanger.ac.uk

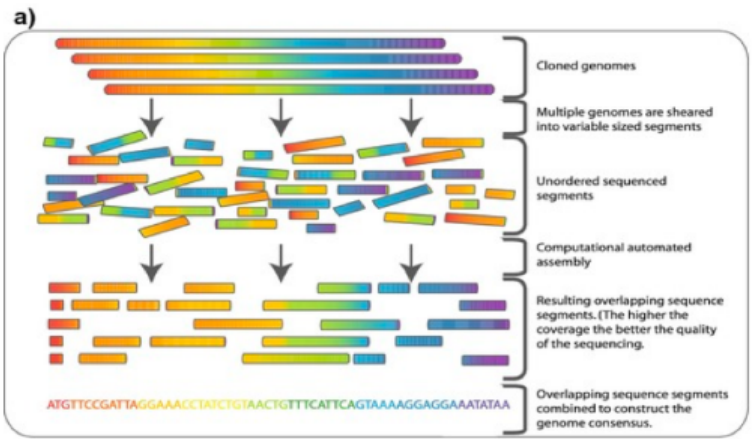
Méthodes de séquençage du génome à haut débit :

- **Hierarchical genome shotgun sequencing - Shotgun phase**



- **Whole-genome shotgun sequencing** (Celera genomics)

- Genome sheared randomly into small fragments (appropriately sized for sequencing) Reassembly.



Encode Project : (https://www.encodeproject.org)
Le projet Encyclop of DNA Elements (ENCODE) vise à fournir "une liste d'éléments fonctionnels du génome humain, y compris des éléments qui agissent au niveau des protéines et de l'ARN, et des éléments régulateurs qui contrôlent les cellules et les circonstances dans lesquelles un gène est actif.
Gencode Project : (http://www.gencodegenes.org)
Le génome humain a fait l'objet d'une annotation manuelle intensive :
Le consortium GENCODE vise à identifier toutes les caractéristiques des gènes dans les génomes humains et murins en utilisant une combinaison d'analyse informatique, d'annotation manuelle et de validation expérimentale.

HapMap project :

Les différences génétiques dans les bases individuelles (SNP) d'un génome sont de loin le type de variation génétique le plus courant.

Objectif : développer une carte des haplotypes du génome humain = identification et catalogage de la plupart des millions de SNP estimés se produire couramment dans le génome humain. hmp://hapmap.ncbi.nlm.nih.gov

1000 Genomes Project : (www.1000genomes.org/)
est devenu plus complet et fiable car de nombreuses nouvelles variantes ont été découvertes !!
Objectifs :
► identifier la plupart des variantes génétiques avec des fréquences d'au moins 1 %.
► ressource librement accessible sur la variation génétique humaine. ensemble de données final = données pour 2 504 individus de 26 populations.
► International Genome Sample Resource (IGSR) pour l'utilisation continue des données générées par le projet 1000 Genomes.

UK10K Project: (www.**uk10k**.org/):
► identification de variants génétiques rares par l'étude de l'ADN de 4 000 individus et leur comparaison avec les zones codant pour les protéines de 6 000 personnes atteintes de maladies documentées.
► lien entre variants génétiques et maladies rares.

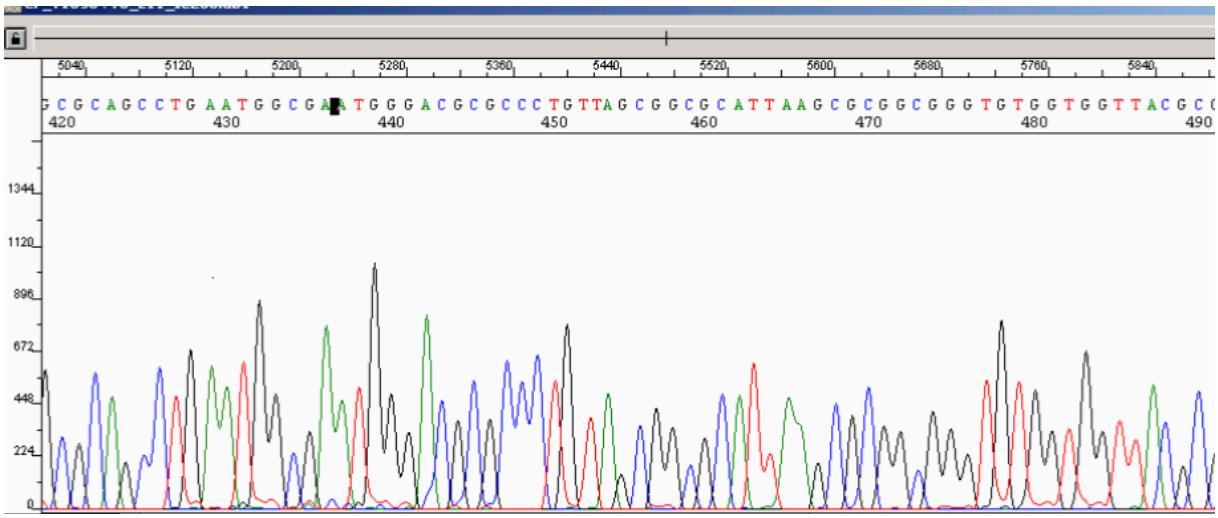
Technologies de séquençage & le NGS

Introduction à la technique de séquençage de l'ADN

What is DNA Sequencing ?

- DNA Sequencing is the process of reading the nucleotides present in DNA : determining the precise order of nucleotides within a DNA molecule.
- DNA-Seq generally refers today to any NGS method or technology that is used to determine the order of the four bases (A, T, C, G) in a strand of DNA.

- In fact, there are two main types of DNA sequencing technologies that are used today: * Sanger sequencing
  * Next-Generation Sequencing
- Each of these technologies has utility in today's genetic analysis environment



Séquençage à haut débit

- Une révolution en génomique fonctionnelle a eu lieu avec l'avènement des technologies de séquençage à très haut débit.
  - Avant depuis les 80 : max 800 à 1000 nucléotides pouvait être séquencé en quelques jours par des techniques lourdes, complexes et dangereuses (utilisation d'isotopes radioactifs)
  - Après qq années: Séquençage simplifié qui séquencent des milliards de nucléotides par expérience, en plus, en temps réel dans des bases de données pour leur analyse.
- En conséquence: de plus en plus de génomes sont séquencés:

⇒ NB: Des biologistes et des informaticiens prévoient que les ressources informatiques nécessaires pour traiter les données liées aux génomes dépasseront à terme celles nécessaires à Twitter et YouTube.
  - On estime que, en 2025, 100 millions à 2 milliards de génomes humains auront été séquencés.
    - A lui seul, le stockage de ces données pourrait nécessiter 2 à 40 exaoctets (1 exaoctet = 10^{18}octets) car les données stockées pour un génome sont 30 fois plus grande que la taille du génome lui-même (données brutes, erreurs, analyse préliminaire ...).
- Le stockage des données ne sera qu'une petite partie du problème : les besoins pour l'acquisition, la distribution et l'analyse des données!!!!!

Caractéristiques élémentaires des NGS :

- Séquençage d'un grand nombre de nucléotides (~ 10^{12} N) par expérience), à un coût << méthode de Sanger.

Séquençage en un temps record : amplifier spécifiquement sans les étapes de clonage.

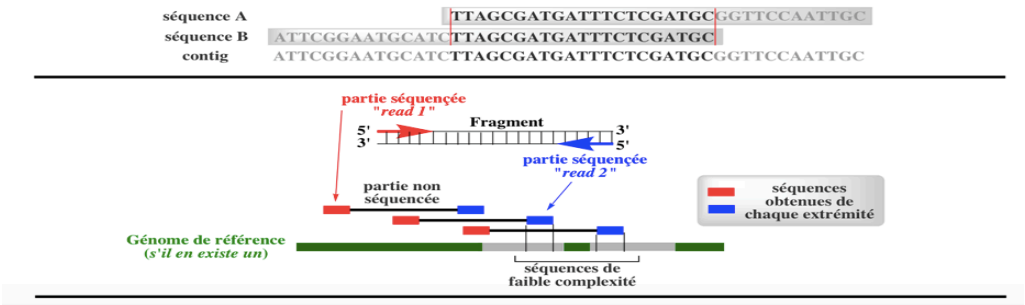
- Ces méthodes sont parallèlisées : des millions de réactions ont lieu en même temps.
- Les fragments séquencés sont courts : actuellement de 30 à environ 250 paires de base selon la technologie.
  - MAIS, la petite taille et le nombre très élevé des fragments séquencés nécessite un travail d'analyse bioinformatique ++ en aval: faut assembler les fragments en contigs.

Avantages du NGS en pratique clinique

- Rapidité : plus de patients, plus de gènes étudiés en simultané
- Coût : inférieur au séquençage de Sanger
- Sensibilité : supérieure au Sanger, détection de sous-clones, suivi de MRD
- Exhaustivité : plus d'exons, gènes entiers
- Économie du matériel biologique : qq ng d'ADN pour un grand nombre de gènes

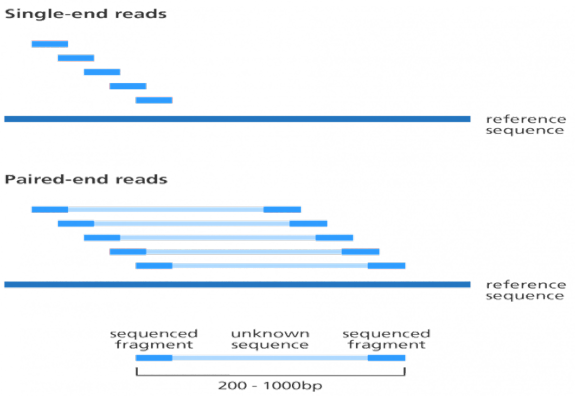
L'inconvénient des NGS :

Est la petite taille des fragments séquencés d'où: un nombre élevé de *gaps*, en particulier pour les régions de faible complexité (exemple : séquences répétées).



- trous ("*gaps*") : parties du génome non séquencées ou dont les séquences ne chevauchent pas avec d'autres et ne peuvent donc entrer dans un contig.=> comblé par un Seq ciblé
- régions de faible compléxité : parties du génome dont les séquences sont très peu diversifiées (exemple : séquences répétées).
  - **Prob**: Lors de l'assemblage, ils peuvent conduire à assembler 2 séquences provenant de régions distantes du génome
  - Solution : , elles sont "masquées" par des programmes informatiques tel que *RepeatMasker*: remplaçent les nucléotides de ces régions par le symbole "N" qui décrit n'importe quel nucléotide.

- ➔ Pour pallier à cette difficulté, on peut séquencer les fragments :
- à partir d'une extrémité ("*single-end sequencing*") : on obtient le début de la séquence du fragment à une extrémité.
- à partir des 2 extrémités ("*paired-end sequencing*") : on obtient le début de la séquence du fragment à une extrémité et le début de la séquence du fragment à l'autre extrémité



- ➔ **Un autre moyen est de construire :**
  - des banques avec des inserts de petites tailles (0,2 - 0,8 kpb)
  - et des banques avec des inserts de grandes tailles (2 - 40 kpb).
  - On obtient ainsi des fragments séquencés de tailles variables **("*short-insert paired end reads*" et "*long-insert paired end reads*" ou "*mate paired*")** qui aboutissent à un meilleur assemblage du fait de contigs plus longs.
- Challenge: +++ de données générées nécessitent le développement d'outils bioinformatiques de plus en plus spécialisés (exemple : Allpaths-LG propose un assembleur pour grands génomes).

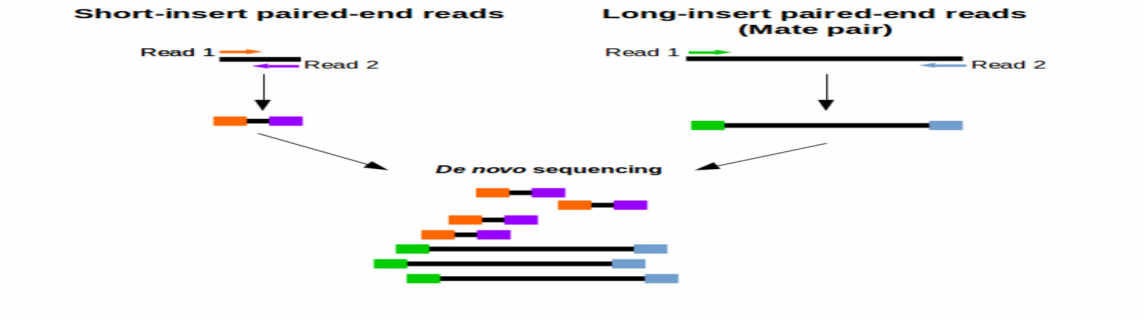
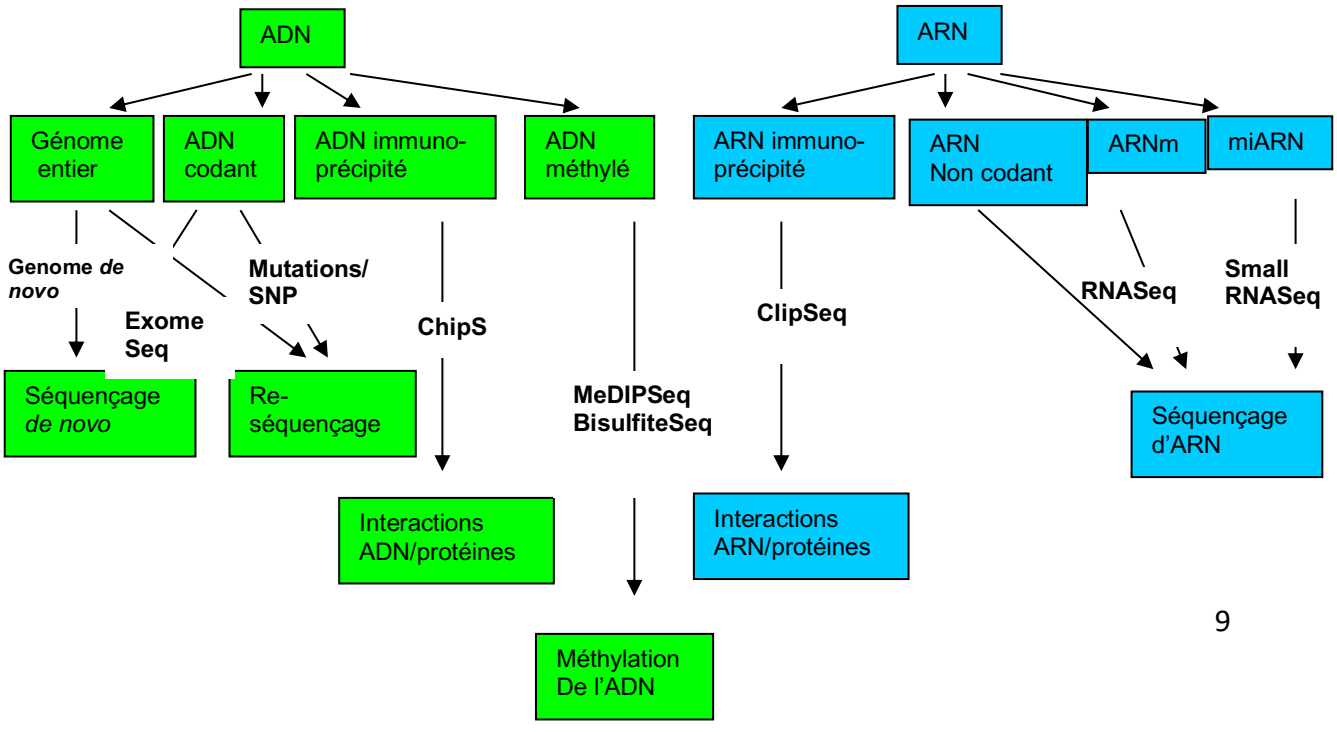


Figure 2: Combining reads from mate pair sequencing with that from short-insert paired-end sequencing for De-novo Sequencing.

https://www.ecseq.com/support/ngs/what-is-mate-pair-sequencing-useful-for

Applications du NGS :



Généralités sur le Protocole du DNA-Seq :

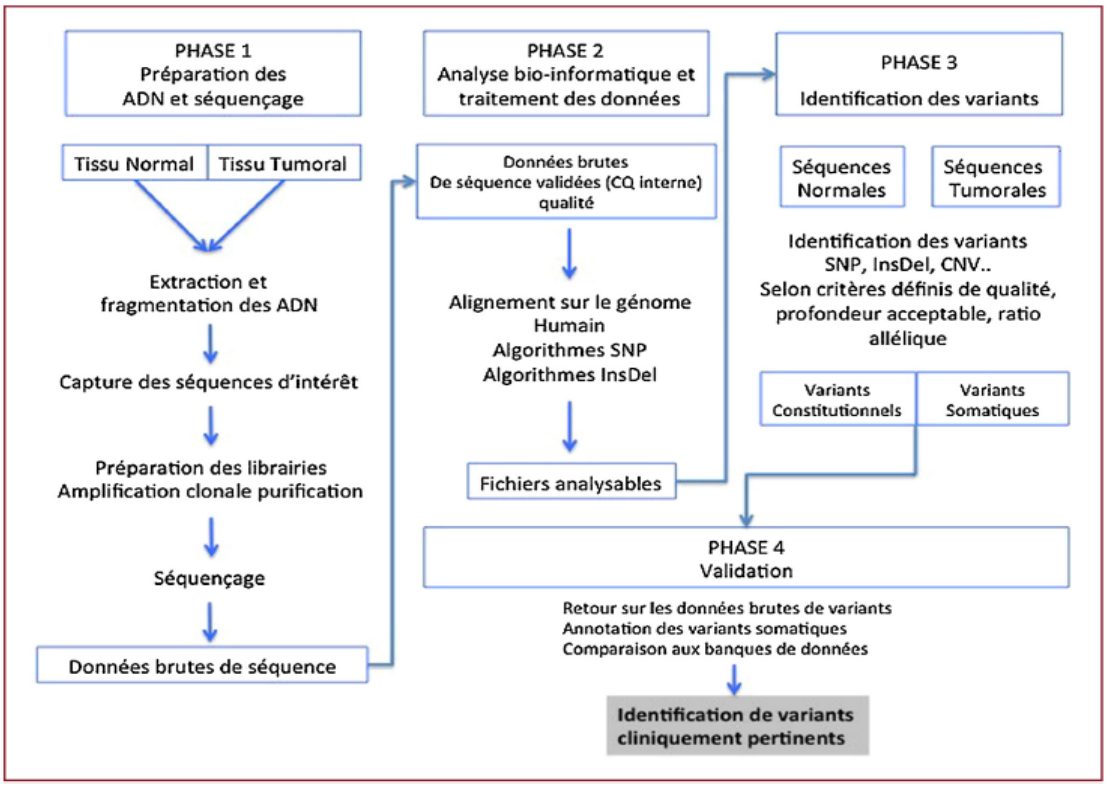


Figure 5. De la technique à l'analyse : les différentes étapes du typage NGS, d'après Ulahannan et al. [1]. Figure résumée des différentes étapes du séquençage et l'interprétation des variants de séquence.
NGS in clinical practice, implementation chart.

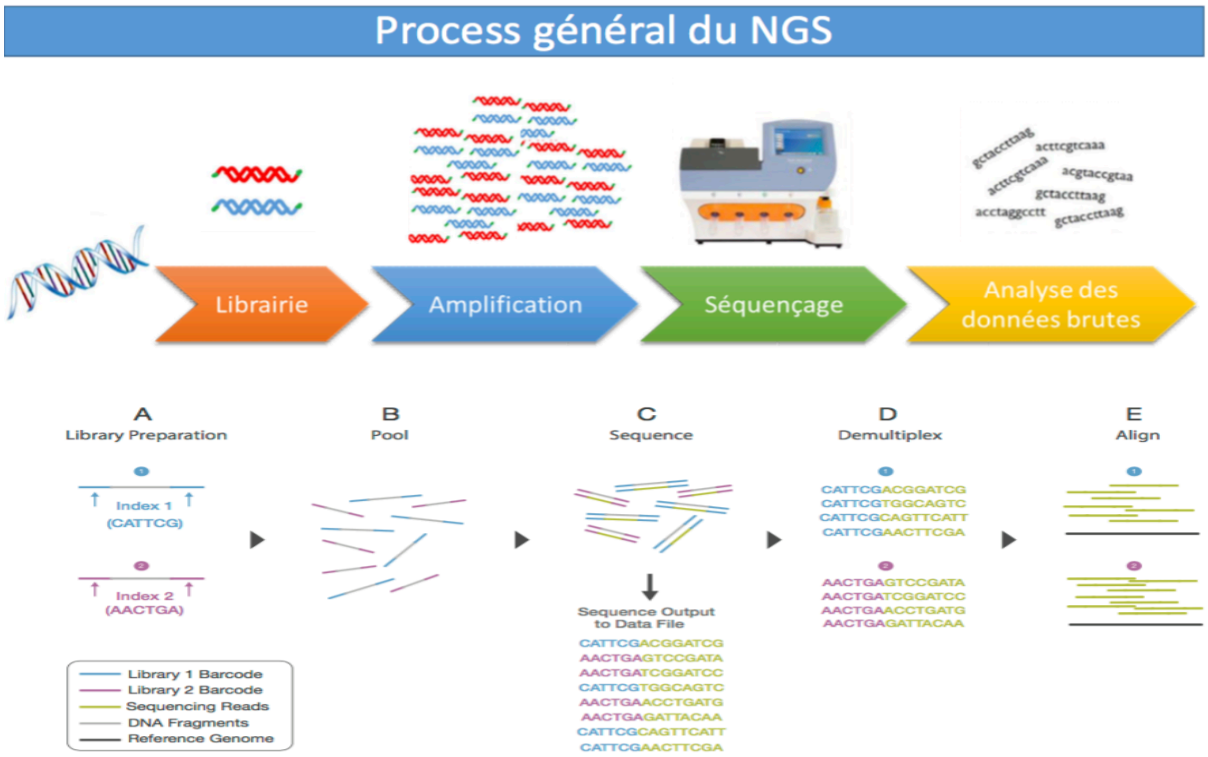


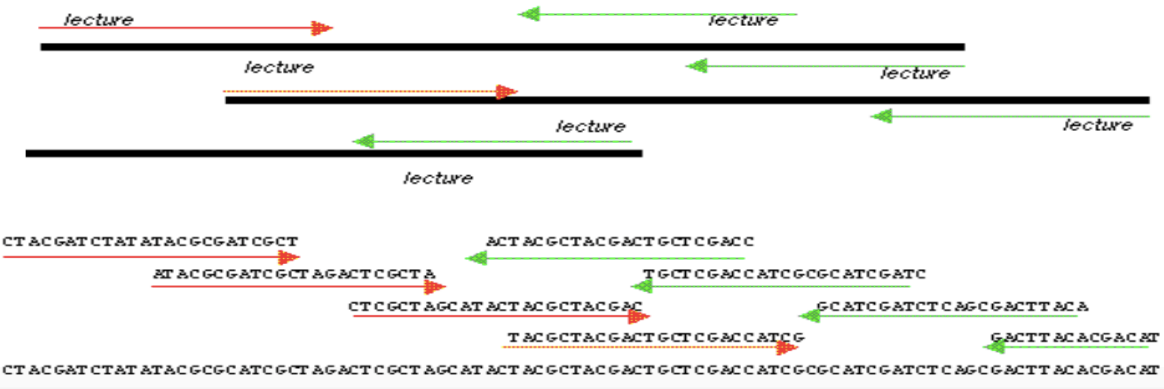
Figure 5: Library Multiplexing Overview.
a. Two distinct libraries are attached to unique index sequences. Index sequences are attached during library preparation.
b. Libraries are pooled together and loaded into the same flow cell lane.
c. Libraries are sequenced together during a single instrument run. All sequences are exported to a single output file.
d. A demultiplexing algorithm sorts the reads into different files according to their indexes.
e. Each set of reads is aligned to the appropriate reference sequence.

L'assemblage du génome

Assemblage = Ensemble de séquences approximant le mieux possible la séquence d'un génome

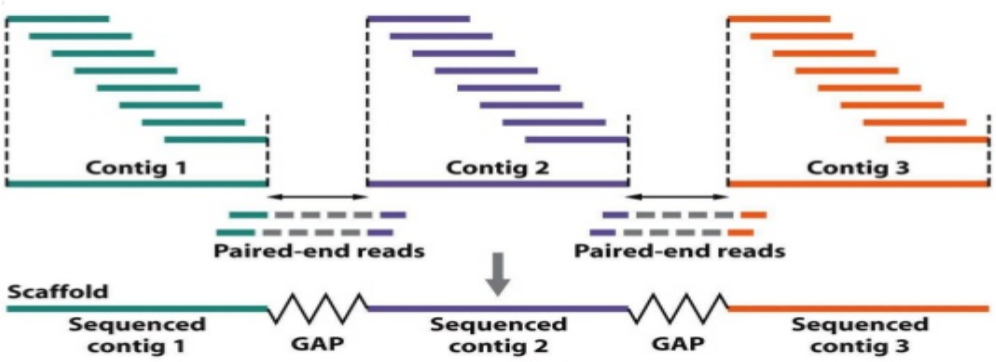
- Avec les technologies encore courantes dans de nombreux laboratoires, chaque séquençage ne permet d'obtenir une lecture que de quelques milliers de paires de base.
  - Il n'est donc pas possible de séquencer en une seule fois des molécules d'ADN aussi grandes que les chromosomes.
  - Pour reconstituer ces immenses séquences:
  - il faut effectuer un grand nombre de séquençages, plusieurs fois >>> la taille du chr.
  - Ces séquençages redondants permettent : de?
    - de raccorder les séquences les unes aux autres
    - de s'assurer de la qualité du résultat de chaque lecture

La comparaison des séquences permet



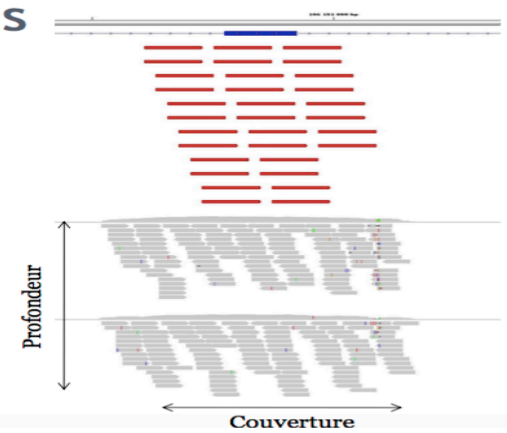
s grands

En reliant l'ensemble des contigs, on reconstitue des séquences de plusieurs millions à plusieurs dizaines de millions de nucléotides
=> Ces opérations sont effectuées par des programmes bioinformatiques.



Quelques définitions

- **Couverture**
  = zone du génome couverte par un nombre suffisant de lecture
- **Profondeur**
  = nombre de lecture de chaque base

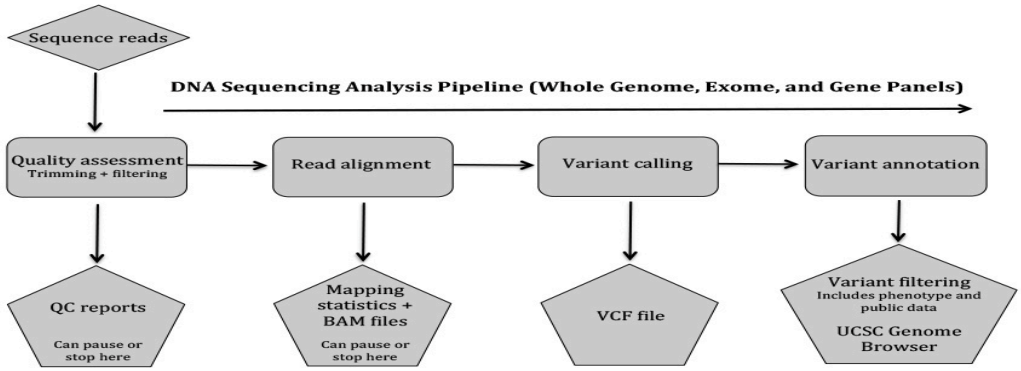


La méthode "*shotgun*"

C'est un processus aléatoire d'échantillonnage de N lectures de taille L, pour un génome de taille G :

- couverture : $a = N . L / G$
- nombre de contig obtenus (N_c) en fonction de la couverture : $N_c = (a . G / L) e^{-a}$
- taille moyenne des contigs : $L_c = (e^a - 1) . L / a$

Le Pipeline d'analyse des données NGS:



DNA sequence data / alignment statistics

Fastqc reads length, reads number, base quality,
Samtool sequence quality, sequence duplication levels,

DNA sequence data aligners

BWA, Bowtie, Lifescope - Indexing Genome with Suffix Array
BFAST, Novoalign - Indexing Genome with Hash Tables
SHRiMP2 - Indexing Reads with Hash Tables

DNA sequence data variant callers

GATK - indel realignment, base recalibration, variant calling (unified genotyper caller and haplotype caller) and filtering
FreeBayes - haplotype-based, Bayesian algorithm designed to find small polymorphisms, specifically SNPs, indels, MNPs, and complex events
Samtools mpileup - scans every position, computes all the possible genotypes, and the probability of these genotypes

RNA sequence data / alignment statistics

Fastqc	reads length, reads number, base quality, sequence quality, sequence duplication levels, etc.
Samtools	alignment quality, number of high quality aligned reads,
Picard	base mismatch rate, strand balance, duplicate reads, etc.

RNA sequence data aligners / Transcriptome assemblers

Tophat/Bowtie/Cufflinks - to identify exon-exon splice junctions, assembles transcripts, estimates their abundances, and tests for differential expression and regulation (2009)
Subread/Subjunc/featureCounts – alignment, exon-exon junction detection and read

Expression data statistical analysis

Differential expression analysis
High dimensional analysis
Enrichment/pathways analysis
Analysis software developed in house

Les formats de fichiers d'Alignement des séquences :
– SAM et BAM (= standards)
– ELAND (specific Illumina)
– MAQ Map

Le fichier SAM

● NGS => a variety of new alignment tools :
Bowtie (Langmead,B. et al (2009), Maq (Li,H. et al (2008), BWA (Li and Durbin, 2009), ...
● SAM : Sequence Alignment/Map format
● SAM : a common alignment format that supports and stores all sequence types and aligners
● A well-defined interface between alignment
and downstream analyses

Le fichier BAM

● BAM = compressed SAM
● BAM indexé: *.bam.bai

Formats « Variant Calling » :

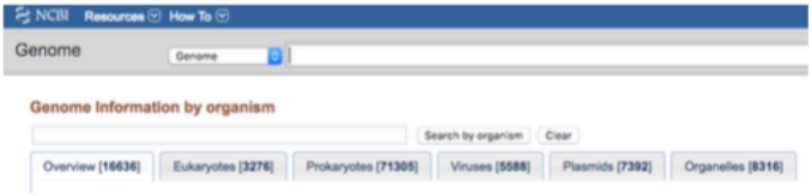
● Deux formats de fichiers sont couramment utilisés :
  – Format Pileup
  – Format VCF
● Le format Pileup est spécifique de l'outil Samtools Mpileup

● Le format VCF est le format par défaut d'un grand nombre de SNP caller Formats « Variant Calling »

Les Genomes Browsers :

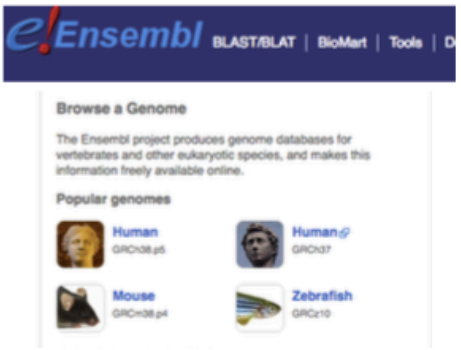
- Systematically integrate omics data by overlaying genomic sequence data with annotations
- We are able to:
    - - Explore regulatory & structural features around a gene or chr region
    - - Explore variation for a specific region/gene
    - - Compare the region to other genomes
    - - Retrieve annotated data for a specific region/gene
    - - Link out to other sources of informa6on

-Web-based or stand alone applications (e.g. GenomeView)
Multi-species (compare across species)
species-specific (focus on 1 model organism)



http://www.ncbi.nlm.nih.gov/mapview/

https://genome.ucsc.edu/index.html

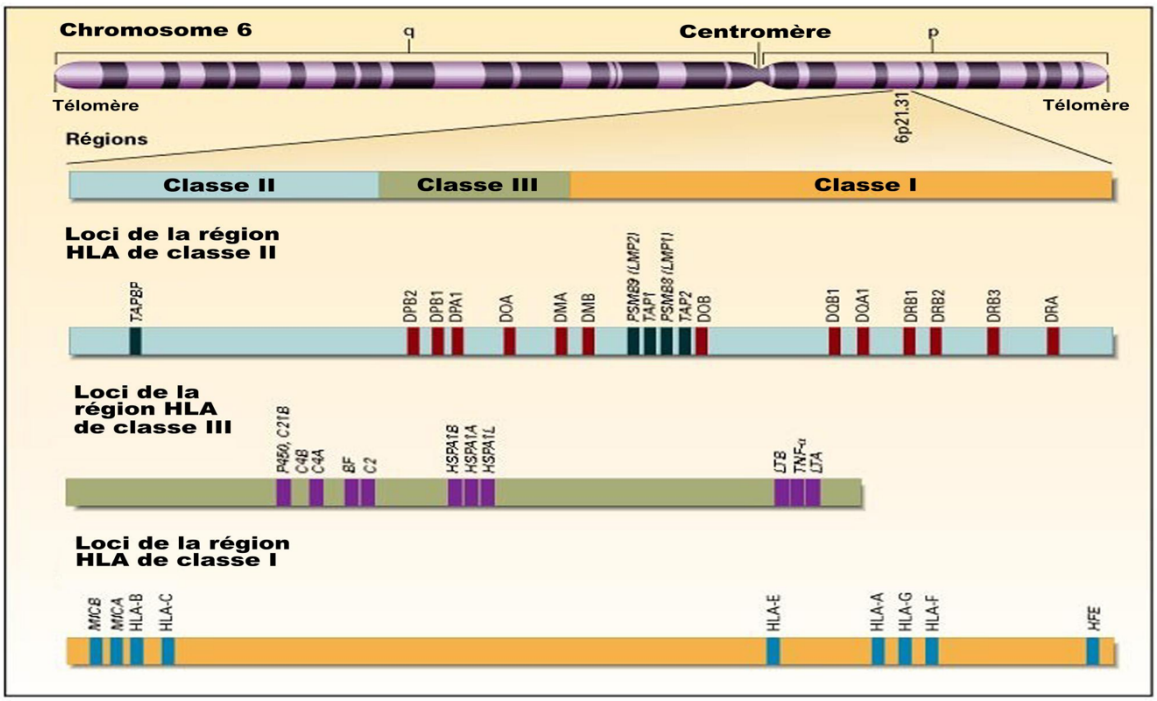


http://www.ensembl.org/index.html

Des séances de travaux pratiques et dirigés

Installation de l'application IGV « Analyses des séquences

IMMUNOGENOMIQUE



Fonctions du système HLA :

- **Sélection thymique et induction de la tolérance**
- **Transport et Présentation des peptides antigéniques**

La Sélection Positive :

Au niveau du cortex thymique:
➔ les thymocytes qui sont CD4+8+ subissent une dichotomie:
- ceux qui reconnaissent le HLA de classe I porté par les cellules épithéliales prolifèrent en CD8+
- que ceux qui reconnaissent le HLA de classe II prolifèrent en CD4+.

➔ Tous les thymocytes dont ce réarrangement n'est pas fait ou mal fait meurent par apoptose.
➔ Les thymocytes (avec TCRαβ-CD3) incapable de reconnaître chacune des molécules du HLA, ne reçoivent aucun signal de différenciation et meurent par apoptose. Ce type de sélection confère aux cellules T la restriction de reconnaissance du soi par le HLA.

La Sélection Négative :

Au niveau de la jonction cortico-médullaire et de la médullaire :
➔ Élimination des cellules T auto-réactives
Les thymocytes T qui reconnaissent les peptides du soi présentés par les molécules HLA de classe I et de classe II portées par les CPA ou les cellules thymiques sont rapidement éliminées par apoptose

==> Induction de la tolérance au soi.

N.B: Certaines de ces cellules autoréactives échappent à la sélection négative et se retrouvent dans le milieu périphérique.

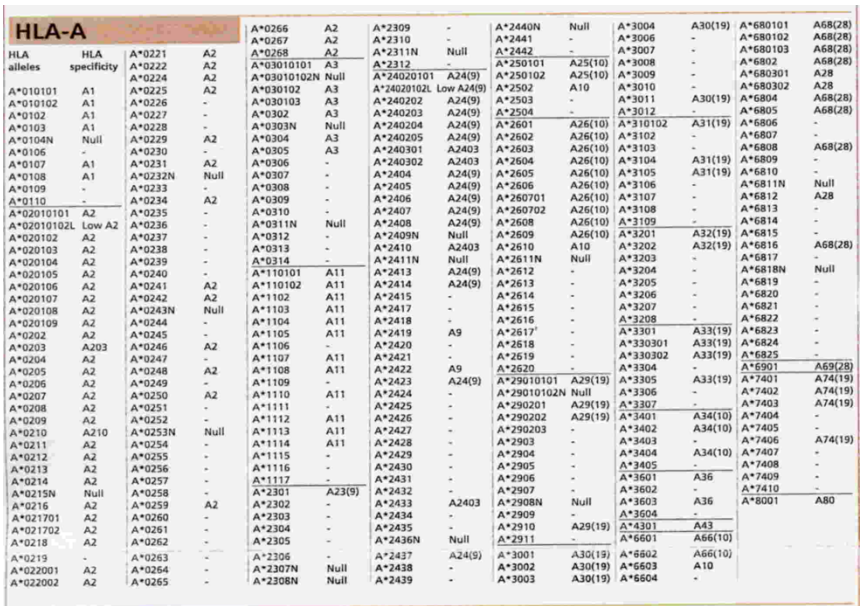
Polymorphisme du système HLA

Le système HLA est très polymorphe tant sur le plan antigénique qu'allélique :

Les antigènes HLA :

A	B		C	DR	DQ	DP
A1	B5	B50(21)	Cw1	DR1	DQ1	DP 1
A2	B7	B51(5)	Cw2	DR103	DQ2	DP 2
A203	B703	B5102	Cw3	DR2	DQ3	DP 4
A210	B8	B5103	Cw4	DR3	DQ4	DP 5
A3	B12	B52(5)	Cw5	DR4	DQ5(1)	DP 6
A9	B13	B53	Cw6	DR5	DQ6(1)	
A10	B14	B54(22)	Cw7	DR6	DQ7(3)	
A11	B15	B55(22)	Cw8	DR7	DQ8(3)	
A19	B16	B56(22)	Cw9(W3)	DR8	DQ9(3)	
A23(9)	B17	B57(17)	Cw10(w3)	DR9		
A24(9)	B18	B58(17)		DR10		
A2403	B21	B59		DR11(5)		
A25(10)	B22	B60(40)		DR12(5)		
A26(10)	B27	B61(40)		DR13(6)		
A28	B35	B62(15)		DR14(6)		
A29(19)	B37	B63(15)		DR1404		
A30(19)	B38(16)	B64(14)		DR15(2)		
A31(19)	B3901	B65(14)		DR16(2)		
A32(19)	B3902	B67		DR17(3)		
A33(19)	B40	B70		DR18(3)		
A34(10)	B4005	B71(70)				
A36	B41	B72(70)				
A43	B42	B73				
A66(10)	B44(12)	B75(15)				
A68(28)	B45(12)	B76(15)				
A69(28)	B46	B77(15)				
A74(19)	B47	B7801				
	B48	Bw4				
	B49(21)	Bw6				

Les Allèles HLA-A:

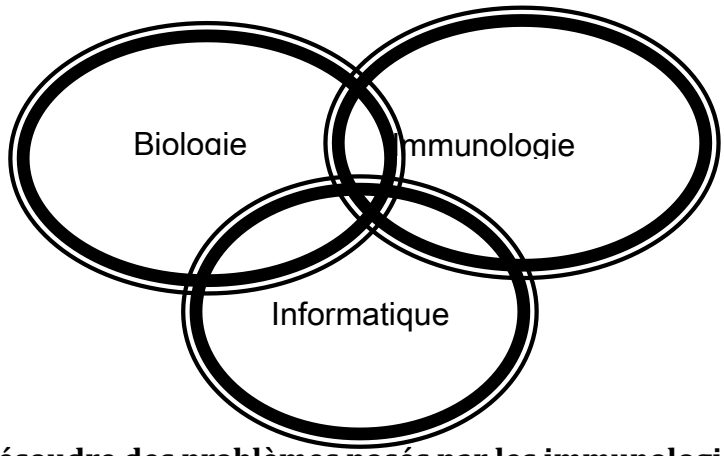


HLA-A

HLA alleles	HLA specificity
A*010101	A1
A*010102	A1
A*0102	A1
A*0103	A1
A*0104N	Null
A*0106	-
A*0107	A1
A*0108	A1
A*0109	-
A*0110	-
A*02010101	A2
A*02010102L	Low A2
A*020102	A2
A*020103	A2
A*020104	A2
A*020105	A2
A*020106	A2
A*020107	A2
A*020108	A2
A*020109	A2
A*0202	A2
A*0203	A203
A*0204	A2
A*0205	A2
A*0206	A2
A*0207	A2
A*0208	A2
A*0209	A2
A*0210	A210
A*0211	A2
A*0212	A2
A*0213	A2
A*0214	A2
A*0215N	Null
A*0216	A2
A*021701	A2
A*021702	A2
A*0218	A2
A*0219	-
A*022001	A2
A*022002	A2
A*0221	A2
A*0222	A2
A*0224	A2
A*0225	A2
A*0226	-
A*0227	-
A*022B	-
A*0229	A2
A*0230	-
A*0231	A2
A*0232N	Null
A*0233	-
A*0234	A2
A*0235	-
A*0236	-
A*0237	-
A*0238	-
A*0239	-
A*0240	-
A*0241	A2
A*0242	A2
A*0243N	Null
A*0244	-
A*0245	-
A*0246	A2
A*0247	-
A*0248	A2
A*0249	-
A*0250	A2
A*0251	-
A*0252	-
A*0253N	Null
A*0254	-
A*0255	-
A*0256	-
A*0257	-
A*0258	-
A*0259	A2
A*0260	-
A*0261	-
A*0262	-
A*0263	-
A*0264	-
A*0265	-
A*0266	A2
A*0267	A2
A*0268	A2
A*03010101	A3
A*03010102N	Null
A*030102	A3
A*030103	A3
A*0302	A3
A*0303N	Null
A*0304	A3
A*0305	A3
A*0306	-
A*0307	-
A*0308	-
A*0309	-
A*0310	-
A*0311N	Null
A*0312	-
A*0313	-
A*0314	-
A*110101	A11
A*110102	A11
A*1102	A11
A*1103	A11
A*1104	A11
A*1105	A11
A*1106	-
A*1107	A11
A*1108	A11
A*1109	-
A*1110	A11
A*1111	-
A*1112	A11
A*1113	A11
A*1114	A11
A*1115	-
A*1116	-
A*1117	-
A*2301	A23(9)
A*2302	-
A*2303	-
A*2304	-
A*2305	-
A*2306	-
A*2307N	Null
A*2308N	Null
A*2309	-
A*2310	-
A*2311N	Null
A*2312	-
A*24020101	A24(9)
A*24020102L	Low A24(9)
A*240202	A24(9)
A*240203	A24(9)
A*240204	A24(9)
A*240205	A24(9)
A*240301	A2403
A*240302	A2403
A*2404	A24(9)
A*2405	A24(9)
A*2406	A24(9)
A*2407	A24(9)
A*2408	A24(9)
A*2409N	Null
A*2410	A2403
A*2411N	Null
A*2413	A24(9)
A*2414	A24(9)
A*2415	-
A*2417	-
A*2418	-
A*2419	A9
A*2420	-
A*2421	-
A*2422	A9
A*2423	A24(9)
A*2424	-
A*2425	-
A*2426	-
A*2427	-
A*2428	-
A*2429	-
A*2430	-
A*2431	-
A*2432	-
A*2433	A2403
A*2434	-
A*2435	-
A*2436N	Null
A*2437	A24(9)
A*2438	-
A*2439	-
A*2440N	Null
A*2441	-
A*2442	-
A*250101	A25(10)
A*250102	A25(10)
A*2502	A10
A*2503	-
A*2504	-
A*2601	A26(10)
A*2602	A26(10)
A*2603	A26(10)
A*2604	A26(10)
A*2605	A26(10)
A*2606	A26(10)
A*260701	A26(10)
A*260702	A26(10)
A*2608	A26(10)
A*2609	A26(10)
A*2610	A10
A*2611N	Null
A*2612	-
A*2613	-
A*2614	-
A*2615	-
A*2616	-
A*2617	-
A*2618	-
A*2619	-
A*2620	-
A*29010101	A29(19)
A*29010102N	Null
A*290201	A29(19)
A*290202	A29(19)
A*290203	-
A*2903	-
A*2904	-
A*2905	-
A*2906	-
A*2907	-
A*2908N	Null
A*2909	-
A*2910	A29(19)
A*2911	-
A*3001	A30(19)
A*3002	A30(19)
A*3003	A30(19)
A*3004	A30(19)
A*3006	-
A*3007	-
A*3008	-
A*3009	-
A*3010	-
A*3011	A30(19)
A*3012	-
A*310102	A31(19)
A*3102	-
A*3103	-
A*3104	A31(19)
A*3105	A31(19)
A*3106	-
A*3107	-
A*3108	-
A*3109	-
A*3201	A32(19)
A*3202	A32(19)
A*3203	-
A*3204	-
A*3205	-
A*3206	-
A*3207	-
A*3208	-
A*3301	A33(19)
A*330301	A33(19)
A*330302	A33(19)
A*3304	-
A*3305	A33(19)
A*3306	-
A*3307	-
A*3401	A34(10)
A*3402	A34(10)
A*3403	-
A*3404	A34(10)
A*3405	-
A*3601	A36
A*3602	-
A*3603	A36
A*3604	-
A*4301	A43
A*6601	A66(10)
A*6602	A66(10)
A*6603	A10
A*6604	-
A*680101	A68(28)
A*680102	A68(28)
A*680103	A68(28)
A*6802	A68(28)
A*680301	A28
A*680302	A28
A*6804	A68(28)
A*6805	A68(28)
A*6806	-
A*6807	-
A*6808	A68(28)
A*6809	-
A*6810	-
A*6811N	Null
A*6812	A28
A*6813	-
A*6814	-
A*6815	-
A*6816	A68(28)
A*6817	-
A*6818N	Null
A*6819	-
A*6820	-
A*6821	-
A*6822	-
A*6823	-
A*6824	-
A*6825	-
A*6901	A69(28)
A*7401	A74(19)
A*7402	A74(19)
A*7403	A74(19)
A*7404	-
A*7405	-
A*7406	A74(19)
A*7407	-
A*7408	-
A*7409	-
A*7410	-
A*8001	A80

L'immunoinformatique

- **C'est une Branche de l'immunologie et de la bioinformatique au ss.**
- **C'est l'intersection entre l'immunologie, la biologie et l'informatique.**



- **Elle permet de résoudre des problèmes posés par les immunologistes.**
- **Offre une meilleure plateforme pour avancer la recherche sur le vaccin et les traitements immunothérapeutiques.**

Comment interagit le peptide de liaison avec la molécule du HLA ?:

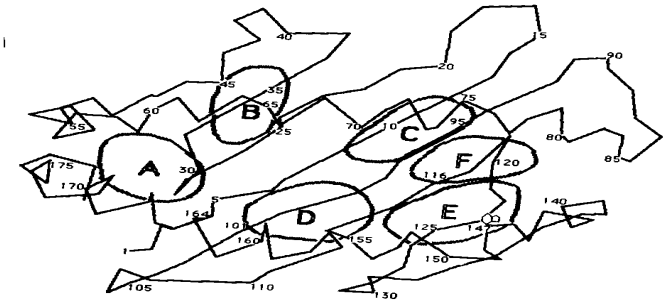
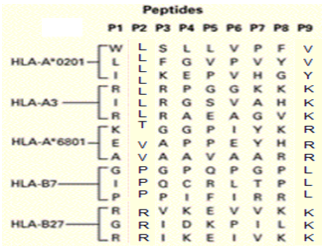


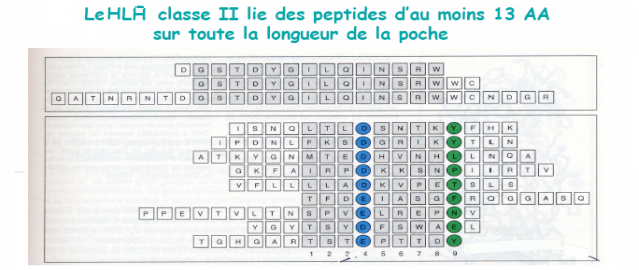
Schéma de la molécule HLA de classe I avec ses différentes poches

- 6 poches : A,B,C,D,E et F
- les résidus des poches A et F :
  ➔ orientation du peptide et de son ancrage dans le sillon
- Les résidus polymorphes des poches B, C, D et E :
  ➔ influencent la spécificité de liaison

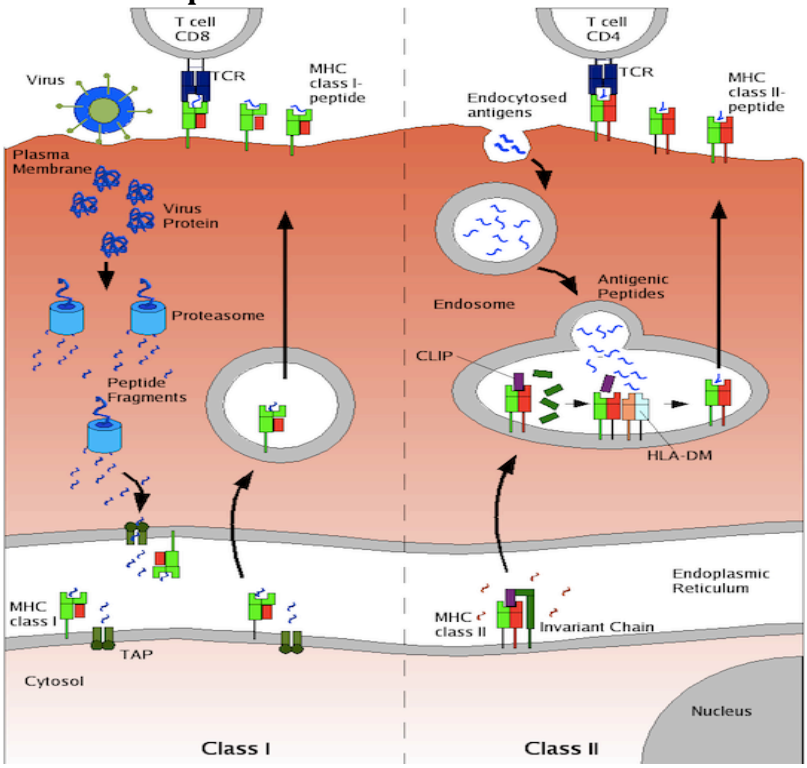
Taille des peptides de liaison

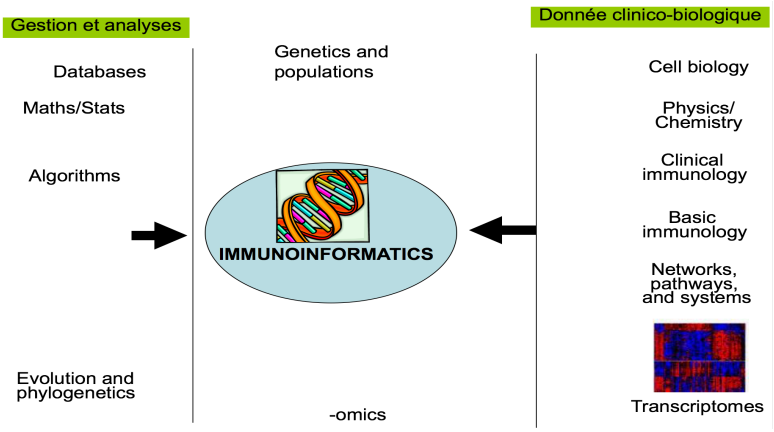


Peptides for specific MHC Class I share sequence motif at anchor residues P2 and P9.



Les Voies de présentation des molécules HLA de Classe I et de classe II

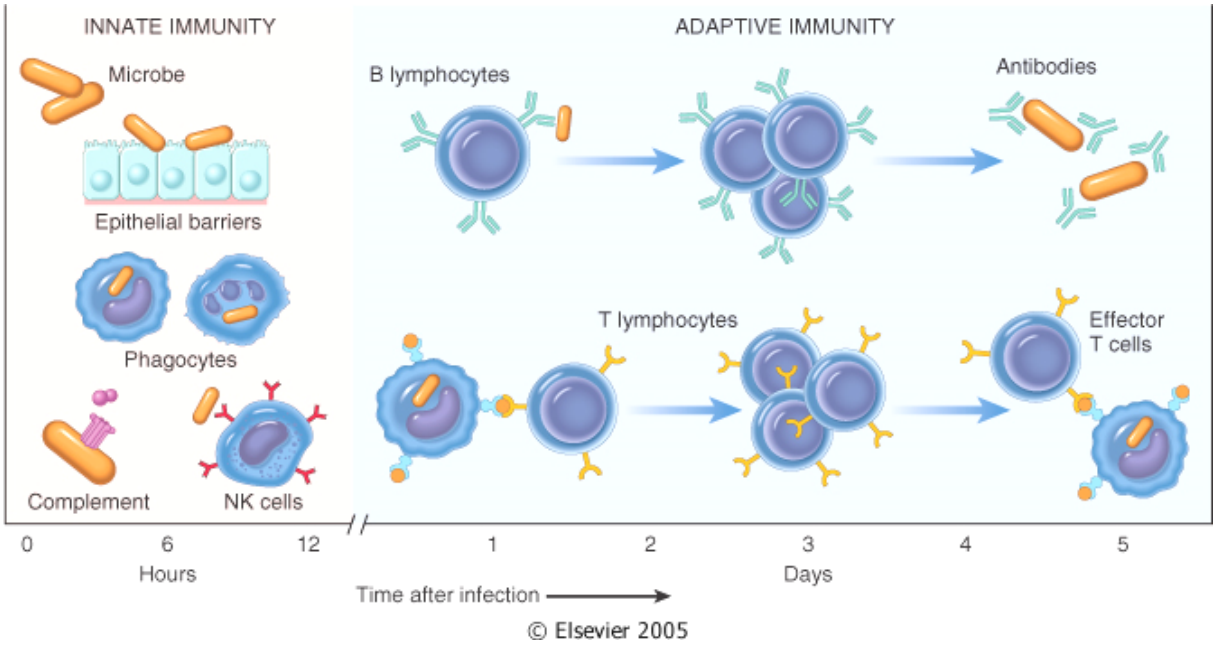




Objectifs globaux: La Lutte contre les maladies :

1. Investigation génome : detection des marqueurs
2. Proteomics/genomics : patients/CTRL
3. Analyses des séquences des antigènes/marqueurs
4. Analyses de la structure des antigènes
5. Analyse de la Structure des épitopes T (lym T)
6. Analyse de la Structure des épitopes B (lym B)
   => analyse des Igs
7. Conception de vaccin

Les principaux mécanismes de contrôle de la réponse immunitaire innée et adaptative :



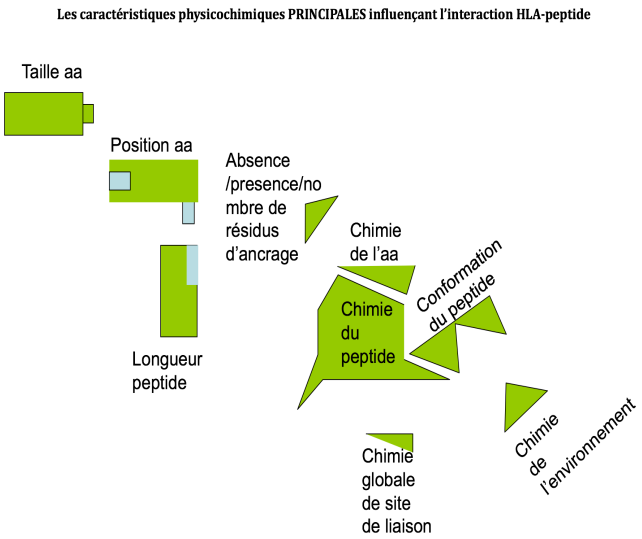
Genomics *vs.* Immunomics

- Genomics: pour une question bien précise
  - 10^4 genes => 10^6 produits
- Immunomics: pour une réponse immune
  - 10^2-10^3 genes => >10^{12} produits
  - Donc: une grande diversité en immunomics controlant la fonction de RI

- >10^{13} MHC class I haplotypes (IMGT-HLA)
- 10^7-10^{15} T cell receptors (Arstila *et al.*, 1999)
- >10^9 combinatorial antibodies (Jerne, 1993)
- 10^{12} B cell clonotypes (Jerne, 1993)
- 10^{11} linear epitopes composed of nine amino acids
- >>10^{11} conformational epitopes

Principe de la selection des T épitopes :

- Interaction peptide-HLA selon le Principe du clé et la serrure.
- Existence de trois critères de sélection de peptides :
  - Forme
  - Taille
  - Paramètre physique: force électrostatique
  - chimique : acide /base



Modélisation in-silico pour l'identification des épitopes T ou B des lymphocytes

- Problème: le nombre d'expériences/temps +++
- Solution:)> Réduire)> utilisation des méthodes in-silico pour prédire les peptides de liaison.
- Economies des expériences du labo.
- Gain du temps pour la découverte de peptides spécifiques a des fins diagnostiques et vaccinales.
- développement des algorithmes

A partir d'une protéine prédire les épitopes T ou B et à l'aide des programmes bioinformatique en tenant compte de la spécificité au niveau :

- allélique (allèles ou sous-types) en Gnr.
- allélique (allèles ou sous-types) par rapport à des spécificités pathologiques

Exp: HLA-B2702, HLA-B5101 et HLA-DR0401….SONT associés à certaines maladies auto-immunes.

- Initialement:
- Prédiction de peptide de liaison avec le HLA
- Decouverte que les peptides se liant specifiquement au HLA SONT FONCTIONNELEMENT liés et partage en commun des propriétés similaires ds différente positions au niveau de leurs séquences Iere.

poche d'ancrage: complémentarité et adaptation des résidus entre ceux des peptides de liaison et leurs homologues au nv des allèles HLA.=> résidus d'ancrage : lie pep au nv diff position de la poche de HLA

==> motif peptidique

Plusieurs auteurs bioinformaticiens ont développé des outils computationnels qui scannent des peptides qui pourraient se fixer dans le HLA : contient le motifs.

➔ SYFPEITHI

➔ Propred

➔ autres …

Des séances de travaux pratiques et dirigés

*** Utilisation de :**

➔ **SYFPEITHI**

➔ **Propred**

*** Présentation des étudiants sur les sujets suivants :**

TITRES DES PRESENTATIONS
Bioinformatics Approach to Identify Significant Biomarkers, Drug Targets Shared Between Parkinson's Disease and Bipolar Disorder: A Pilot Study
Assessing the utilization of high-resolution 2-field HLA typing in solid organ transplantation
Distribution of HLA-DRB1 alleles in BRICS countries with a high tuberculosis burden: a systematic review and meta-analysis
HIV1 and human genetic variation
The HLA-B 221 dimorphism impacts on NK cell education and clinical outcome of immunotherapy in acute myeloid leukemia
Spondyloarthritis and the Human Leukocyte Antigen (HLA)-B∗27 Connection
PD-L1 Expression, Tumor Mutational Burden, and Cancer Gene Mutations Are Stronger Predictors of Benefit from Immune Checkpoint Blockade than HLA Class I Genotype in Non–Small Cell Lung Cancer
Low Hydrophobic Mismatch Scores Calculated for HLA-A/B/DR/DQ Loci Improve Kidney Allograft Survival
Polymorphisme HLA et maladie de Behçet dans la population marocaine HLA polymorphism and Behc¸et's disease in Moroccan population
Immunopeptidomic Profiling of HLA-A2-Positive Triple Negative Breast Cancer Identifies Potential Immunotherapy Target Antigens

Ressources bibliographiques

- Genomes. 2nd edition.Oxford: Wiley-Liss; 2002.
https://www.ncbi.nlm.nih.gov/books/NBK21134/
- https://www.zoology.ubc.ca/~bio463/lecture_2.htm

https://telum.umc.edu.dz/enrol/index.php?id=646

The Human Genome And What We Do With It! Dr. Kaitlin Wade 22nd June 2016: https://www.bristol.ac.uk/alspac/external/presentations/The-human-genome-and-what-we-do-with-it.pdf

_ Immunoinformatics? Bioinformatic Strategies for Better Understanding of Immune Function? Gregory R. Bock, Jamie A. Goode.
https://www.perlego.com/book/2769772/immunoinformatics-bioinformatic-strategies-for-better-understanding-of-immune-function-pdf

- Sadki K, Bakri Y, Tijane M and Amzazi S. MHC Polymorphism and Tuberculosis Disease. Understanding Tuberculosis. 343-354, 2012. Pere-Joan Cardona, IntechOpen, DOI: 10.5772/29742. Available from: https://www.intechopen.com/books/understanding-tuberculosis-analyzing-the-origin-of-mycobacterium-tuberculosis-pathogenicity/mhc-polymorphism-and-tuberculosis-disease