

Séquençage et NGS

C'est une fiche complète et MAJ avec le présentiel de cette année. Le cours n'a pas changé depuis l'année dernière mais je vous ai rajouté quelques remarques que la prof a fait en présentiel (+++ notamment sur ce qu'elle veut que vous reteniez +++). Le cours peut paraître long et compliqué à comprendre mais elle ne vous interrogera pas sur les détails alors essayez de comprendre les grandes lignes !!! Et n'hésitez pas à me poser des questions (forum ou discord) pour que je vous réexplique certaines notions !!!

Séquençage de l'ADN

BUT : Le séquençage d'un fragment d'ADN est une technique qui permet de déterminer la succession des nucléotides qui le compose.

Pour rappel, l'ADN est cette **double hélice** composée des 4 nucléotides A/T/C/G reliés par des **liaisons phosphodiester** (dans un même brin, entre les nucléotides), avec ces deux brins antiparallèles, liés entre eux par des **liaisons hydrogènes** (entre les deux brins).

Le principe du séquençage d'ADN (aussi appelée **méthode de Sanger**) est basé sur les **di-désoxynucléotides** (on explique juste après pourquoi).

C'est une méthode enzymatique, aujourd'hui **méthode de référence pour séquencer l'ADN**. Malgré l'apparition de nouvelles technologies que l'on abordera à la fin du cours, comme le Séquençage Haut Débit, la méthode de Sanger reste la méthode de référence.

1. Principe

Globalement le principe est le même que la PCR, les étapes sont très proches. L'amplification sera un peu différente.

On va utiliser :

- Le brin d'ADN à séquencer
- **Une seule amorce (= primer)**



Puisqu'on ne veut lire qu'un seul brin d'ADN à la fois. Dans le séquençage ce sont exactement les mêmes amorces qu'on a vu dans la PCR, sauf que là au lieu d'en mettre une en amont et une en aval, il n'y en a qu'une !

- Une ADN polymérase (pour la synthèse du brin d'ADN complémentaire) qui synthétise de 5' en 3' en incorporant les nucléotides. Comme pour la PCR sauf qu'on le fait dans un seul sens

2. Etapes

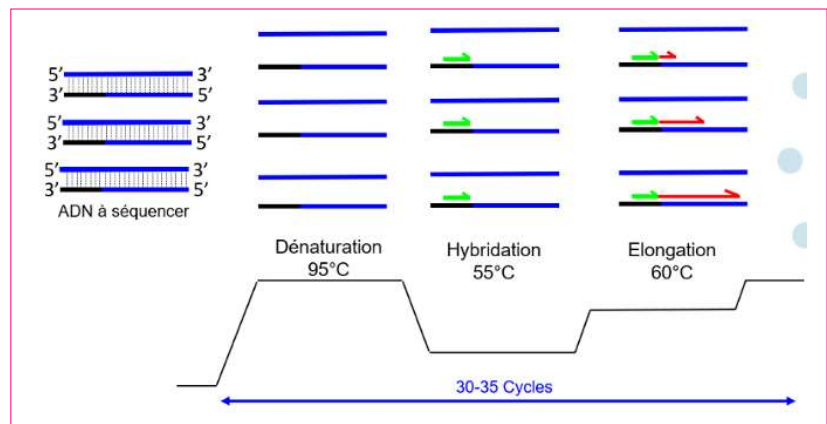
Mêmes étapes que la PCR avec le même jeu sur la température :

- **Dénaturation** à 95°

- **Hybridation** à 55°

(Un seul primer va s'hybrider)

- **Élongation** à 60°



Explications :

Dans un tube, on va mettre :

- **L'ADN à séquencer**, qui peut être le résultat d'un produit PCR
- **La Taq Polymérase**
- **L'amorce (1 seule)** qui doit être spécifique du brin que l'on veut séquencer (si on veut en séquencer deux, alors on utilise une autre amorce dans un autre tube : ici, une seule amorce)
- Le matériel nécessaire à la Taq polymérase et à la synthèse du nouveau brin, à savoir **les nucléotides (dNTPs)**
- **Les tampons**
- **Des ddNTPs (spécificité du séquençage)**

Nous allons ensuite avoir des variations successives de températures de façon à :

- 1) **Dénaturer** à 95°C notre double brin d'ADN : les liaisons hydrogènes qui lient les deux brins cassent à cette température.
- 2) **Hybrider** selon la complémentarité notre amorce sur les ADN devenus simples brins, à 55°C (ATTENTION : ici il n'y a qu'une seule amorce complémentaire à l'un des deux brins / Dans la PCR, vous aviez 2 amorces : la complémentaire et la reverse)
- 3) **Élonger** notre brin, à 60°C : le brin complémentaire au brin que l'on veut séquencer se forme.

+++ L'autre particularité pour l'étape de séquençage est que l'on va utiliser un mélange de dNTPs (désoxyribonucléotides ATCG), et de ddNTPs (didésoxyribonucléotides) +++

Lors de la synthèse, la polymérase va utiliser aléatoirement, en fonction de la complémentarité des bases, des dNTPs ou des ddNTPs.

L'utilisation des dNTPs permet la création des liaisons phosphodiester, entre un nucléotide et le suivant, et donc de continuer la synthèse du brin.

En revanche, un **ddNTP** est un dNTP mais avec un groupement H au lieu d'un groupement OH : la liaison avec le nucléotide suivant ne peut pas se faire.

À partir du moment où un ddNTP est introduit par la polymérase, la synthèse s'arrête, à la différence de l'introduction de dNTPs.

3. Cycles successifs

L'introduction d'un dNTP ou d'un ddNTP se fait **au hasard**. On va réaliser des cycles successifs de dénaturation des brins - hybridation de l'amorce - élongation du brin complémentaire.



On se retrouve à la fin avec **de très nombreux fragments d'ADN**, de tailles différentes, complémentaires de l'ADN qu'on veut séquencer.

ATTENTION : Le brin d'ADN copié par la polymérase par complémentarité des bases est le brin complémentaire. À la fin, il ne faudra pas oublier que ce n'est pas le brin que l'on veut séquencer que l'on a mais bien son brin complémentaire !

4. Histoire

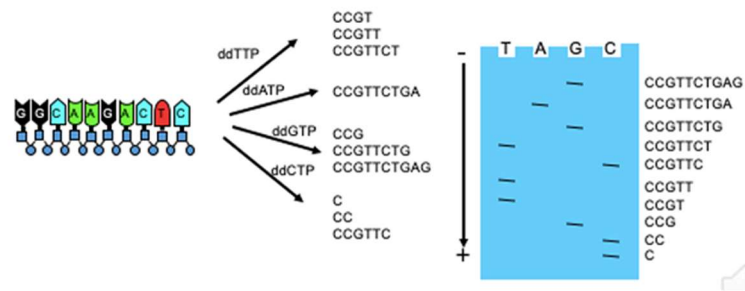
Cette partie n'a pas été évoquée en cours cette année (la prof a directement parlé de la méthode automatisée et pas de « l'historique »). Je vous la remets quand même pour votre compréhension mais ne vous attardez pas trop dessus.

Historiquement, nous utilisons la Méthode Sanger avec laquelle nous faisons **4 réactions distinctes**, avec dans chacune des réactions, un seul type de DDNTP (soit A, soit C, soit T, soit G), et tous les DNTPs.

Chacun des tubes contenaient tout ce qu'il fallait pour que la polymérase puisse séquencer

- Le fragment d'ADN
- La polymérase
- L'amorce
- Les DNTPs (ne stoppant pas la synthèse)

- Un seul type de DDNTPs (stopnant la synthèse)



On avait donc 4 tubes comme on peut le voir sur ce schéma.

Ces fragments obtenus étaient séparés sur un gel par électrophorèse, puisqu'on a vu que si l'on veut séparer l'ADN en fonction de sa taille, on réalise une migration électrophorétique.

Nous avons donc ce genre de résultats, avec sur les différentes pistes, différents morceaux d'ADN migrants plus ou moins loin en fonction de l'endroit où le DDNTP était introduit (on avait donc **pleins de brins de tailles totalement différentes**).

L'identité des nucléotides est apportée en fonction des pistes dans laquelle on va lire le fragment (T, A, C ou G) et l'ordre des nucléotides se fait par rapport à la taille des fragments.

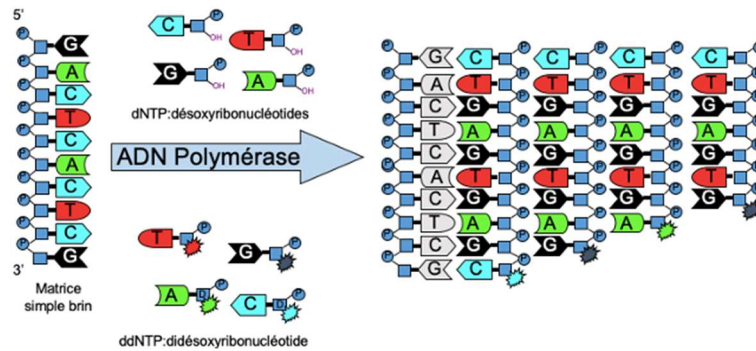
De plus, il faut savoir que **plus un morceau d'ADN est lourd, plus il a de difficultés à migrer**. Donc on lit de bas en haut (du plus léger / plus court au plus lourd / plus long).

C'est ainsi que, dans cet exemple, le **brin à séquencer étant GGCAAGACTC**, nous obtenons du C (le plus bas) CCGTTCTGAG (soit son +++ **brin complémentaire** +++).

La première synthèse à s'être arrêtée, correspondant au DDNTPs complémentaire au premier DNTP de l'ADN à séquencer. Ce qui donne le fragment qui ira le plus loin, soit le premier nucléotide complémentaire (ici la C tout en bas).

5. Méthode automatisée

La méthode Sanger a, par la suite, été **automatisée** et **simplifiée** grâce à l'utilisation de **ddNTPs fluorescents**. Chaque ddNTP est couplé à un fluorochrome de couleur différente en fonction du nucléotide. La nomenclature veut que le T soit rouge, le C bleu, le A vert et le G noir. C'est grâce à ce code couleur que l'on va pouvoir savoir quel nucléotide va être incorporé. Les 4 ddNTPs sont ajoutés dans **le même tube réactionnel**.



L'ADN polymérase synthétise, à partir d'une amorce, un brin complémentaire fidèle à la séquence d'ADN à étudier en incorporant au hasard des dNTPs ou des ddNTPs. L'incorporation d'un ddNTP stoppe la synthèse du brin complémentaire.

Les fragments obtenus seront de taille différente en fonction de si la polymérase incorpore que des dNTPs jusqu'à la synthèse complète du brin ou si elle incorpore un ddNTP dès le début (et toutes les combinaisons intermédiaires sont bien sûr possibles).

On a toutes les combinaisons possibles d'incorporation de ces ddNTPs stoppant la synthèse donc on a toutes les synthèses possiblement stoppées à n'importe quel moment de la synthèse de notre brin complémentaire par la polymérase.

Les produits synthétisés sont **séparés**, en fonction de leur taille par **migration électrophorétique**.

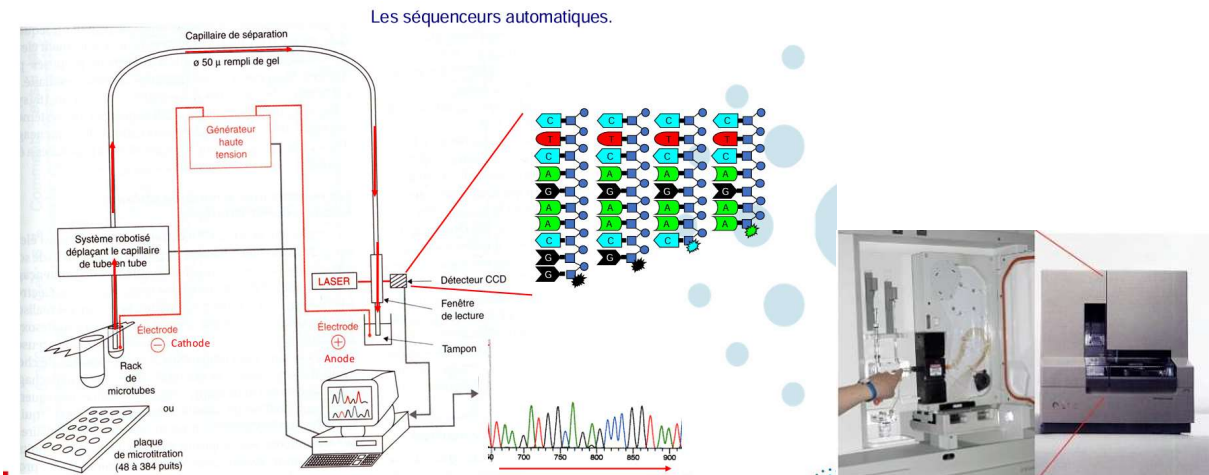
Cette migration électrophorétique et l'identification des nucléotides se fait dans des « **Séquenceurs Automatiques** » dans lesquels on retrouve des petits capillaires. Chaque capillaire correspond à un gel d'agarose et donc à une migration électrophorétique.

Les fragments d'ADN vont ensuite passer devant une **caméra** qui va lire le fluorophore qui a été incorporé.

Plus en détails, la machine trempe chacune de ses extrémités dans un tampon avec deux électrodes : d'un côté une cathode et de l'autre l'anode de façon que la migration puisse faire de la cathode vers l'anode (du - vers le +). Ce capillaire est également rempli d'un gel de polymère particulier (équivalent à de l'agarose vu précédemment). Ceci va permettre de séparer l'ADN en fonction de sa taille, il va migrer dans un champ électrophorétique.

- À gauche → tubes de séquençage, injectés dans ces capillaires (1 tube/une réaction = 1 capillaire).
- Les préparations vont migrer dans un champ électrique.
- Les fragments vont ensuite passer devant une caméra/un laser qui identifie les fragments et la couleur des fragments qui passent. Donc en fonction du ddNTP, la caméra va voir plutôt un A incorporé si c'est du vert...

- Par **traitement informatique**, on a ensuite la séquence qui apparaît sous forme d'électrophorégramme.



L'enchaînement des nucléotides est apporté par **migration électrophorétique** qui va séparer les fragments d'ADN en fonction de leur taille (+ le fragment est petit, + il migre vite et donc + il sera lu tôt par la caméra) et **l'identité des nucléotides** est apportée par **la couleur des différents ddNTPs incorporés**.

Couleur → Nom du ddNTPS
Poids sur électrophorèse, vitesse de migration → Position

Les séquenceurs automatiques sont encore beaucoup utilisés et permettent de vérifier, par exemple, que notre fragment d'ADN a la séquence qu'on attendait.

6. Retour sur l'exemple de l'achondroplasie

Après la **PCR-RFLP** (= **PCR suivi d'une digestion enzymatique**), on séquence de cette manière cette même PCR et on va lire sur l'électrophorégramme l'enchaînement de nucléotides du patient en fonction de la couleur affichée. On pourra alors connaître le nucléotide pour chacune des positions sur l'ADN et donc savoir s'il y a un mauvais nucléotide donc une mutation.

On utilise ici le séquençage pour vérifier notre digestion enzymatique : on a défini qu'on avait une mutation GA en position 1138, on vérifie en séquence qu'on a bien GA = double pic au niveau de la position 1138.

Exemple d'analyse d'un gène par PCR et séquençage Sanger : le syndrome de Wolfram

Nous allons voir maintenant une application de l'utilisation de la PCR et du séquençage, mais cette fois-ci dans le but de screener la totalité d'un gène. On ne sait pas quelle mutation est présente, donc on va séquencer **l'intégralité des exons de ce gène pour identifier le variant pathogène**. Ici on ne veut pas chercher une seule mutation comme pour l'achondroplasie. = on cherche bien une mutation mais vu qu'il en existe pleins sur ce gène on les cherche toutes donc pas qu'une.

1. Tableau clinique

Le **syndrome de Wolfram** associe d'un point de vue clinique :

- Diabète
- Surdit 
- Atrophie optique
- Troubles neurologiques

C'est une maladie   transmission **autosomique r cessive**, ce que signifie que les individus atteints ont forc ment les 2 all les mut s pour que les signes apparaissent (*les parents sont porteurs mais non atteints*).

Le **g ne WFS1** est un g ne constitu  de 8 exons, et l'ATG (codon permettant le d but de la traduction de l'ARN messager) se trouve dans le deuxi me exon du g ne. Cela signifie donc que le premier exon est non codant. La **Wolframine** (prot ine cod e par le g ne WFS1) a une fonction encore peu connue. Elle jouerait un r le dans le flux calcique.

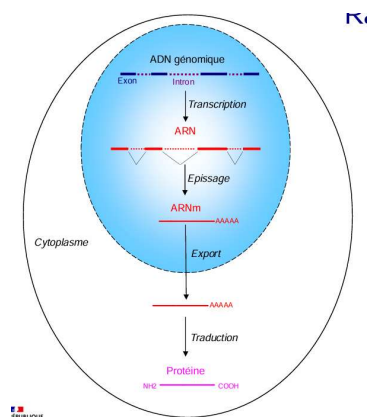
2. Rappels : organisation, expression des g nes et nomenclature

Dans le noyau se trouve l'**ADN g nomique**, organis  en **chromosomes** (contenant les g nes constitu s d'exons et d'introns).

Ces g nes vont  tre transcrits en **ARN**, et apr s les **m canismes d' pissage**, il ne restera **plus que les r gions codantes**.

Les introns seront donc  limin s pour obtenir un **ARN messager mature** avec,   leur extr mit  3', une queue Poly A.

Ces ARN messagers vont ensuite  tre export s du noyau vers le cytoplasme o  ils seront traduits en prot ines.

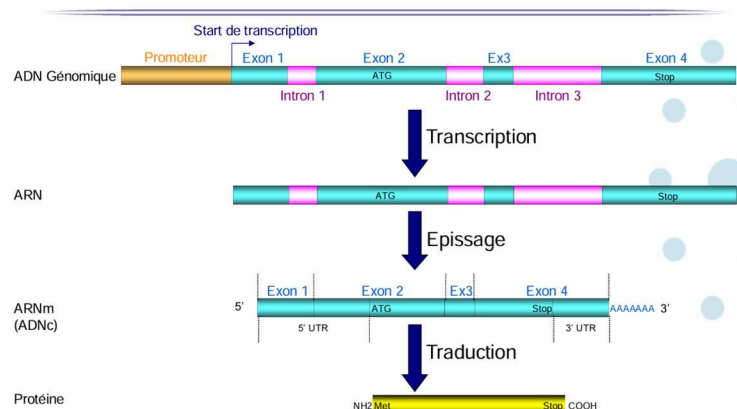


La terminologie est importante : Il faut bien distinguer ce qui correspond à la protéine = par ce que finalement c'est ce qui nous intéresse. L'unité fonctionnelle du gène c'est la protéine, un gène en tant que tel, tout seul, si on ne sait pas à quoi correspond la protéine pour lequel il code, ne va pas nous permettre d'en déduire grand-chose.

Si on regarde de plus près l'organisation et l'expression des gènes : on représente sur la 1ère ligne l'organisation génomique avec en 5' **la région promotrice**.

Ce 5' est le lieu où vont se fixer les protéines qui vont permettre le démarrage de la transcription de ce gène en ARN. Le **start de transcription** définit le cadre de lecture. On retrouvera ainsi **les régions codantes (exons)**, séparées par les **régions non codantes (introns)**.

L'organisation et l'expression des gènes



/!\ Attention : la traduction ne commence pas forcément aux premiers nucléotides du premier exon /!

Elle commence à partir du **codon ATG** (sur **l'exon 2** dans cet exemple : le premier n'est donc pas traduit). Ceci sous-entend que si le codon ATG se trouve dans l'exon 2, tout ce qui est en amont de cet exon ne sera pas traduit, même si la séquence est présente dans l'ARN messager.

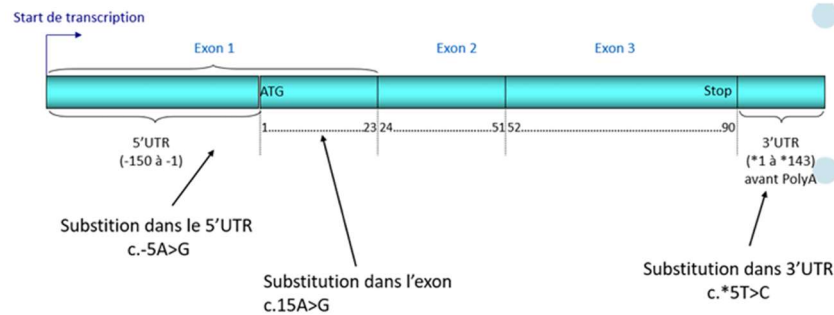
Un exon est sur l'ARNm mais n'est pas forcément codant pour la protéine.

La région en amont du codon ATG (codon START) est appelée 5' UTR (non codante) et la région en aval du codon STOP est appelée 3' UTR, cette région est également non codante.

Les **régions 5' UTR et 3' UTR** sont donc des **régions transcrites mais non codantes donc non traduites**.

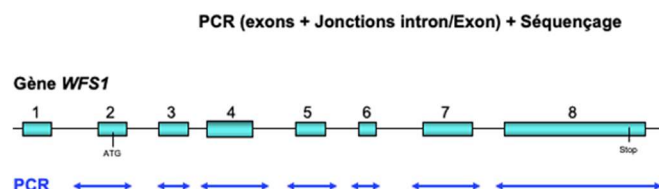
En général ces régions jouent un rôle dans la **stabilité de l'ARNm**, son **exportation** ou autre **(elles ne servent pas à rien)**.

La protéine commence au niveau de l'ATG (= méthionine, premier codon codant = c.1) et se termine au niveau du codon stop.



Les substitutions situées en 5' UTR seront numérotées « c. - » en partant de l'ATG et en remontant (donc de droite à gauche) et celles en 3' UTR seront numérotées « c. * » en partant du codon stop (donc de gauche à droite)

3. Recherche de mutations dans un gène



Si on reprend notre exemple de départ, à savoir le **screening du gène WFS1**, on remarque que ce gène est organisé en **8 exons**, et l'ATG se situe dans l'exon 2. L'exon 1 est donc transcrit mais non traduit.

Pour diagnostiquer cette pathologie, il faut **identifier le variant nucléotidique responsable d'un dysfonctionnement de la Wolframine**, qui va modifier la fonction de cette protéine.

De ce fait, on va particulièrement s'intéresser **aux régions codantes**. On va **amplifier les exons codants** parce **qu'on sait facilement interpréter des variants nucléotidiques situés dans les exons**. On va facilement pouvoir savoir si tel ou tel variant nucléotidiques modifie un acide aminé et la conséquence de cette modification sur la protéine.

On va réaliser **7 PCR** pour amplifier les régions codantes : les PCR seront faites sur les exons 2 / 3 / 4 / 5 / 6 / 7 / 8.

On n'amplifie pas l'exon 1 un car il est non codant. S'il y avait une mutation dans cet exon, il serait **très difficile de l'interpréter** et de savoir quelles conséquences seraient engendrées puisqu'il n'est pas exprimé. Aujourd'hui, **on n'a pas encore les outils pour prédire ce qu'il va se passer lorsque le variant nucléotidique est dans une région non codante**.

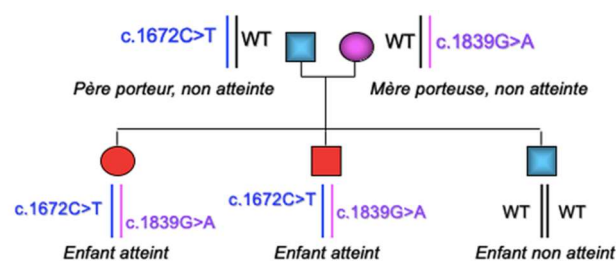
Malgré tout, on sait qu'un variant sur les **tous premiers nucléotides de l'intron** a généralement un effet sur **l'épissage** (et donc sur la protéine finale) donc on prend des

amorces pour la PCR qui amplifient l'exon mais aussi **une partie des introns qui l'encadrent** (à 50 nucléotides de cette jonction intro/exon).

RECAP

On amplifie : les exons codants (on interprète facilement les changements nucléotidiques dans les parties codantes) + une partie des introns qui les encadrent (ont potentiellement des effets sur l'épissage)

On n'amplifie PAS : les exons non codants (car on n'a pas encore les outils pour prédire ce qu'il va se passer lorsque le variant nucléotidique est dans une région non codante)



➤ Diagnostic de Syndrome de Wolfram

Prenons l'exemple de cette famille : on amplifie les 7 exons par PCR, on les séquence, et on identifie chez le père un variant en position 1672 (qui modifie un C en T) à l'état hétérozygote.

Chez la mère, un deuxième variant apparaît, toujours sur le gène WFS1, mais en position 1839.

On voit donc que les deux enfants atteints ont hérité des deux gènes mutés : celui du père et celui de la mère.

En revanche le troisième enfant a hérité des gènes non mutés.

Le résultat permet de poser le diagnostic de **Syndrome de Wolfram**. Cette pathologie étant **autosomique récessive**, les parents ne présentent donc aucun signe. En revanche les deux enfants atteints développeront la maladie cliniquement car ils n'ont aucun de leurs deux allèles qui peut donner la bonne protéine (les deux sont mutés).

Là encore on était dans un « cas simple », on regarde UN gène donc ce n'est pas très difficile à séquencer, on y arrive facilement. Mais pour d'autres pathologies, il y a plusieurs gènes à regarder. Jusqu'à il a une dizaine d'année, on séquencer avec la méthode Sanger les gènes les uns après les autres mais est apparu une autre méthode :

Séquençage Haut Débit ou NGS (Next Generation DNA Sequencing)

Aujourd'hui le **NGS** est de plus en plus utilisé car c'est une **technique extrêmement puissante** qui coûte **de moins en moins cher** et est **de plus en plus accessible**.

Rappel historique :

- Le « Séquençage Manuel » (dit Sanger) inventé en 1977
- Le « Séquençage Automatique » avec ses DDNTPs fluorescents, datant de 1993
- Depuis 2007, on a le « Séquençage Haut Débit » ou NGS qui a révolutionné la biologie moléculaire

En **2001**, 95% du génome humain avait été séquencé. C'est un **énorme projet mondial** débuté en **1989** et achevé en **2003**, réalisé à l'aide de **séquenceurs capillaires**, car initié bien avant la fabrication des machines NGS.

Le **génome humain** correspond à environ **3 milliards de pdb**. On y trouve **30 000 gènes** et les **régions codantes (exons)** ne représentent que **1 à 2% du génome**.

Il nous reste donc encore beaucoup de choses à comprendre car les 98% restants ne servent certainement pas à rien mais aujourd'hui on n'est pas capable de comprendre exactement à quoi servent toutes ces régions non codantes. On aurait d'ailleurs pu les trouver moins intéressantes car pas codantes. On s'attendait bien sûr en séquençant le génome à retrouver plus de régions codantes que 1 ou 2 %.

Les régions non codantes doivent certainement avoir un rôle au niveau de la régulation de l'expression des gènes voir dans d'autres mécanismes qu'on n'a pas encore compris. La difficulté aujourd'hui, quand on fait un séquençage complet du génome pour retrouver le variant responsable, n'est donc plus au niveau du séquençage mais dans l'interprétation des variants.

Pour réaliser le séquençage complet d'un génome humain il aurait fallu :

- **12 ans** avec **1000 machines PCR** avec la méthode Sanger en 1980
- **1 an** avec **120 machines** avec le **séquenceur automatisé** dans les années 2000
- Il faut **1 semaine** (voire moins : **2-3 jours**) avec **1 machine**, avec le **NGS**, aujourd'hui pour réaliser un séquençage du génome humain de plusieurs dizaines d'individus en même temps.

Les capacités de séquençage ont donc augmenté de manière exponentielle ces dernières années.

Définition : Le séquençage au débit est un **séquençage massif** en parallèle de molécules d'ADN individuellement séparées et amplifiées sous forme de clones ou de molécules uniques. *= on séquence en même temps beaucoup de molécules d'ADN, vous visualiserez mieux à la fin du cours à quoi correspond le NGS. Je vous conseille de faire une première lecture du cours avec comme but de comprendre et vous apprendrez ensuite.*

*La définition peut paraître un peu compliquée comme ça mais une fois que vous aurez vu les différentes étapes vous comprendrez que ce qu'on va faire c'est : +++ **fragmenter notre ADN, l'isoler et le séquencer** +++.*

On fait un **NGS** lorsqu'on veut séquencer énormément de gènes voire l'exome ou le génome au complet.

DEF = EXOME = ensemble des exons // GENOME = ensemble des gènes (donc exons mais aussi introns)

Le NGS sera utilisé pour rechercher les mutations dans les maladies génétiques qui peuvent être dues à un très grand nombre de gènes.

On pourrait utiliser cette technique pour diagnostiquer l'achondroplasie, mais **elle ne sera pas en adéquation avec l'utilité de cette dernière** car l'achondroplasie est une maladie qui atteint un gène bien précis. (*= le NGS donnerait bien sûr le diagnostic donc en théorie on pourrait l'utiliser mais on ne le fait pas car ce serait inutile de séquencer tout le génome alors qu'on recherche un seul gène, c'est comme si vous lisiez tout le dictionnaire alors que vous cherchez un mot en p)*

+++ On fera donc une PCR-RFLP (= PCR suivie d'une digestion enzymatique), plutôt qu'un NGS pour diagnostiquer l'achondroplasie +++

Les **outils de biologie moléculaire** utilisés pour réaliser un NGS sont pratiquement les mêmes que pour les techniques vues auparavant : polymérase, ligase etc....

+++ Ce qui a permis l'arrivée du séquençage massif c'est l'amélioration des technologies, la miniaturisation des systèmes, et la bio-informatique qui s'est développé et se développe encore aujourd'hui +++

Actuellement il existe **deux plateformes** (2 sociétés) qui commercialisent deux technologies différentes au niveau des étapes d'amplification et de séquençage les sociétés :

- **Illumina** -> celui de plus en plus fréquent
- **ThermoFisher**

2 plate-formes:

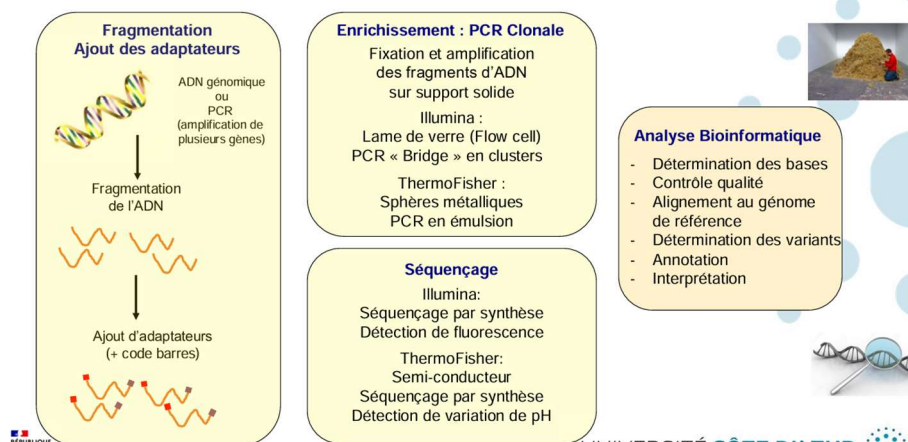
illumina®
<https://www.youtube.com/watch?v=fCd6B5HRaZ8>

ThermoFisher **ion torrent**
 S C I E N T I F I C
<https://www.youtube.com/watch?v=zBPKj0mMcDg>

Vous avez les liens de vidéos youtube expliquant toutes les étapes de ces techniques sur le diapo de la prof qui est sur moodle.

1. Etapes globales du NGS

Les étapes du NGS



Les étapes sont les mêmes pour les deux systèmes : ce qu'il faut que vous compreniez c'est comment on fait un NGS et quelles sont les grandes étapes. Je ne vous demande pas de retenir les détails de ces étapes qu'on présentera après.

- 1) La préparation des échantillons/librairies → fragmentation de l'ADN et ajout des adaptateurs et des barres-codes (BC) à la séquence d'ADN génomique.
- 2) La PCR clonale → amplification de chacune de ces molécules d'ADN génomique d'intérêt.
- 3) Le séquençage de ces molécules d'ADN
- 4) L'analyse bio-informatique → nécessaire au traitement des données générées par les séquenceurs de façon à identifier les variants et à rechercher d'éventuelles mutations (par ex).

1ère étape : identique quelle que soit la plateforme utilisée (illumina ou ThermoFischer). Elle consiste à **fragmenter notre ADN** (ADN génomique ou des fragments PCR)

Les techniques de NGS permettent aujourd'hui de séquencer des fragments d'environ 200 à 400 pb en fonction des kits et des plateformes utilisés. A ces petits fragments d'ADN

vont être ajoutés des adaptateurs (= primers) et des BC, c'est-à-dire des oligonucléotides dont la séquence est connue.

Ils vont permettre :

- **L'amplification PCR pour les adaptateurs** (identiques pour toutes les séquences et donc tous les patients)
- **L'identification des patients pour les BC** (donc différents/spécifiques pour chaque patient)

En effet ces **BC** ont une séquence bien déterminée et chaque BC va être attribué à un patient, soit +++ **une séquence de BC = un patient** +++

Dans les étapes de séquençage on va donc pouvoir mélanger nos patients dans un même run, dans une même réaction. Lors de l'analyse bio-informatique, ces séquences vont être triées grâce à ces BC ce qui va permettre d'identifier et d'attribuer chacune des séquences à chacun des patients.

+++ Les adaptateurs sont ajoutés aux extrémités 5' et 3' des fragments d'ADN et ils servent à ce que toutes les extrémités 5' et 3' soient identiques et donc qu'on puisse amplifier l'ensemble de nos fragments d'ADN avec uniquement un seul couple de primer (1 primer sens + 1 primer reverse) +++

2e étape : enrichissement de nos fragments d'ADN par PCR clonale, (car on part d'un seul fragment et on l'amplifie x fois) il va y avoir **une étape de fixation** et **d'amplification** de chacun des fragments d'ADN sur un support solide, différents selon les plateformes :

➤ Technologie Illumina = **Lame de verre**

➤ Technologie Thermo Fischer = **sphères métalliques**

L'idée c'est de séparer tous les fragments d'ADN de manière à pouvoir les amplifier de manière clonale (à en avoir des copies, des clones). En imaginant qu'on le fait pour nos milliards de fragments en même temps.

Ces supports solides (lame de verre, sphère métallique) vont être **séquencés individuellement**. On va retrouver un peu ce qu'on avait avec le séquençage Sanger mais avec des **détections** qui vont être **différentes en fonction des plateformes** :

➤ Illumina : **détection de fluorescence** (ptit mémo : illumina => lumière ça fait penser à la fluorescence)

➤ ThermoFischer : **des variations de pH**

« Peu importe ce qu'il faut que vous compreniez c'est +++ qu'on fragmente, on isole, on amplifie et on séquence +++ »

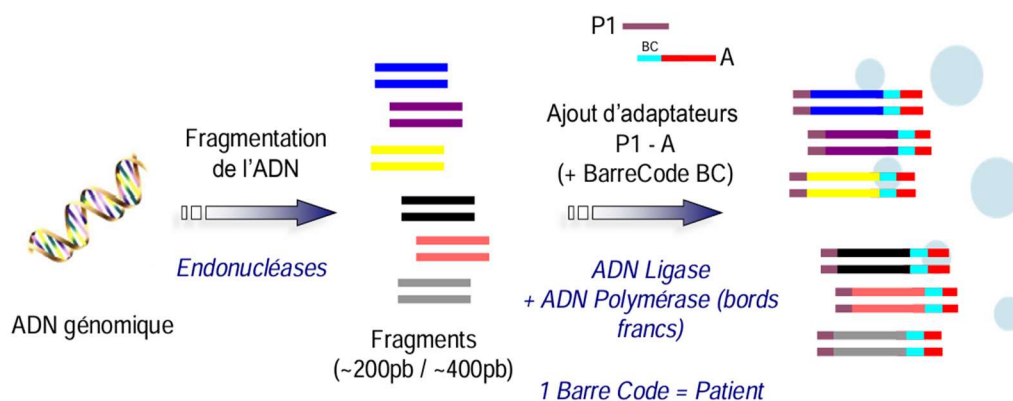
Dernière étape : commune et extrêmement importante, **l'analyse bio-informatique**, puisqu'on va générer des données massives et il va falloir ensuite transformer ces signaux lumineux (détection fluorescence → Illumina) ou transformer ces variations de pH (ThermoFischer) en « nucléotides », en données de séquences.

C'est cette partie qui est la grosse difficulté, elle est compliquée car il va falloir beaucoup de systèmes informatiques avec des algorithmes de plus en plus puissants (car on arrive à séquencer de plus en plus). Ce sont des fichiers de plus en plus gros qui engendrent des difficultés de stockage de données. Et c'est surtout des problèmes d'annotation et d'interprétation (on reverra ça à la fin).

Vous avez séquencé vos 3 milliards de nucléotides sur votre génome, vous cherchez la faute d'orthographe, le nucléotide variant pathogène parmi tous les polymorphismes (car on a tous des variants nucléotidiques d'un individu à l'autre mais qui n'entraînent aucun caractère pathogène, qui ne donnent pas de maladies → yeux marrons/verts, cheveux bruns ou blonds ... = caractères physiologiques).

2. Préparation des échantillons (= préparation des librairies) *Je sais que ça peut paraître compliqué mais essayez de bien visualiser avec les schémas !!*

On va voir **comment on peut fragmenter et ajouter les adaptateurs** sur ces échantillons (ADN génomique). On est en mesure de récupérer uniquement les régions qui nous intéressent (des régions exoniques en général, pas toujours les mêmes) grâce à un système/technique de capture. On capture/extrait du milieu réactionnel uniquement les régions d'ADN génomique qui nous intéressent.



Cet ADN génomique va être tout d'abord coupé en fragments de 200 à 400 pb grâce à des **endonucléases** qui ont comme propriété de couper l'ADN double brin en son milieu. On fragmente avec plusieurs endonucléases qui vont générer des fragments aléatoirement sur l'ADN génomique.

Ensuite à ces fragments d'ADN on va rajouter **des adaptateurs** (P1 et A) **et des barres codes** (BC).

Pour tout cela on retrouve **les outils vus précédemment** combinés différemment :

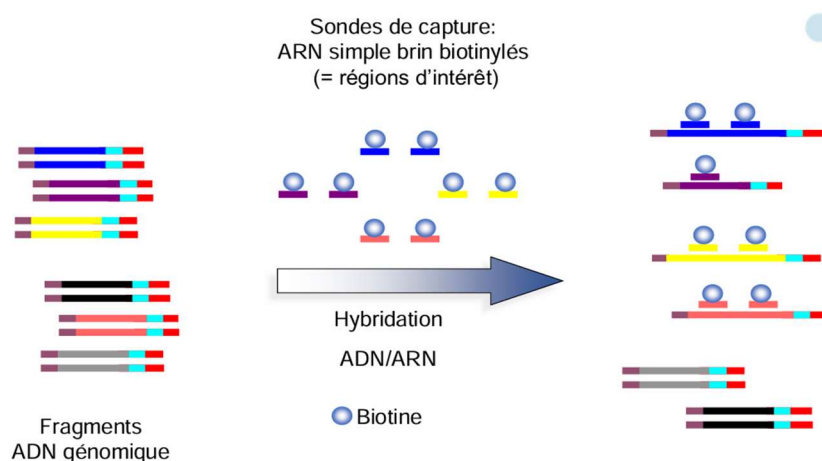
- Des **ADN ligases** pour lier les fragments d'ADN aux adaptateurs et aux barres codes (qui sont aussi des séquences d'ADN)
- Des **ADN polymérases** → en fonction de la digestion réalisée par les endonucléases, de leur site de restriction, les extrémités ne sont pas forcément toutes à bords francs.

Grâce à ces outils on va pouvoir couper et liguer nos fragments d'ADN qui auront alors **les mêmes extrémités 5'** et **les mêmes extrémités 3'**.

+++ Toutes les extrémités 5' sont identiques, toutes les extrémités 3' sont identiques mais les extrémités 5' et 3' sont différentes entre elles +++ « c'est ça que vous devez savoir ! On n'ira pas vous demander si l'adaptateur en 5' c'est le P1 ou le A »

Maintenant qu'on a obtenu nos fragments d'ADN munis d'un BC et des adaptateurs, **on peut mélanger plusieurs patients**, car ils seront reconnus par la séquence des BC. On va alors chercher à récupérer les séquences d'ADN qui nous intéressent. C'est aussi pour ça qu'on est capable de séquencer énormément, on peut mélanger nos patients ! Par contre attention il ne faut pas les mélanger avant d'avoir mis le BC !

« Peu importe que le BC soit en 5' ou en 3', peu importe comment il s'appelle mais retenez qu'à ce moment-là dans le NGS vous pouvez mélanger vos patients car l'ADN a été barrencodé. »

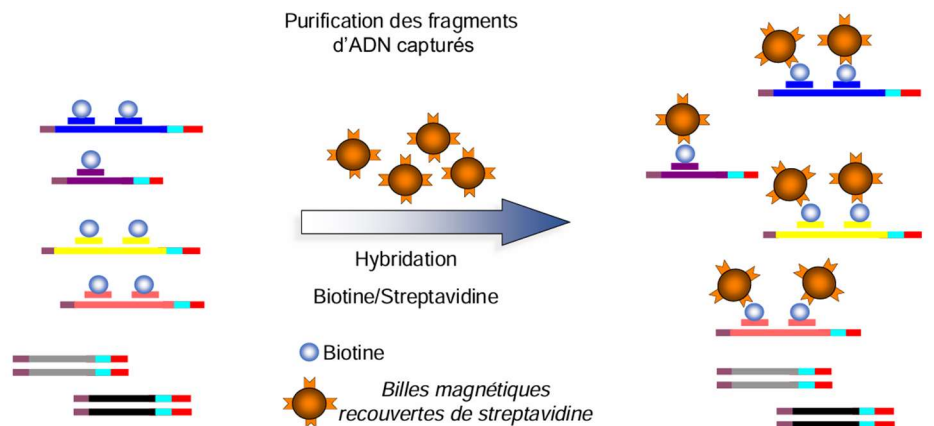


On ajoute à ces fragments **des sondes de capture** = sondes ARN simple brin. Ce sont des petits fragments d'ARN complémentaires de la séquence d'ADN que l'on veut séquencer. On a un mélange de différentes sondes ARN spécifiques aux fragments d'ADN ciblés.

Ces sondes sont **biotinyllées** = de la biotine est fixée sur ce fragment d'ARN.

Comme pour la PCR, **on dénature** = on casse les liaisons hydrogènes qui maintiennent les deux brins d'ADN entre eux. **Grâce à des variations de température**, on aura la formation

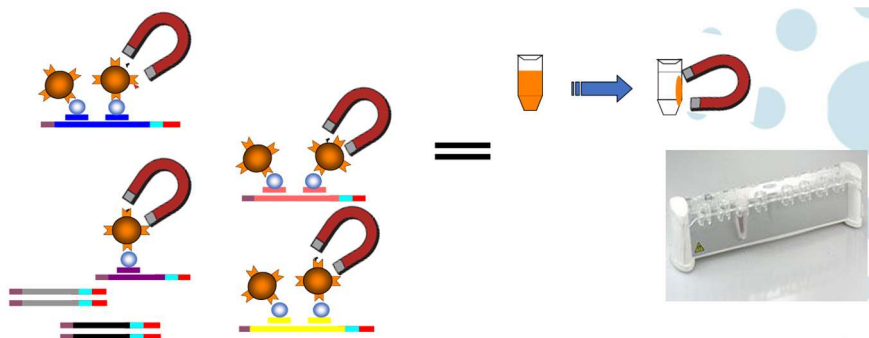
d'hybrides ADN/ARN, avec ces ARN biotinylés qui vont se fixer sur les fragments d'ADN qui leur sont spécifiques et qui nous intéressent.



Nous allons ensuite **purifier ces fragments d'ADN capturés** en ajoutant **des sondes = des billes magnétiques recouvertes de streptavidine**.

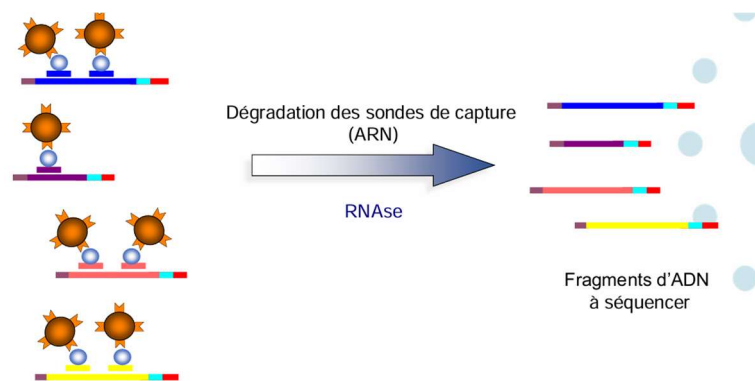
L'interaction « **biotine-streptavidine** » est une interaction très forte (protéine-protéine) qui est utilisée dans beaucoup de techniques différentes. Lorsqu'on rajoute ces billes recouvertes de streptavidine, elles vont se fixer sur la biotine donc sur les hybrides ADN/ARN.

Tous les fragments d'ADN reconnus vont être piégés sur ces billes magnétiques



Les billes recouvertes de streptavidine étant **des billes magnétiques**, si on ajoute des **aimants sur le bord du tube**, ces billes vont être attirées sur les côtés, se coller sur le bord du tube → coloration marron sur le bord du tube.

On va pouvoir réaliser des lavages (ajouter des liquides, des tampons), aspirer, retirer le surnageant de notre tube qui ne nous intéresse pas de façon à nettoyer et ne récupérer que les billes qui auront fixées ces sondes ARN elles-mêmes hybridées sur nos régions d'intérêt = fragments d'ADN qu'on veut séquencer.



Une fois qu'on a **purifié les régions d'intérêt**, on va dégrader les sondes ARN hybridées à l'ADN d'intérêt.

Pour cela on va **ajouter des RNAses**, enzymes qui dégradent l'ARN hybridé dans le complexe ADN/ARN, **pour qu'il ne reste plus que l'ADN à séquencer**.

On obtient donc nos régions d'ADN ciblées qui sont barre codées et avec leurs adaptateurs.

On rappelle que quelle que soit la plateforme utilisée, la préparation des échantillons avec cette étape de capture est identique.

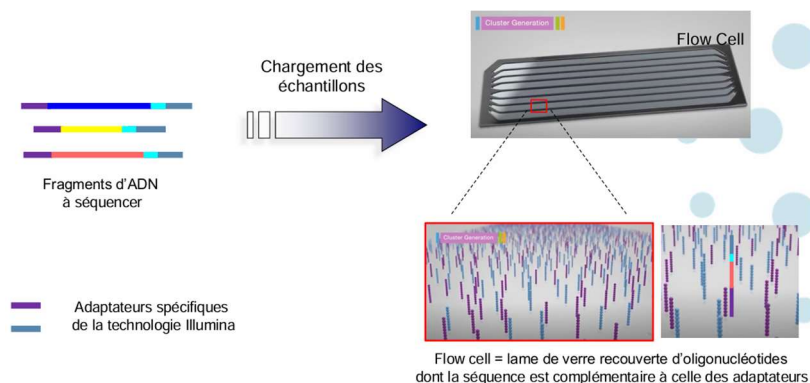
Surtout n'essayez d'apprendre par cœur cette partie ou les parties qui suivent sans ++ comprendre ++ Là tout ce que l'on vient de voir c'est beaucoup d'explications pour quelque chose de simple au final !

- On a TOUS nos ADN fragmenté mais on ne veut que ceux qui nous intéressent
- Comment on peut les récupérer ? Avec des sondes ARN (car on peut facilement prendre des ARN avec la séquence qu'on veut et hop ça s'hybride)
- Oui mais après faut les récupérer ces fragments ARN/ADN !?
- Alors on n'utilise pas que des ARN mais des ARN biotinylés !
- PK de la biotine ? Par ce qu'elle a une très forte interaction avec la streptavidine
- Alors on recouvre des billes magnétiques de streptavidine
- A la fin l'aimant permet de récupérer les billes magnétiques, recouvertes de streptavidine ce qui récupère donc l'ARN biotinylé, ce qui récupère donc nos fragments d'ADN qu'on voulait séquencer
- Ensuite on lave tout et on dégrade les ARN (avec RNase) car on en a plus besoin !

RECAP : Aimant > Billes magnétiques > Streptavidine > Biotine > ARN > ADN (qu'on veut séquencer)

3. Enrichissement par PCR clonale avec Illumina (PCR en pont, clusters)

Pour la technologie **Illumina**, l'enrichissement/la PCR clonale est réalisée sur **une lame de verre**, également appelée **Flow Cell**.



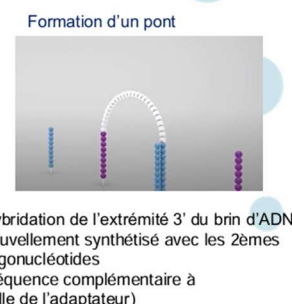
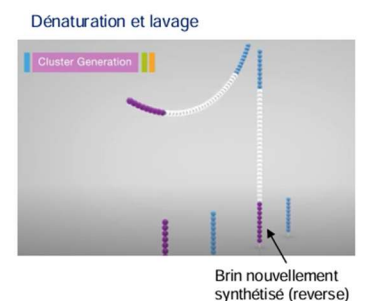
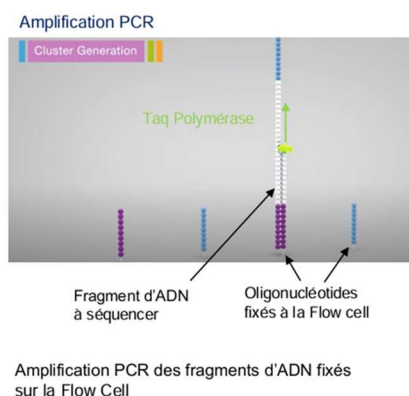
Cette lame est recouverte de **2 types d'oligonucléotides** (= amorces) fixés :

- Oligonucléotides spécifiques des adaptateurs ajoutés en 5' = leur séquence est complémentaire aux adaptateurs ajoutés en 5'
- Oligonucléotides spécifiques des adaptateurs placés en 3' = avec une séquence complémentaire aux adaptateurs placés en 3'.

Les échantillons préparés précédemment sont déposés sur cette lame de verre et, par complémentarité des bases, les fragments d'ADN préparés vont s'hybrider sur les différents oligonucléotides.

Il faut donc imaginer que l'ensemble de nos échantillons vont être hybridés à différents endroits, sur cette lame de verre de sorte à **générer des clusters**.

À partir de ces fragments fixés sur cette lame de verre, on réalise une amplification de chacun de ces fragments par PCR.



<https://www.youtube.com/watch?v=fCd6B5HRaZ8>

On va donc retrouver tout ce qu'il faut pour amplifier un fragment d'ADN et notamment une **Taq polymérase**.

Elle va synthétiser de 5' en 3' un brin complémentaire au brin hybridé sur notre lame de verre.

On a ensuite une étape de **dénaturation et de lavage** qui va permettre de libérer, de notre lame, le premier fragment qui a été hybridé sur celle-ci. Les fragments nouvellement synthétisés, eux, ne sont pas enlevés par les lavages puisque l'oligonucléotide qui a permis d'amorcer cette synthèse est fixé sur la lame de verre. Par contre, le fragment que nous avons ajouté et qui provenait de notre échantillon va être enlevé de cette lame lors du lavage son amorce n'est pas fixée de manière covalente sur la lame de verre (il était simplement hybridé par complémentarité des bases à l'aide de l'oligonucléotide).

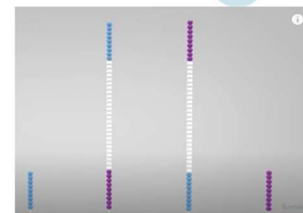
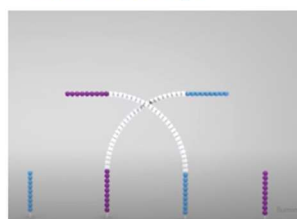
Seul le brin nouvellement synthétisé est attaché à la lame de verre puisque celui-ci a été formé à partir des oligonucléotides fixés sur cette lame.

Amplification PCR

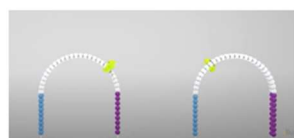
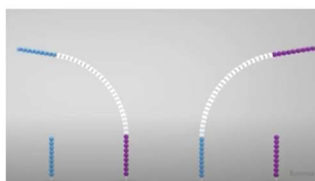


Synthèse du brin complémentaire
= brin sens

Dénaturation et lavage

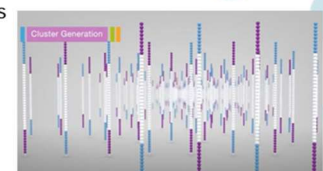


⇒ Seulement les brins nouvellement synthétisés restent fixés sur la Flow Cell



Formation de millions de clusters

Cycles successifs d'amplification PCR



On a ensuite **une étape de ré amplification de ces fragments isolés = PCR bridge ou PCR en ponts**.

Lors de cette ré amplification, on va avoir la **formation de ponts** puisque l'adaptateur non fixé de notre fragment s'hybride à l'oligonucléotide complémentaire à cette séquence, localisé juste à côté. *(Regardez sur les diapos : le premier pont qu'on voit avec les oligonucléotides en bleu)*

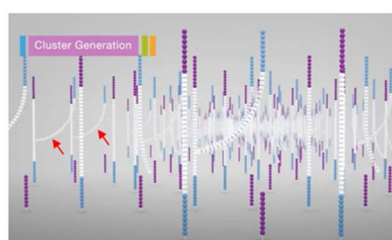
C'est pour cette raison qu'on parle de PCR en pont : notre molécule d'ADN nouvellement synthétisée forme des ponts grâce à l'hybridation de chacun de ses adaptateurs en 5' et en 3'.

Suite à la formation de ce pont, la Taq polymérase va pouvoir synthétiser le brin complémentaire de 5' en 3'. On a ensuite de nouveau une étape de dénaturation et de lavage qui va permettre d'obtenir deux fragments d'ADN simple brin, fixés sur notre lame de verre. Nous obtenons donc 2 nouvelles molécules d'ADN nouvellement synthétisées (un brin sens et un brin reverse).

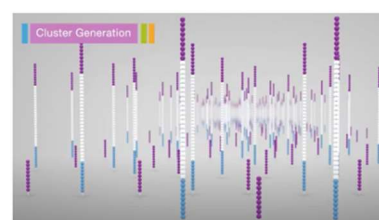
Ces étapes de formation de ponts vont se répéter, ainsi que la formation du brin complémentaire grâce à la Taq polymérase. Cette répétition se fait **jusqu'à la formation d'un grand nombre de brins** : c'est la **formation d'un cluster d'amplification** (formation de groupes d'amplification de produits PCR).

+++ Sur ce support solide, à la fin des cycles successifs de PCR, on obtient des millions de clusters qui correspondent initialement à un seul fragment d'ADN +++

À la fin de nos différents cycles d'amplification, on a une étape qui va permettre de cliver les brins reverse, qui sont les brins complémentaires du brin que nous avons mis initialement sur la lame. On ne garde donc que les brins sens, c'est-à-dire les brins qui ont la même séquence que notre ADN original et initialement hybridés sur la lame. Ce sont ces brins sens que l'on va séquencer. « peu importe reverse ou non, retenir qu'on a amplifié notre ADN »



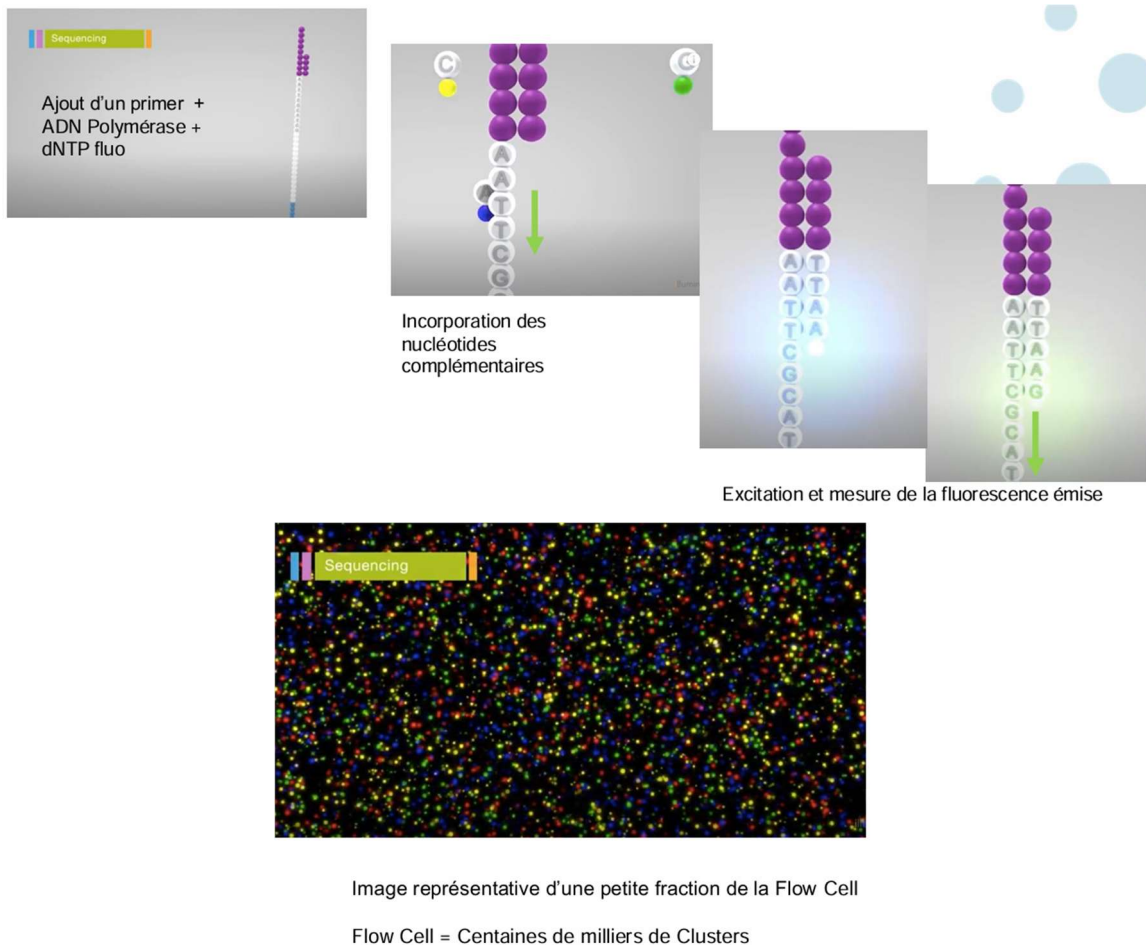
Clivage des brins reverse



Elimination des brins reverse par lavage
Conservation des brins sens uniquement

Je sais que c'est dur à se représenter cette partie, regardez bien les diapos et ne vous prenez pas trop la tête sur les détails ! Globalement, on a isolé chaque fragment d'ADN en le fixant à une lame de verre et ensuite on a amplifié = formation de cluster. Retenez déjà ça c'est très bien !

4. Séquençage avec la plateforme Illumina



Le séquençage est réalisé directement sur notre lame de verre. Pour cela on rajoute **un primer**, qui correspond à un oligonucléotide complémentaire de l'adaptateur (violet sur la diapo). On ajoute une **polymérase** pour la synthèse du brin de 5' en 3', et des **nucléotides fluorescents** (= dNTPs fluorescents). Lorsque la polymérase introduit un nucléotide par complémentarité des bases, une fluorescence d'une couleur différente selon le dNTP est émise.

L'automate va lire/mesurer en direct la fluorescence émise par le dNTP ajouté sur chacun des fragments et dans les différents clusters au cours de la synthèse. L'appareil lit la couleur émise et donc on saura par déduction quel nucléotide est incorporé. On obtient ce type d'image. Nous avons ici une toute petite partie d'une Flow Cell. On voit bien les signaux lumineux émis chaque fois qu'on incorpore un nouveau nucléotide. Le système arrive à les lire au fur et à mesure et à les transformer en données de séquence.

On parle de **Séquençage Massif**, puisque chaque cluster peut représenter un tube de séquençage. Ils sont séquencés et lus en parallèle.

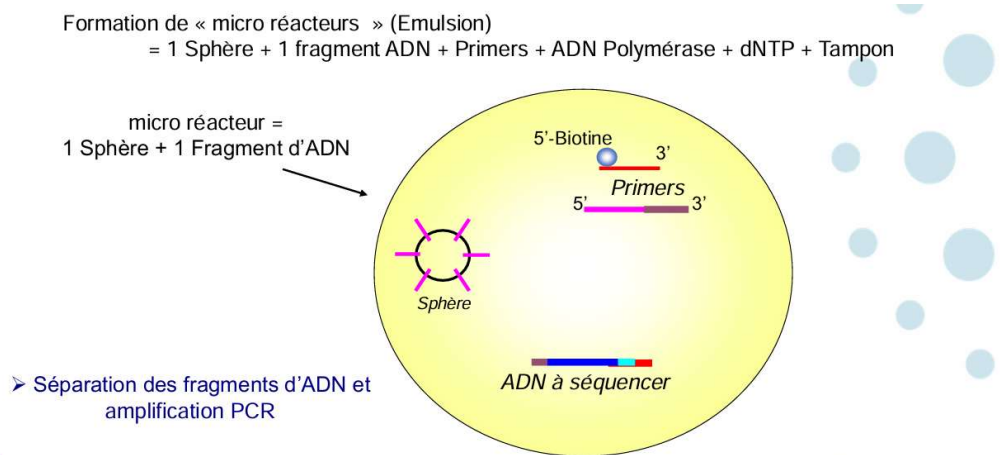
C'est fini pour illumina mtn, on va voir la même chose mais chez ThermoFischer (oui vous pouvez prendre une pause entre les deux c'est mérité !!)

5. Enrichissement par PCR clonale avec ThermoFischer

Avec les techniques d'amplification clonale sur la plateforme **ThermoFischer**, l'ADN capturé n'est pas fixé sur une lame de verre, mais sur **une sphère métallique mais le principe reste le même (amplification clonale puis séquençage)**.

Contrairement à Illumina, l'amplification clonale ne va pas se faire par cluster mais dans un **microréacteur créé par un système d'émulsion**.

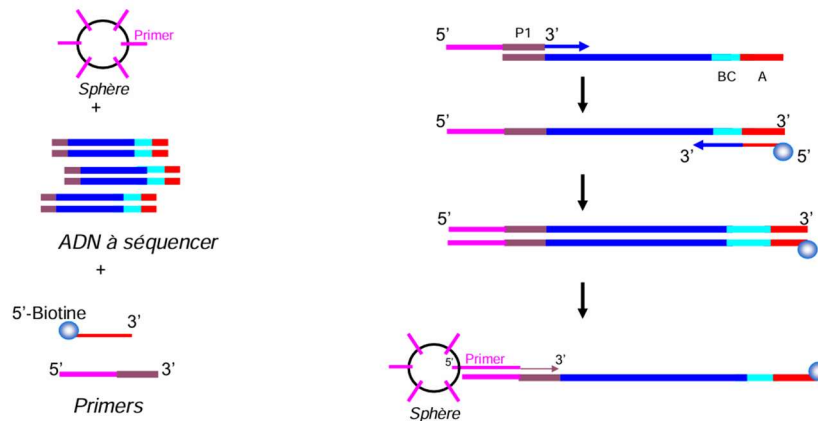
Lorsqu'on mélange une substance grasse (ex : huile) avec une solution aqueuse, il y a la formation de microgouttelettes. Ce sont ces microgouttelettes qui vont former nos microréacteurs.



Le système est fait de sorte que chaque microréacteur soit l'équivalent d'un tube contenant :

- Une sphère avec des oligonucléotides spécifiques et complémentaires des adaptateurs de nos fragments d'ADN à séquençer
- Un fragment d'ADN à séquençer
- Deux primers : un biotynilé, l'autre non
- L'ADN polymérase
- Les dNTPs
- Un tampon (pour que la Taq Polymérase fonctionne correctement).

On réalise ensuite la PCR clonale avec les mêmes étapes de dénaturation, hybridation, élongation que la PCR classique.

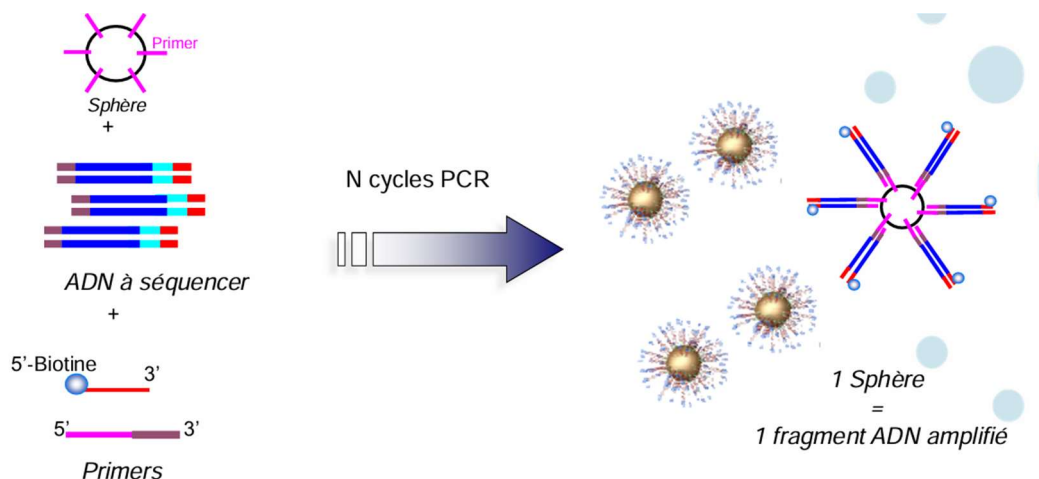


Lors du **premier cycle PCR**, le 1er primeur s'amorce au niveau de l'extrémité 3' de notre fragment d'ADN. On a l'élongation du brin complémentaire à notre fragment d'ADN d'intérêt de 5' en 3'. Puis **on dénature**. Lors du deuxième cycle, c'est le deuxième primer qui s'hybride en complémentarité de l'adaptateur. Cette fois-ci, le primer est biotynilé en 5'.

Le brin complémentaire est généré grâce à la polymérase : **on a alors un double brin**.

Le fait d'avoir synthétisé ce brin reverse (lors du 2e cycle) permet à ce nouveau brin de pouvoir se fixer toujours par complémentarité des bases sur le primer localisé sur une petite sphère magnétique. C'est de cette manière qu'on peut **piéger nos fragments d'amplification sur un support solide**.

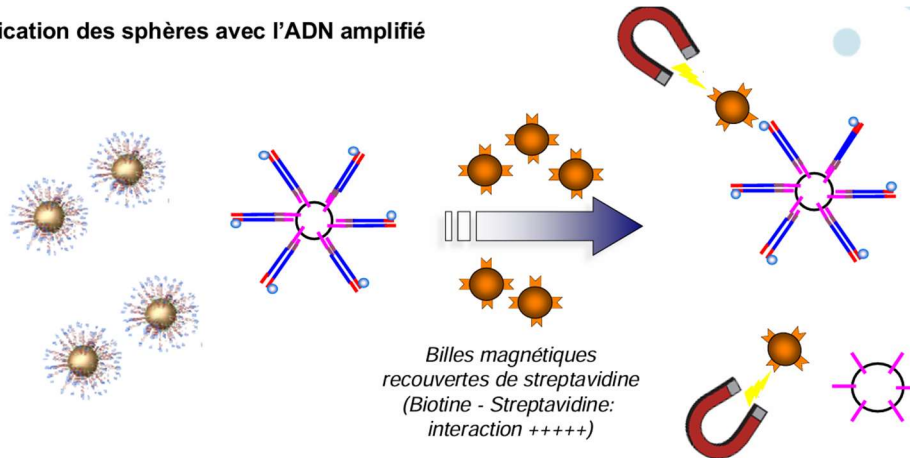
À chaque étape, nous avons amplification de chacun des fragments et fixation de ces fragments sur une sphère magnétique.



Après n cycles de PCR, on se retrouve avec cette sphère recouverte d'une multitude de fragments d'ADN, correspondant à notre fragment initial présent dans notre microréacteur qu'on a amplifié.

On parle de **PCR clonale** parce que c'est réellement un clone d'une molécule d'ADN amplifiée dans chaque microréacteur avec, sur chaque sphère différente, des produits PCR uniques → 1 sphère contient de multiples copies identiques d'un même fragment d'ADN (sphères « pures ») mais d'une sphère à l'autre les fragments d'ADN diffèrent.

Purification des sphères avec l'ADN amplifié



On va ensuite devoir **purifier ces sphères recouvertes d'ADN pour les récupérer.**

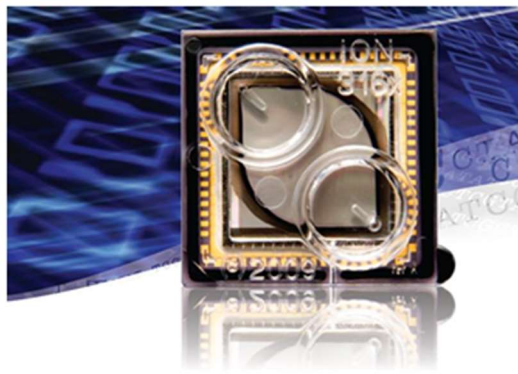
Pour cela on va utiliser la propriété d'un des primers qui avait été recouvert de biotine : on va ajouter nos billes magnétiques recouvertes de streptavidine. Encore une fois, nous avons une **interaction biotine / streptavidine ++**, sur laquelle, si nous appliquons un aimant, nous allons pouvoir récupérer et purifier uniquement les sphères recouvertes d'ADN.

Il faut garder à l'esprit que dans le microréacteur nous pouvons avoir tout type de situation, comme des microréacteurs n'ayant pas de fragment d'ADN amplifié. Il n'y aura donc pas de fixation de la streptavidine (car pas d'ADN fixé = pas de biotine = pas d'interaction biotine/streptavidine) : ces sphères ne nous intéressent pas. Grâce à cette interaction, nous allons donc pouvoir purifier uniquement les sphères qui nous intéressent.

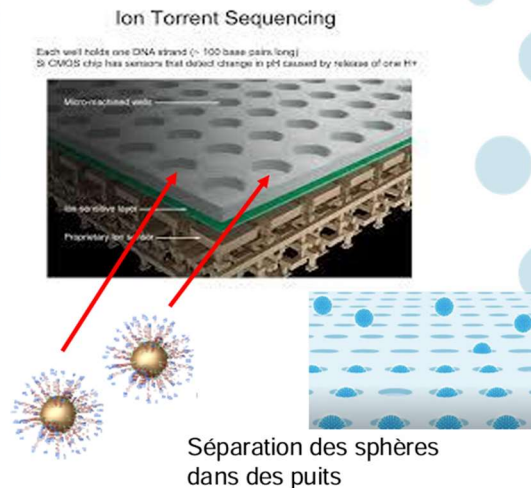
6. Séquençage individuel des sphères avec ThermoFischer

Nous allons ensuite **récupérer nos sphères** et les **séparer physiquement** en les plaçant individuellement dans des puits qui leur seront propres. Une seule sphère sera placée dans un seul puit.

On a des toutes petites puces (quelques cm) contenant 1 à 11 millions de puits, dans lesquelles nous allons donc pouvoir séquencer 1 à 11 millions de sphères en parallèle. C'est donc **système extrêmement puissant**, qui génère énormément de données de séquençage. Le séquençage se fait directement dans ces puces.



1 à 11 Millions de puits

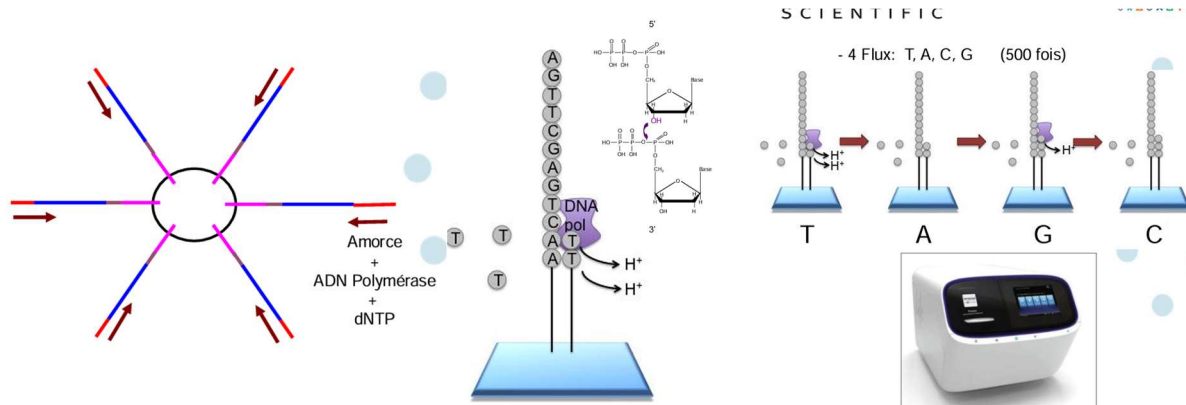


Si on reprend la définition du NGS, on comprend bien maintenant pourquoi ce séquençage est dit :

- **Massif** → 11 millions de réactions en même temps
- **Clonal** → isolation puis amplification d'un même fragment identique de très nombreuses fois → C'est ce qui fait la puissance du système

On va ajouter dans notre milieu réactionnel ce qu'il faut pour réaliser le séquençage :

- Une ADN polymérase
- Des dNTPs
- Une amorce (pour s'hybrider à l'un des adaptateurs et à partir de laquelle la polymérase synthétisera le brin complémentaire)



Lorsque la polymérase ajoute un nouveau nucléotide pour synthétiser le brin complémentaire, il y a formation d'une **liaison phosphodiester** entre les nucléotides. La formation de **cette liaison phosphodiester** libère un ion H⁺ dans le milieu réactionnel, entraînant **des variations de pH**.

L'automate va lire/mesurer ces variations de pH dans tous ces puits en direct. Il lit dans chaque puits s'il y a eu une variation de pH. Les nucléotides sont ajoutés séquentiellement : on envoie d'abord le nucléotide T dans le milieu réactionnel. S'il y a un A en face, la

polymérase va ajouter ce T. On aura alors une variation de pH dans le puits où le nucléotide a été incorporé. Ensuite on rince, entre chaque nucléotide envoyé, un lavage est réalisé. Puis on ajoute le A, pareil : dans les puits où il y a un T en face, le A sera ajouté par la polymérase donc on aura une variation de pH dans ces puits. Et ainsi de suite, ...

Cette succession de 4 nucléotides (T, A, G, C) se fait environ 500 fois dans un run de séquençage. **Le système lit dans chaque puits à quel moment il a lu une variation de pH.**

7. Analyse bio-informatique

Il va falloir transformer nos données obtenues (pH, fluorescence) en données de séquence. **Le travail bio- informatique réalisé à la fin de notre séquençage est énorme.**

+++ C'est grâce à ces outils que nous avons pu développer nos nouvelles techniques de NGS : à la fois par la miniaturisation des systèmes (milieux réactionnels, automates...) mais également par le développement des outils bio-informatiques +++

Étapes :

- Détermination des bases (base calling)

Transformation des signaux générés (pH, fluorescence) en données de séquence

- Étape de contrôle qualité

Élimination des lectures de séquences de **mauvaise qualité** (données générées : des millions → il y a forcément des puits dans lesquels la mesure ne s'est pas faite correctement car pas de produit PCR ou alors 2 produits PCR présents initialement donc l'amplification n'a pas été clonale, etc.). On ne garde que les lectures de bonne qualité.

- Étapes d'alignement

Positionnement des séquences d'ADN générées sur la séquence de référence : on va essayer d'aligner nos différents « reads » / lectures sur notre séquence de référence.

Après avoir tout fragmenté et tout séquencé individuellement, on « reconstruit le puzzle » grâce à l'informatique. On prend pour référence le génome humain puisqu'il est entièrement connu. On pourra ensuite chercher les variants / les différences.

- Étape d'analyse

On regarde si les gènes que l'on voulait étudier, ont été séquencés correctement.

Est qu'il a été séquencé dans sa totalité ? Quel a été le nombre de lecture pour chacune des bases ?

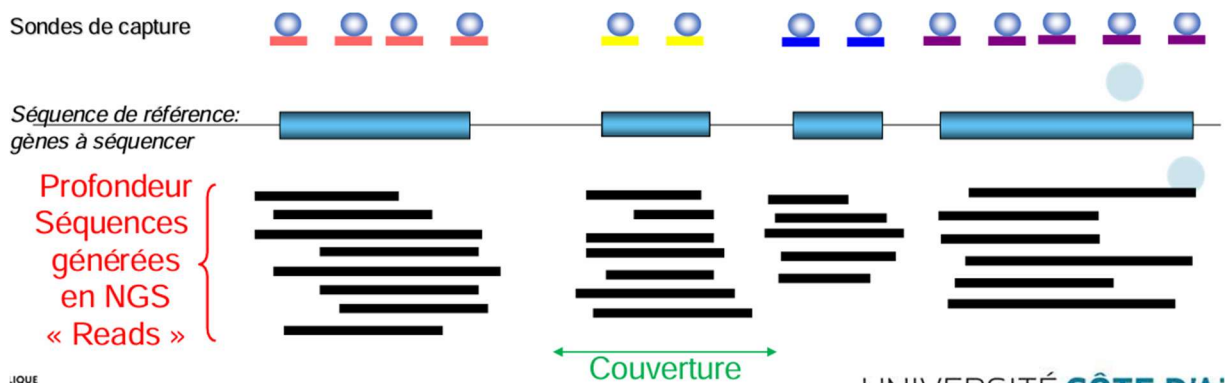
2 notions très importantes :

- **Profondeur de lecture** : nombre de lectures « reads » indépendantes d'une base. C'est le nombre de fois où chacune des bases d'intérêt a été lue sur des reads différents.

- **Couverture** = est ce que je lis bien toute ma séquence d'intérêt :

$$\frac{\% \text{ de bases séquencées par lecture indépendante}}{\text{Nombre total de bases de la région d'intérêt}}$$

→ Valeur donnée pour une profondeur de lecture donnée.



On vérifie donc que tout l'exon a été lu correctement et si la profondeur est suffisante. On considère un séquençage de bonne qualité lorsque chacune des bases de l'ensemble des exons a été lue **au moins 20 à 30 fois** (au moins 20 à 30 séquences indépendantes).

- Étape d'annotation des variants

Consiste à rechercher les localisations chromosomiques des variants sur la séquence, en confrontant la séquence de notre patient et celle de référence. Les annotations, permettant une classification, vont être :

- Localisation sur le chromosome ? Sur quel gène ? À quel endroit du gène ?
- Variants nucléotidiques exoniques ? introniques ?
- Conséquences sur la protéine après traduction ? Caractère pathogène ?

Si on a une mutation qui donne un codon stop très tôt dans le gène, on n'aura pas la protéine.

- Fréquence de ce variant ? Dans quelles populations ? (Caucasiennes, asiatiques...)

Si votre variant est retrouvé chez 70 % de la population on ne va pas pouvoir le tenir responsable d'une maladie rare.

Le gros de ce tri et de l'annotation (et quelques prédictions) est heureusement **fait par l'informatique**, par des algorithmes de plus en plus performants. Ça aide à l'interprétation du variant, et à créer des bases de données.

On classe donc les variants en 5 classes : Bénin (polymorphisme) / Probablement bénin (on en est pas trop sûr mais à priori on en tient pas compte) / Variants de signification inconnus (ce sont ceux-là aujourd'hui qui posent le plus de problème par ce qu'on est pas capable de les classer) / Probablement pathogène (on a pas tous les arguments en faveur mais on pense quand même bien que c'est ce variant qui est responsable) / Pathogènes (ce sont les variants connus dans la littérature et associé à une maladie)

- **Étape d'interprétation des variants = expertise clinico-biologique**

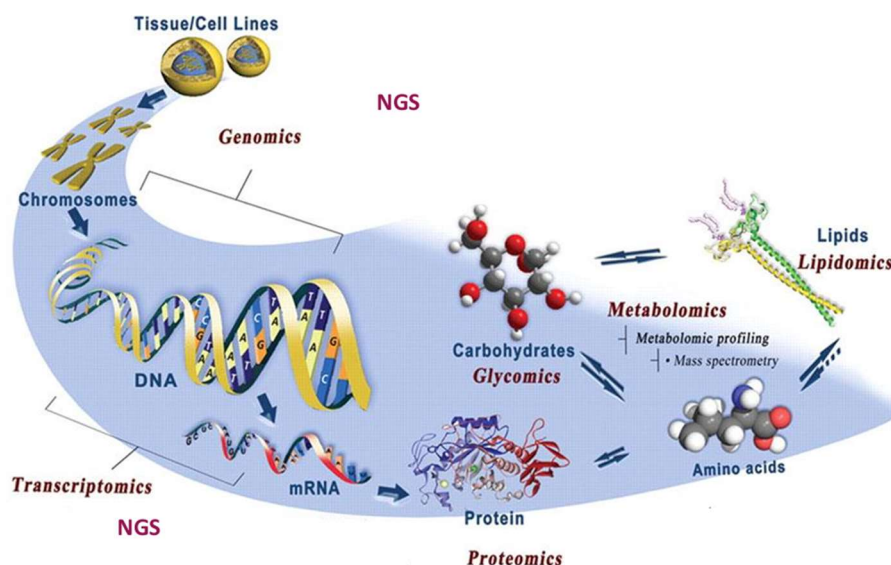
Étape la plus importante (+++) Nous avons une confrontation biologistes/cliniciens, puisque pour bien interpréter un variant, **il faut l'interpréter dans son contexte biologique propre** mais aussi en regardant le **dossier du patient** et notamment en prenant en compte **les signes cliniques du patient, l'histoire familiale...**

Il nous faut répondre à la question suivante : est-ce que les variants identifiés par NGS permettent d'expliquer le phénotype et la clinique que le patient présente et de lui donner un diagnostic ?

C'est l'étape aujourd'hui, la plus difficile, mais aussi la plus importante : séquencer est devenu facile grâce au NGS, mais la difficulté se trouve au niveau de l'interprétation. Il faut trouver le variant responsable de la pathologie parmi de nombreux variants.



L'ère des Multi-OMICS



Le NGS est l'une des briques du multi-OMICS : génomique (génomique, exome), transcriptome.

D'autres techniques s'intéressent aux protéines (protéomique), au métabolisme (métabolomique).

Ce n'est pas tout de séquencer le génome mais il faut savoir l'interpréter. Pour une mutation déjà connue, un syndrome bien défini, l'identification et l'interprétation d'un séquençage peut aller très vite, c'est simple. Cependant, dans beaucoup de maladies génétiques et notamment les maladies rares, tous les gènes ne sont pas connus. On connaît la séquence complète du génome mais pas encore le rôle de tous les gènes.

Utilité du NGS :

- Séquençage de plusieurs gènes (= panel de gènes puisqu'on s'intéresse à plusieurs centaines de gènes)
- Séquençage de l'exome en entier (ce sont toutes les régions codantes d'un génome) (WES)
- « Whole genome sequencing » (on séquence tout le génome de manière intégrale) (WGS)
- DPNI : dépistage prénatal non invasif, qui permet de rechercher les trisomies 13, 18 et 21
- Analyse des ARN : RNA-Seq (expression, épissage alternatif...) (Transcriptome).

On va séquencer les ARN, pour avoir une idée sur l'expression des gènes (quantitatif) mais aussi pour appréhender les variants d'épissage (qualitatif) dans tel ou tel tissu. On va ainsi pouvoir apercevoir un peu plus précisément la fonctionnalité d'un gène puisque ce sont les ARN qui seront traduits en protéines.

On s'aide d'un aspect plus fonctionnel pour comprendre. Au niveau de l'ARN on va voir les modifications au niveau de l'épissage mais ça ne nous donnera pas la mutation responsable. On peut donc croiser les données (génomique + transcriptome par ex.) pour avoir plus d'informations.

Si on trouve un problème au niveau de l'ARN, on revient à notre séquence d'ADN pour voir si on peut trouver la variation nucléotidique responsable de cette anomalie dans notre ARNm.

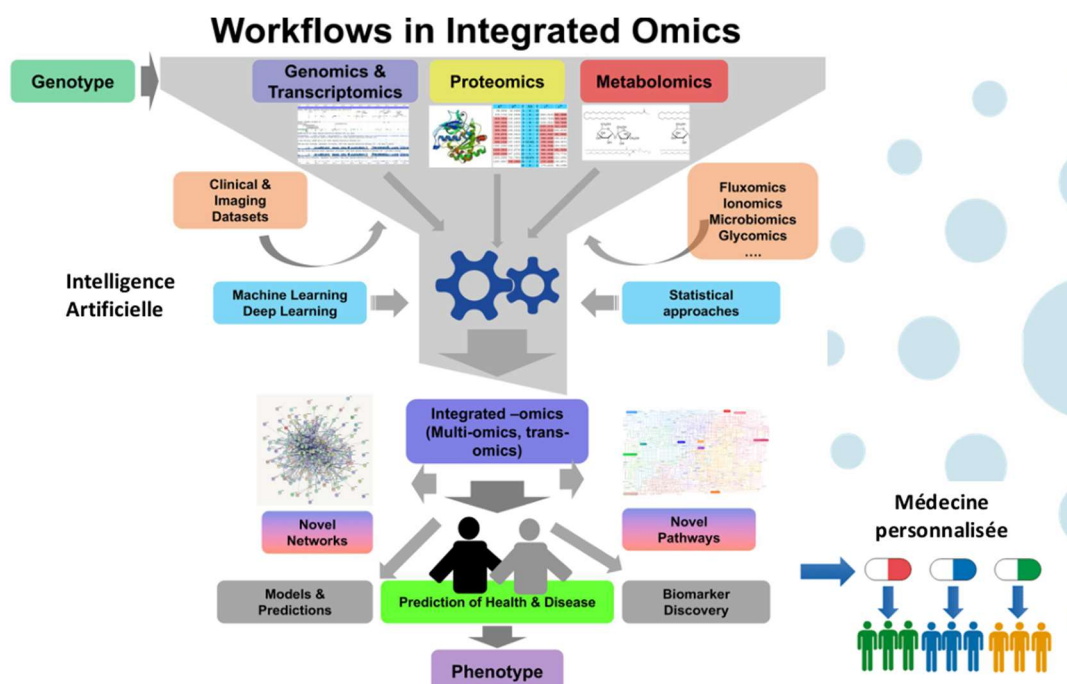
Les multi-omics sont vraiment de nouvelles technologies dont on entendra beaucoup parler dans les années à venir, puisque l'avenir n'est pas dans le séquençage (déjà fait), mais dans l'interprétation et l'application clinique. Il va falloir obtenir la génomique d'un individu dans son intégralité, pour comparer avec sa transcriptomique, sa protéomique, sa métabolomique.

Grâce au développement de l'informatique et de l'intelligence artificielle, on pourrait être capable de croiser l'ensemble de ces données ce qui nous permettrait d'obtenir **un panel complet du métabolisme modifié**. Les voies métaboliques seront, à l'avenir, de plus en plus étudiées et utilisées.

La NGS pourra également nous aider à développer des **thérapeutiques ciblées**, notion dont découle la **«Médecine Personnalisée»** : à chaque patient son traitement, développé à partir de ses voies métaboliques, de ses marqueurs, de ce qui caractérisent l'individu.

L'idée à long terme c'est de ne pas prescrire le même médicament à tout le monde mais de l'adapter en fonction de chacun. Et cela sera possible à partir du moment où on arrivera à tout combiner. Aujourd'hui, on combine pas mal le génome et le transcriptome et petit à petit on rajoute les données protéomiques et métabolomiques. Ça reste encore un petit peu séparé mais de plus en plus vous aurez des approches multiomics.

En France il existe aujourd'hui 2 plateformes nationales qui font du séquençage de génome et dans lesquelles on envoie les échantillons lorsque nos analyses génétiques plus classiques n'ont pas permis d'identifier la mutation génétique responsable de la clinique du patient. Il y a plein d'outils d'intelligence artificielle qui sont en train de se mettre en place de façon à traiter et croiser l'ensemble de ces données (d'une quantité énorme) qu'on a obtenu et qu'on continue d'obtenir en séquençant le génome, transcriptome, protéome, métabolome etc.



En conclusion, la génétique moléculaire est une **discipline qui évolue extrêmement rapidement** → il faut toujours être à l'affut des nouvelles technologies, de ce qui sort dans la bibliographie, des derniers résultats (on peut être amenés à reprendre un dossier 10 ans plus tard car une nouvelle publication est sortie).

Exemple des plateformes vues précédemment : ThermoFischer et Illumina voient déjà une nouvelle génération de séquenceurs arriver alors que ceux actuels sont mis en routine dans les laboratoires depuis à peine une dizaine d'années.

On est très dépendants et les connaissances sont dépendantes de ces nouvelles technologies.

Ce n'est pas parce qu'on séquence le génome/l'exome/une grande quantité de gènes en 2 jours qu'on a les résultats d'un test génétique en 3 jours car difficulté d'interprétation des variants et nécessité de *concertations pluri/interdisciplinaires entre plusieurs spécialistes dont des cliniciens et des biologistes* (expertise clinico-biologique) pour l'interprétation des variants identifiés en NGS (beaucoup sont de signification inconnue). Selon les variants qu'on va trouver, on peut mettre 4 à 5 ans voire plus à donner un diagnostic car il a fallu trouver la fonction de la protéine, comprendre, faire le lien avec la clinique (rendre des variants pour rendre des variants n'a pas de sens).

FIN !!! (de la fiche et du programme)

Bravo d'être arrivés jusqu'ici !! Surtout n'hésitez pas à me faire un retour sur la fiche ! Et aussi à me poser toutes vos questions ou incompréhensions (il n'y a vraiment aucune question bête !!), au pire envoyez moi un message sur messenger si vraiment vous n'osez pas ailleurs (my name is naomi berger). Si je ne peux pas y répondre je transmettrai à la prof (c'est elle-même qui m'a demandé de lui transmettre les questions des étudiants donc elle vous répondra).

C'était la dernière fiche de l'année !! (snif) Le programme est maintenant au complet ! Vous avez trois cours qui peuvent paraître long mais c'est énormément de compréhension, donc comprenez les et entraînez-vous !! J'essayerai quand même de vous poster des fiches récap !

Dédi à ceux qui lisent les dédis avant de commencer le cours (je le faisais aussi et pire je les lisais au début et à la fin)

Dédi à SPARTA !!!!! Donne tout t'es la meilleure (après moi)

Dédi à Clémendocyte (même si tu vas finir par me rendre cardiaque)

Dédi à Maëvacuole, cette star !

Dédi à Marina, la plus chou des anciennes tutrices <3

Dédi à (tes) laulau (love u ma go) ps : je suis contente qu'on puisse se parler mtn que tu penses plus que je te déteste XD

Dédi à Iris (en vrai je t'aurai préféré toi comme marraine mais on va pas le dire à elly)

Dédi à elly quand même et à la banana team = maëva et lila <3

Dédi à Manon geek et à tous les CT (merci et bon courage)

Dédi à Eloise (mais qu'à moitié dédi à la toi qui m'invite au bike par ce que ça fait mal)

Dédi à Alessandra ! ET A SON SUPER PSEUDO : Alestomac !!!!

Dédi à Alicia (t'es trop bien pour la biostat)

Dédi à el papino et à la physio bien sûr !!

Dédi à **phéphé** et à son amour pour la VF

Dédi à Sarah et à ton talent pour la batterie !

Dédi à Amandine (mais pas dédi aux qcm de kiné)

Dédi à Alexandra (bossez la pharmaco les gars)

Dédi à tous mes fillots (vous êtes tous choupis) : Charlotte, Lisa, Anissa, Sarah, Celia, Jade et Mat !!

Dédi au p'tit reuf et à son aide pour le dm annales (il arrive bientôt les gars / meufs)

Dédi à la personne qui habite avec moi : elle s'appelle Tam (je crois)

Dédi aux 4 fantastiques !!! (et encore plus dédi à la personne qui nous a donné ce surnom)

En gardant le meilleur pour la fin dédi à moi !