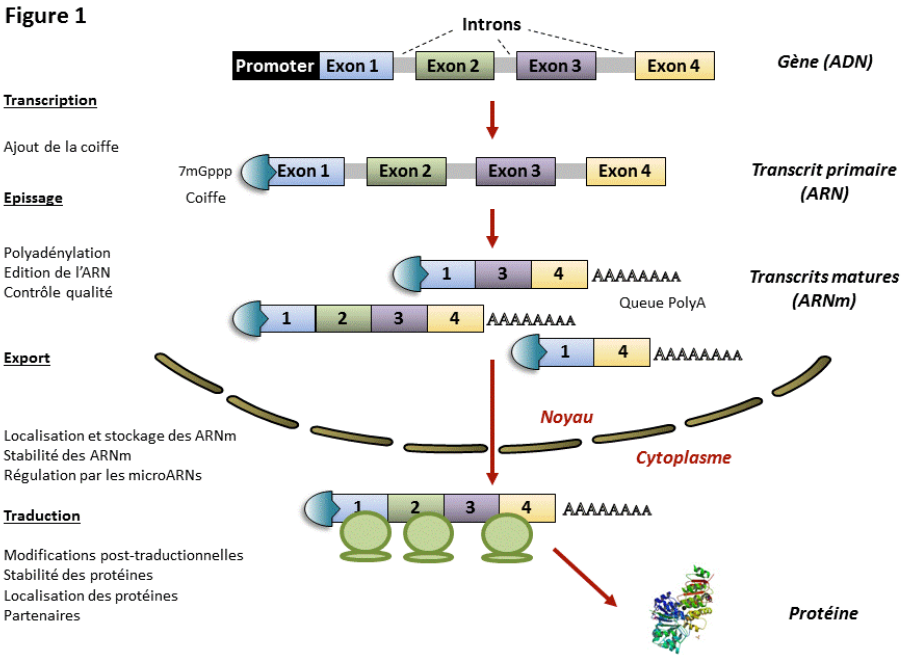
L'épissage alternatif : un gène, combien de protéines ?

Avant la publication de la séquence complète de l'ADN du génome humain, au début des années 2000, on estimait le nombre de gènes à environ 300.000. Aujourd'hui, ce chiffre est tombé à environ 22.000, un résultat étonnant car finalement très voisin de celui d'autres espèces parmi les préférées des généticiens : la souris, le poisson-zèbre ou même le simple ver nématode, qui possède également plus de 20.000 gènes ! En d'autres termes, le nombre de gènes d'un organisme vivant ne reflète pas sa réelle complexité biologique. Ce paradoxe résulte de la combinaison de plusieurs phénomènes, dont ce qu'on appelle l'épissage alternatif des ARN pré-messagers, une étape fondamentale de l'expression des gènes.

Les différentes étapes de l'expression d'un gène

L'expression génique est une succession d'étapes finement contrôlées à différents niveaux (**Figure 1**). En effet, bien que toutes les cellules d'un organisme contiennent le même génome, elles n'expriment pas les mêmes gènes selon l'environnement dans lequel elles se trouvent, ou selon les stimuli auxquels elles sont exposées. La première étape de l'expression d'un gène consiste en sa transcription à partir de son promoteur, une séquence précisant le début de la transcription et dont l'activité est régulée par des protéines spécifiques de chaque type cellulaire. La transcription s'effectue dans le noyau par une machinerie composée de plusieurs dizaines de facteurs protéiques qui copient la totalité de la séquence d'ADN du gène pour donner naissance à un transcrit primaire appelé ARN pré-messager (pré-ARNm).

Figure 1



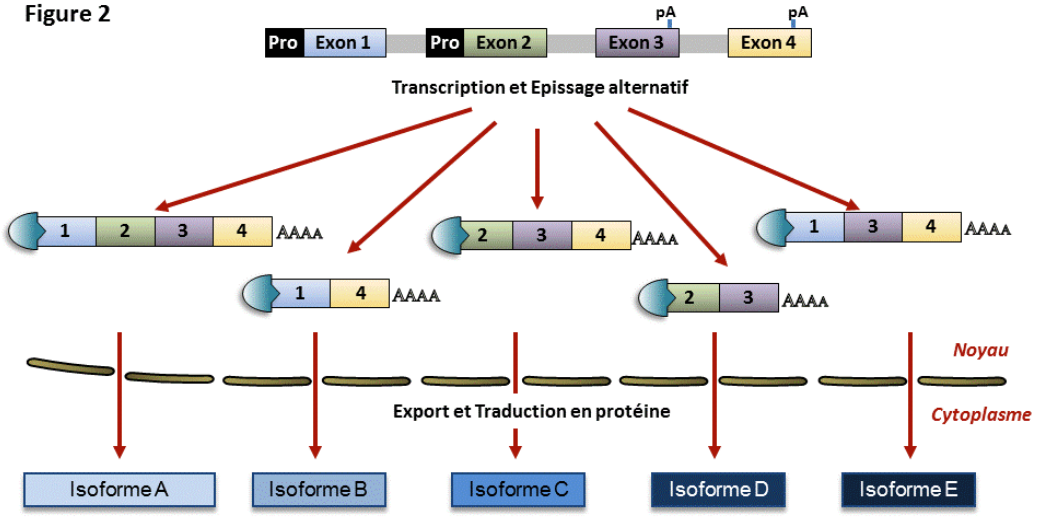
Ce transcrit primaire subit ensuite une série de modification (« maturation ») à ses extrémités : l'ajout d'une coiffe à une extrémité et d'une queue d'environ 200 molécules d'Adénines à l'autre (« capping » et polyadénylation, respectivement). Cependant, la modification principale subie par l'ARN pré-messager consiste en l'élimination d'une grande partie de sa séquence, ce qu'on appelle l'épissage. On sait depuis 1977 que les gènes humains sont discontinus et que les exons, c'est-à-dire les séquences contenant l'information génétique à exprimer en protéines, sont séparés par de larges séquences généralement non codantes appelées introns. On parle ainsi de gènes multi-exoniques. Les exons internes d'un gène mesurent en moyenne 150 nucléotides tandis que les introns mesurent 10 fois cette taille, et parfois bien plus encore. L'épissage consiste en l'élimination des introns et à la mise bout à bout des exons. Le transcrit mature ainsi formé, nommé ARN messager (ARNm), peut alors être exporté dans le cytoplasme (voir ci-dessous) où il sera traduit en protéine. Il est important également de souligner que de nombreux contrôles visant à valider la qualité des ARNm sont réalisés

avant leur export, de façon pour la cellule d'éviter de poursuivre l'expression de transcrits aberrants, ce qui pourrait lui être dommageable.

Vers une nouvelle définition du gène ?

A l'heure où l'emploi de ce mot est devenu courant, il est très difficile de définir ce qu'est un gène. C'est en premier lieu un segment d'ADN qui peut être transcrit en ARN (**Figure 2**). Un gène commence par un promoteur permettant l'initiation de sa transcription et se termine par une

Figure 2

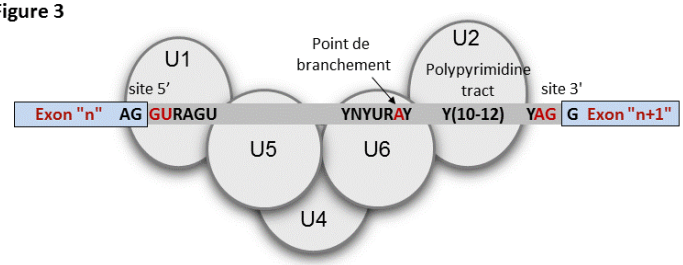


séquence terminatrice. L'initiation et la terminaison peuvent avoir lieu à différents endroits au sein du gène et ainsi mener à la production de transcrits plus longs ou plus courts. Le pré-ARNm peut être épissé d'une manière alternative, il s'agit du processus d'épissage alternatif. En effet, il n'est pas obligatoire que tous les exons d'un gène soient inclus dans l'ARN messager mature, certains exons sont donc considérés comme « alternatifs ». Ainsi un seul gène peut produire différents ARNm matures et par conséquent, plusieurs isoformes protéiques ayant des fonctions biologiques différentes, voire opposées. Au vu de ces données, le traditionnel dogme résumant le flux de l'information génétique à la transcription de l'ADN en un ARNm unique puis à la traduction de ce dernier en protéine, doit être revisité. Il est inconcevable de définir "un gène" comme une entité codant pour une seule protéine. Aujourd'hui, l'unité de l'information génétique n'est plus le gène mais l'exon. Un gène doit plutôt être considéré comme une succession d'exons sélectionnés alternativement et permettant la production d'un ensemble de transcrits matures codant autant d'isoformes protéiques. Ainsi un gène est un message qui peut être interprété par la cellule de différentes manières pour assurer différentes fonctions biologiques selon ses conditions environnementales, son stade de développement et sa spécialisation.

La réaction d'épissage

Les exons du pré-ARNm doivent être réunis ensemble pour maintenir, d'une manière précise, le cadre de lecture (la séquence codante pour la protéine). Par conséquent, le mécanisme moléculaire de l'épissage exige une extrême précision pour reconnaître efficacement les frontières délimitant les introns et les exons dans le pré-ARNm (ce qu'on appelle les séquences 5' et 3' des introns, ou sites 5' et 3' d'épissage) (**Figure 3**). La réaction d'épissage est effectuée par le « spliceosome » (dérivé du mot anglais « splicing », épissage), une machinerie moléculaire complexe constituée de plus de 200 protéines.

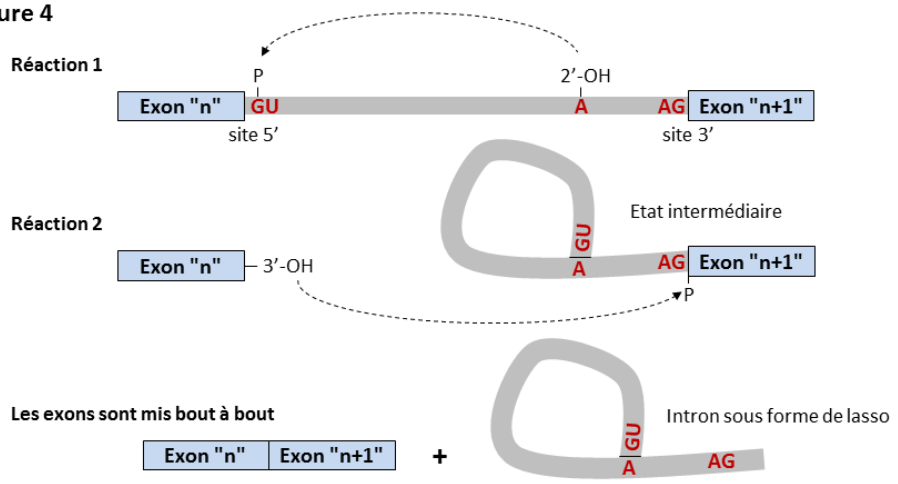
Figure 3



La réaction d'épissage fait appel à 5 snRNPs (petites ribonucléoprotéines nucléaires) : U1, U2, U4/U6 et U5 qui reconnaissent des courtes séquences de nucléotides caractérisant les sites 5' et 3'. Ces séquences ont des caractéristiques (le point de branchement et la séquence riche polypyrimidique) et une localisation particulières aux extrémités de chaque intron.

La snRNP U1 s'associe au site 5' d'épissage et U2 se lie au point de branchement, le tout constitue le pré-spliceosome. L'entrée des snRNP U4/U6 et U5 dans le complexe d'épissage s'accompagne d'une profonde réorganisation du complexe, qualifié alors de « mature », qui permet en son sein le rapprochement physique des sites d'épissage et la réaction catalytique proprement dite (**Figure 4**). Deux réactions chimiques successives assurent la ligation des deux exons et l'élimination de l'intron sous forme d'un lasso. Il est intéressant de savoir que la réaction de l'épissage s'effectue d'une manière synchrone à la transcription du gène et que certains composants protéiques des deux machineries interagissent entre eux.

Figure 4



L'épissage alternatif, une source de diversité des transcrits.

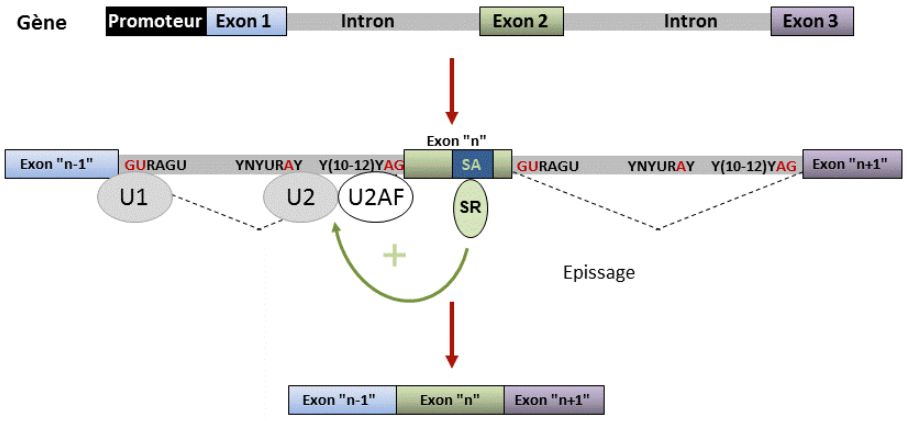
Si l'épissage correspond à l'élimination des introns et à la liaison des exons bout-à-bout dans l'ARNm mature, le choix de l'inclusion des exons peut être alternatif. L'épissage alternatif fournit de larges possibilités pour enrichir le transcriptome (l'ensemble des transcrits d'une cellule) et le protéome (l'ensemble de ses protéines), sans élargir pour autant le nombre de gènes, relativement limité. Grâce à l'inclusion alternative des exons, un seul pré-ARNm peut ainsi produire plusieurs ARNm matures codant pour différentes isoformes protéiques structurellement et fonctionnellement distinctives.

Il a été suggéré que des mutations conduisant, de façon naturelle au cours de l'évolution, à l'affaiblissement des sites d'épissage ou au renforcement de sites d'épissage cryptiques (« cachés » et normalement peu utilisés) peuvent permettre à la machinerie d'épissage de changer le répertoire des ARNm matures générés à partir du même gène. En effet, les exons alternatifs ont des sites d'épissage plus faibles (moins bien reconnus par la machinerie d'épissage) que les exons « constitutifs » (qui sont systématiquement retrouvés dans toutes les formes d'ARNm matures). Néanmoins, il est important de noter que les sites d'épissage sont incapables, à eux seuls, d'initier l'assemblage des composants du spliceosome. En effet, d'autres séquences régulatrices, agissant en *cis* (c'est-à-dire au sein même de la molécule d'ARN, au contraire d'un facteur externe agissant en *trans*), ont été découvertes. Ces séquences, qui peuvent être exoniques ou introniques, sont capables de favoriser ou d'inhiber la reconnaissance d'un exon à l'aide des facteurs d'épissage, dits régulateurs en *trans*.

Il existe deux grandes familles de facteurs d'épissage se liant aux séquences régulatrices en *cis* : les protéines riches en acides aminés serine et arginine, nommées protéines SR, et les protéines des particules ribonucléoprotéiques hétérogènes nucléaires (hnRNP). Conceptuellement et de façon

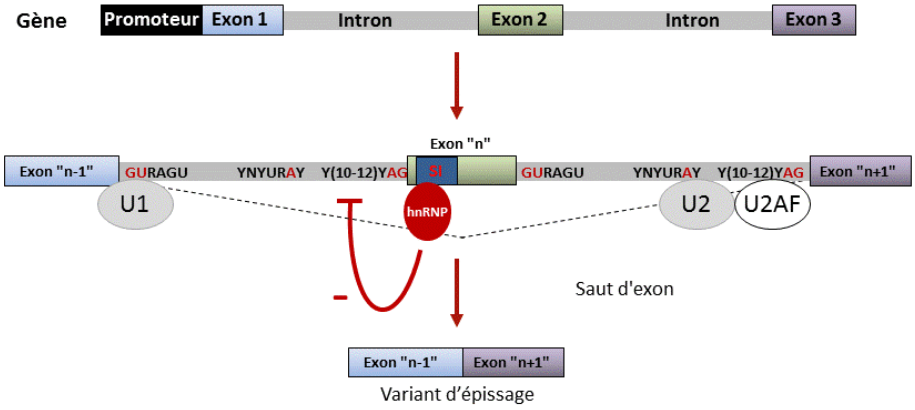
très simplifiée, les protéines SR, en se fixant sur les séquences activatrices (SA), favorisent la reconnaissance des sites d'épissage et facilitent le recrutement des snRNP U1 et U2 (**Figure 5**).

Figure 5



A l'inverse, les protéines hnRNP se lient sur les séquences inhibitrices (SI), jouant le rôle de répresseur en masquant les sites d'épissage (**Figure 6**). L'inclusion ou l'exclusion d'un exon dépend donc de la liaison compétitive de ces différents facteurs à leurs séquences-cibles. D'autres conditions peuvent influencer le choix d'exons alternatifs telles que des modifications affectant les histones (les protéines qui soutiennent et protègent l'ADN dans le noyau), ou la vitesse à laquelle le complexe de transcription copie l'ADN d'un gène en ARN.

Figure 6



Les différents types d'épissage alternatif

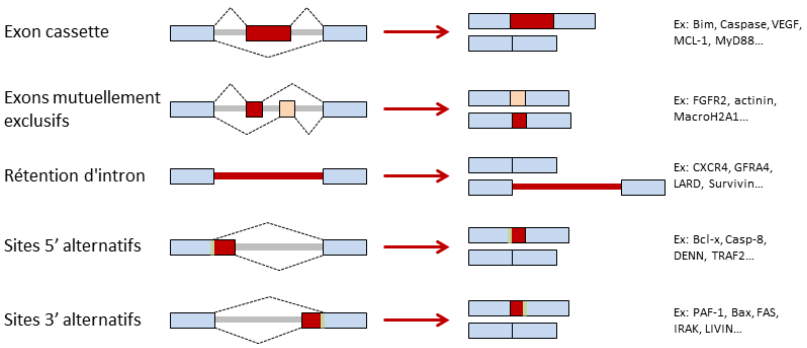
L'épissage alternatif augmente significativement la diversité des transcrits grâce à l'utilisation alternative d'exons ou d'introns, mais aussi de signaux de polyadénylation situés dans des exons alternatifs. La majorité de ces événements se produit dans un maintien du cadre de lecture (la succession de trinucléotides qui permet de décoder la séquence d'acides aminés de la protéine correspondante), résultant en l'expression de différentes isoformes protéiques.

Les exemples d'épissage alternatif peuvent être classés en cinq modèles différents (**Figure 7**) :

1. L'exon cassette (ou le saut d'exon) : il s'agit d'un évènement où un exon est entièrement inclus ou retenu dans le transcrit mature. C'est le cas d'épissage alternatif le plus simple, qui prédomine chez les mammifères, donnant lieu à des transcrits plus ou moins longs.

2. Les exons mutuellement exclusifs : l'un des deux exons, qui sont généralement de taille similaire, est retenu dans l'ARNm mature, mais jamais les deux ensemble. Cet événement change peu la taille du transcrit mature mais modifie les propriétés de la protéine produite.

Figure 7



3. La rétention d'intron : un segment de pré-ARNm peut être épissé comme un intron ou simplement retenu. Cela se distingue du saut d'exon puisque la séquence intronique conservée n'est pas flanquée par des introns. Si l'intron ne modifie pas le cadre de lecture, il code pour des acides aminés comme les exons qui l'envoisinent, produisant ainsi une protéine d'un poids moléculaire supérieur. Dans le cas contraire, il entraine l'apparition d'un codon stop prématuré, ce qui provoque la dégradation du transcrit mature (voir également **Figure 13**). La rétention d'intron est l'événement d'épissage le plus rare chez les mammifères.

4. Site d'épissage alternatif en 5' (site donneur d'épissage) : ce variant d'épissage modifie la borne de l'exon dans sa partie aval.

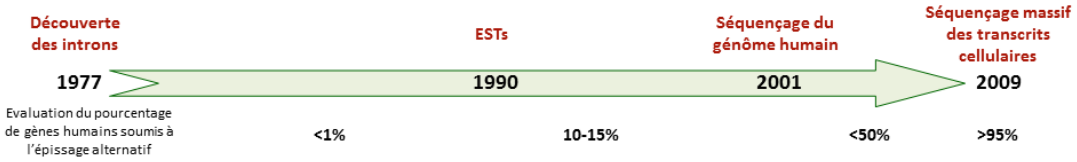
5. Site d'épissage alternatif en 3' (site accepteur d'épissage) : d'une manière similaire au variant précédent, cet événement modifie la borne de l'exon dans sa partie amont, incluant ainsi un segment de l'extrémité 3' de l'intron épissé.

D'autres exemples peuvent être ajoutés aux 5 événements majeurs de l'épissage alternatif, tels que l'usage de promoteurs alternatifs ou de signaux alternatifs de polyadénylation, qui sont connus respectivement comme premier et dernier exon alternatif. Toutefois, en soi ces variants ne sont pas des événements d'épissage puisque leur régulation dépend du choix des sites d'initiation et de la terminaison de la transcription. Néanmoins, ils augmentent également la diversité des ARNm matures générés par un seul gène.

L'épissage alternatif, une règle et non une exception

Les introns, les séquences éliminées des transcrits, ont été découvert en 1977 chez des virus, et rapidement après dans le génôme des vertébrés (**Figure 8**). Auparavant, les études sur

Figure 8



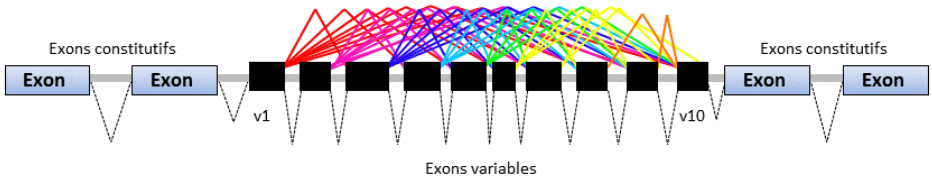
l'organisation des gènes et de leur transcrits avaient principalement été réalisées chez les bactéries, des organismes dépourvus d'introns. Certaines avancées technologiques ont permis de mieux caractériser l'organisation des gènes et de leurs transcrits, et ainsi de mieux évaluer la proportion de gènes soumis à l'épissage alternatif. Au milieu des années 1990, le séquençage de quelques centaines de transcrits humains permettait de prédire à environ 10-15% le nombre de gènes humains produisant au moins deux transcrits par épissage alternatif. Un grand bond en avant a été le séquençage du génome humain au début des années 2000, qui a permis de déterminer le nombre de gènes humains. Les études qui ont suivi et le séquençage de très nombreux transcrits ont fait monter à 50% le nombre de gènes humains produisant au moins deux transcrits. Depuis quelques années, de nouvelles techniques de séquençage massif sont apparues, permettant le séquençage systématique de tous les transcrits dans un état cellulaire donné. Ces méthodes permettent le séquençage parallèle de millions d'ARN. A titre de comparaison, le premier séquençage du génôme humain à duré dix ans pour un coût d'1 milliard d'euros. Aujourd'hui, le séquençage de tous les transcrits de

tous les gènes humains ne coûte que quelques milliers d'euros et peut être obtenu en quelques semaines. Ces nouvelles technologies permettent aujourd'hui d'affirmer que plus de 95% des gènes humains produisent plusieurs variants d'épissage. L'épissage alternatif chez l'Homme est donc une règle, et non pas une exception.

L'épissage alternatif, une source de diversité fonctionnelle des produits des gènes.

Grâce à l'épissage alternatif, un gène peut générer de nombreux ARNm matures et par conséquent coder pour plusieurs isoformes protéiques. La majorité des gènes possèdent en moyenne une dizaine d'exons, les possibilités du saut ou d'inclusion des exons alternatifs sont donc quasi-illimitées. Ainsi, chaque gène peut produire potentiellement des dizaines, voire des centaines de transcrits distinctifs. Un gène comme *DSCAM*, impliqué dans l'adhésion cellulaire, peut produire plus de 15.000 transcrits différents (**Figure 9**), ce qui correspond, à titre de comparaison, au nombre total de gènes chez la Drosophile !

Figure 9

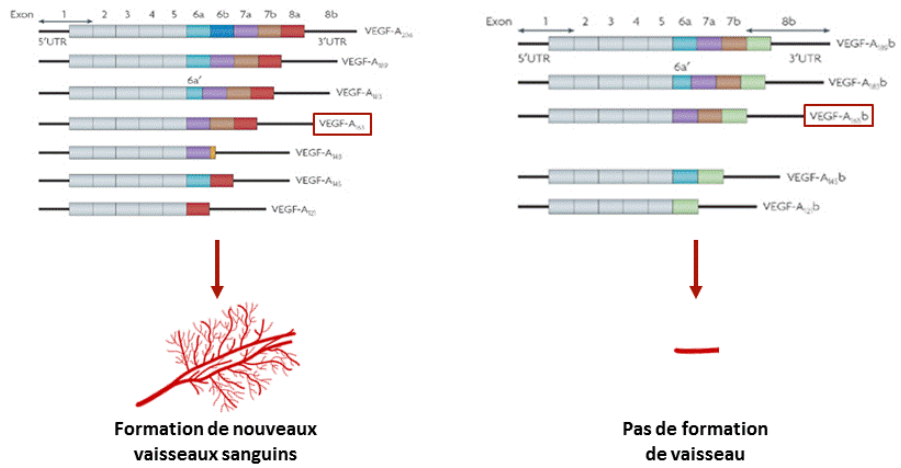


CD44 : plusieurs transcrits
Neurexine : 2.000 transcrits
DSCAM : 15.000 transcrits (autant que de gènes chez la drosophile)
Basonuclin 2 : 90.000 transcrits (100 validés) : >2.000 isoformes protéiques !

La production de plusieurs variants d'épissage rend le message d'un gène plus complexe en y ajoutant un réel impact fonctionnel. Au sein du même processus biologique, un gène peut générer des protéines aux fonctions biologiques différentes, c'est le cas du gène codant pour le récepteur aux oestrogènes alpha. Par épissage alternatif, la protéine produite peut changer de taille et de localisation dans la cellule ; le récepteur dit canonique, d'un poids moléculaire de 66 kiloDaltons (kDa), se trouve dans le noyau où il active l'expression d'un grand nombre de gènes en réponse à l'estrogène. En revanche le récepteur de 36 kDa est cytoplasmique, assurant ainsi les effets non-génomiques de l'hormone estrogénique.

D'une manière intéressante, un même gène peut donner naissance à des protéines aux fonctions opposées au sein d'un processus physiologique. Par exemple, les différentes isoformes générées par épissage alternatif du facteur de croissance des cellules endothéliales vasculaires (VEGF) sont capables de se lier à leur récepteur avec la même affinité, mais elles ne l'activent pas complètement de la même manière (**Figure 10**). Ainsi, l'isoforme VEGF-165b inhibe les effets pro-

Figure 10 **Les différents transcrits produits par le gène *VEGF* (Facteur de Croissance Vasculaire Endothélial)**



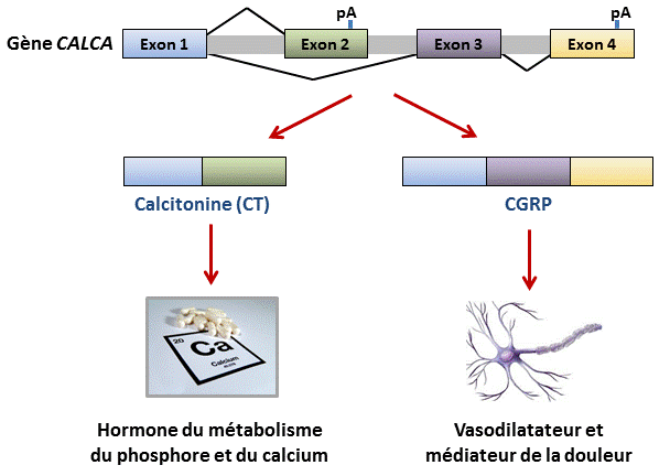
Formation de nouveaux vaisseaux sanguins

Pas de formation de vaisseau

angiogéniques (favorisant la formation de nouveaux vaisseaux sanguins) médiés par l'isoforme VEGF-165.

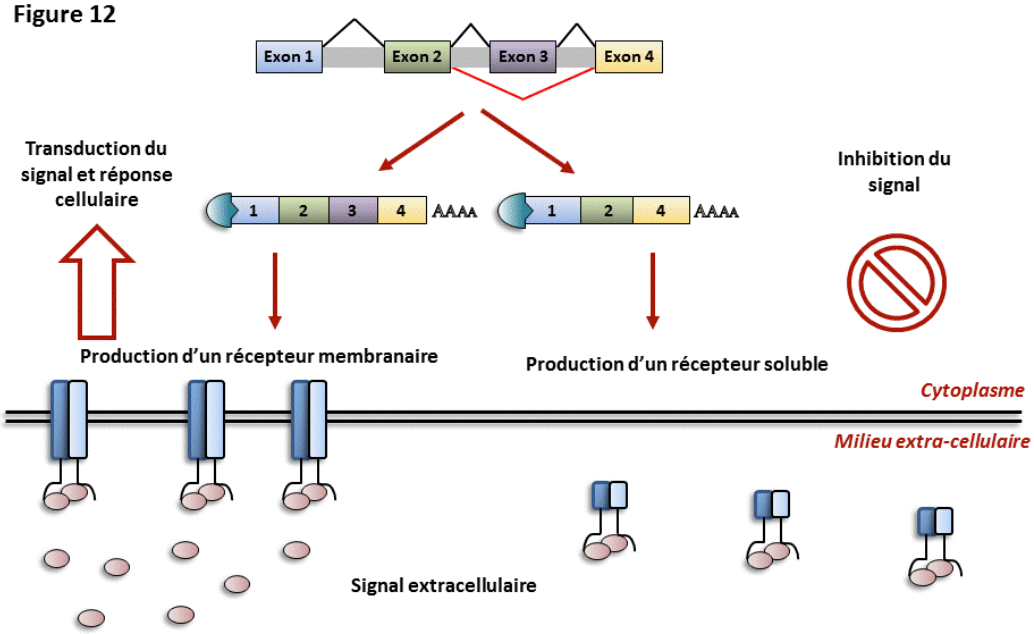
Parfois, la diversité fonctionnelle offerte par l'épissage alternatif peut affecter des processus biologiques complètement différents. Par exemple, les transcrits primaires du gène CALCA produisent, par épissage alternatif dans la glande thyroïdienne, la Calcitonine (CT), une hormone impliquée dans le métabolisme du phosphore et du calcium. En revanche, dans le système nerveux central, les transcrits CALCA produisent un peptide nommé CGRP, un vasodilatateur et médiateur de la douleur (**Figure 11**).

Figure 11



Une autre conséquence possible de l'épissage alternatif peut être la localisation différente des isoformes protéiques produites. Par exemple, un récepteur transmembranaire contient un domaine protéique qui lui permet de s'insérer au travers de la membrane plasmique des cellules, ce qui lui permet de reconnaître des molécules à l'extérieur des cellules (par exemple une hormone ou un neurotransmetteur) et de transmettre des signaux nécessaires à la réponse cellulaire (**Figure 12**). Si un épissage alternatif prive l'ARNm de la région codant pour le signal de localisation transmembranaire, la protéine résultante va par exemple être exportée dans le milieu extracellulaire sous une forme soluble, empêchant la transmission du signal.

Figure 12

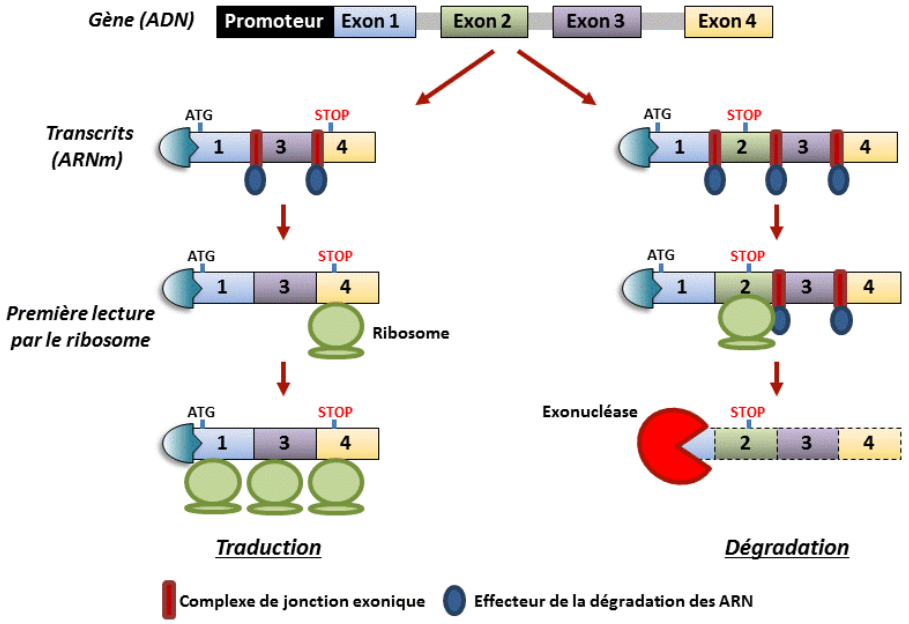


En conclusion, l'ancien dogme basé sur l'hypothèse « un gène, une protéine, une fonction » n'est plus concevable. Aujourd'hui, un gène correspond à un ensemble de fonctions souvent différentes et parfois opposées. La fonction d'un « gène » ne peut donc être réellement appréhendée qu'en analysant la fonction de chacun de ses variants d'épissage. Ainsi, la déception des généticiens devant la révélation du nombre restreint de gènes (environ 22.000) suite au séquençage du génome humain doit donc être tempérée et remise en perspective du fait de la diversité quasi-illimitée offerte par l'épissage alternatif.

Epissage alternatif et dégradation

L'épissage alternatif, source d'une grande diversité des transcrits et du protéome (l'ensemble des protéines codées par ces transcrits), est donc avant tout un mécanisme de régulation qualitative de l'expression des gènes. Cependant l'épissage alternatif peut également assurer un contrôle quantitatif de l'expression d'un gène, en produisant un variant d'ARN « non sens » qui va être reconnu et dégradé par une machinerie cellulaire spécifique (appelée NMD pour *non-sense mediated decay*) (**Figure 13**). Cette machinerie du NMD est en fait responsable d'un processus complexe de

Figure 13



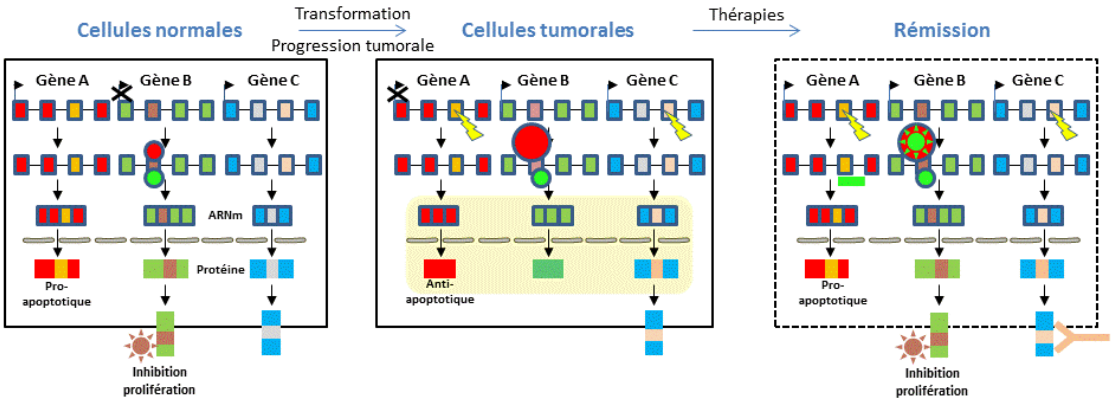
contrôle-qualité des ARNm permettant à la cellule de ne décoder que des ARNm correctement épissés, en particulier dépourvus d'introns, dont la longueur pourrait modifier considérablement la nature des protéines produites et entraîner des dysfonctionnements cellulaires majeurs. La phase codante de lecture est la séquence d'ARN qui va être décodée en protéine par un complexe appelé ribosome. Elle commence par un codon initiateur (ATG) et se termine par un codon de terminaison ou codon STOP. Lors de l'épissage, un complexe protéique appelé complexe de jonction exonique reste associé au niveau des jonctions exon-exon, marquant ainsi l'absence d'intron. Lors de la traduction de l'ARN en protéine, le ribosome va dans un premier temps enlever ces complexes de jonction exonique et s'arrêter naturellement au codon STOP final. Cependant, dans certains cas particuliers, l'épissage alternatif d'un transcrit peut faire apparaitre un codon STOP localisé bien avant le codon STOP final. Ce codon sera alors reconnu et considéré comme prématuré par la machinerie de NMD, entraînant la dégradation de l'ARNm par des exonucléases. Ainsi l'inclusion ou l'exclusion d'un simple exon peut entrainer la dégradation d'un ARN entier. L'épissage alternatif est donc bien un mécanisme de régulation quantitative de l'expression génique.

Epissage alternatif et cancer.

Des changements de l'épissage alternatif participent à la transformation et la progression tumorale (**Figure 14**). Les anomalies d'épissage peuvent être dues à des mutations affectant la reconnaissance d'exons (éclair jaune sur Gènes A et C) ou l'altération de l'expression et/ou de l'activité de facteurs régulateurs de l'épissage, affectant le taux d'inclusion/d'exclusion d'exons (cercle rouge, Gène B).

La détection de variants d'épissage au niveau ARN ou protéique dans les cellules tumorales (carré jaune dans cellules tumorales) permettrait un meilleur diagnostic et sans doute une meilleure classification des tumeurs.

Figure 14



Par ailleurs, de nouvelles stratégies thérapeutiques ciblant l'épissage sont à l'étude. Tout d'abord un épissage aberrant dû à une mutation (Gène A) peut être corrigé par l'utilisation d'oligonucléotides modifiés (trait vert dans cadre rémission), permettant de forcer l'inclusion (ou l'exclusion) d'exons, et donc de changer l'activité du produit du gène cible. Par exemple, forcer à produire une isoforme pro-apoptotique c'est-à-dire entrainant la mort cellulaire au détriment d'une forme anti-apoptotique. De même, l'utilisation de molécules ciblant les facteurs d'épissage (cadre rémission Gene B) permettra de favoriser l'expression de formes inhibant la prolifération cellulaire par exemple. Enfin, si les cellules tumorales expriment des isoformes aberrantes (Gène C), des anticorps dirigés spécifiquement contre ces formes permettront non seulement de mieux détecter les cellules tumorales, mais surtout de cibler ces cellules pour que l'organisme les détruise.

De la plasticité du transcriptome à la plasticité phénotypique

Au-delà de la définition du gène, l'épissage alternatif bouleverse ainsi notre conception du transcriptome (ensemble des transcrits exprimés par une cellule) et du protéome (ensemble des protéines exprimées par une cellule). En effet, chaque gène contenant en moyenne une dizaine d'exons, dont un grand nombre peut être épissé alternativement, cela offre une variabilité combinatoire quasi-infinie permettant à chaque cellule d'exprimer un transcriptome et donc un protéome unique. Ce concept est supporté par l'analyse récente du transcriptome de cellules uniques, qu'il est désormais technologiquement possible de réaliser. Si on analyse des cellules issues d'une même population, c'est-à-dire considérées comme identiques, on s'aperçoit que ces cellules expriment globalement les mêmes gènes mais pas exactement les mêmes variants d'épissage. Ainsi deux cellules de même origine présentes dans le même environnement n'auront pas exactement le même transcriptome. Simplement, ces deux transcriptomes auront une probabilité plus grande d'être similaires que les transcriptomes de deux cellules d'origines différentes ou placées dans des conditions différentes. L'étude de l'épissage alternatif permet donc de concevoir le transcriptome, et par extension le protéome, comme étant « plastique » et « probabilistique », en contraste avec une vision ancienne de l'expression génique, plus rigide et statique, basée uniquement sur le niveau d'expression des gènes. Cette plasticité du transcriptome, associée à la plasticité phénotypique des cellules, pourrait s'avérer fondamentale dans la compréhension de maladies telles que le cancer.

La combinatoire associant variations du niveau d'expression des gènes et variations d'épissage laisse supposer qu'aucune cellule d'un même organisme n'exprime le même protéome. Cette plasticité pourrait contribuer à la très grande diversité des cellules neuronales par exemple. Ainsi, parmi les 100 milliards de neurones présents dans le cerveau, aucun n'a exactement le même protéome. Supportant cette hypothèse, de nombreuses études ont montré que la plus grande hétérogénéité et diversité des variants d'épissage exprimés se retrouve justement dans le cerveau. La plasticité neuronale est donc probablement liée à la plasticité du transcriptome, et il ne fait aucun doute que l'étude de l'épissage alternatif apportera beaucoup dans la compréhension des maladies neuronales et neurodégénératives.

Contributeurs : Simon Samaan, Etienne Dardenne, Didier Auboeuf, Cyril Bourgeois.