# Nomenclature des variations de séquence en génétique : applications en génétique somatique

*F. Escande*, E. Rouleau***

La nomenclature des variations de séquence permet de définir les événements moléculaires de façon univoque. Une uniformisation de la description des variations de séquence a été envisagée dès 1993 *(1-4)*. Ces règles ont ensuite été revues et modifiées régulièrement par la Human Genome Variation Society (HGVS) *[5]*. Dans cet article, nous présenterons les principales recommandations de nomenclature utilisées pour nommer les gènes et les variations de séquence ainsi que les sources et les bases de données auxquelles le lecteur pourra se référer.

## Nomenclature des gènes et des protéines

La nomenclature des gènes suit les recommandations du HUGO Gene Nomenclature Committee (HGNC) *[6]*. Ainsi, les gènes se nomment exclusivement en lettres latines et en chiffres arabes, sans espaces, notamment entre les lettres et les chiffres. Concernant les gènes humains, le symbole doit être écrit exclusivement en **majuscules** et en **italique**, pour être différencié du produit de traduction qu'est la protéine. Par exemple, le gène *KRAS* ne doit pas être écrit "k-ras", et son produit protéique s'écrit "KRAS". De même, le gène *MET* ne s'écrit pas "c-MET", et son produit protéique s'écrit "MET".

## Nomenclature des variants alléliques

Un variant allélique correspond, par définition, au changement d'un ou de plusieurs nucléotides dans la séquence de référence. Les termes "variant" ou "variation de séquence" doivent être préférés à ceux de "mutation" ou "polymorphisme", qui sous-entendent une classification ou une interprétation des conséquences du changement de nucléotide(s).

### Séquences de référence *(tableau I)*

La description des variants doit **toujours** être faite au niveau de l'**ADN** en lien avec une **séquence de référence** qui peut être :

**Tableau I.** Indication de la séquence de référence.

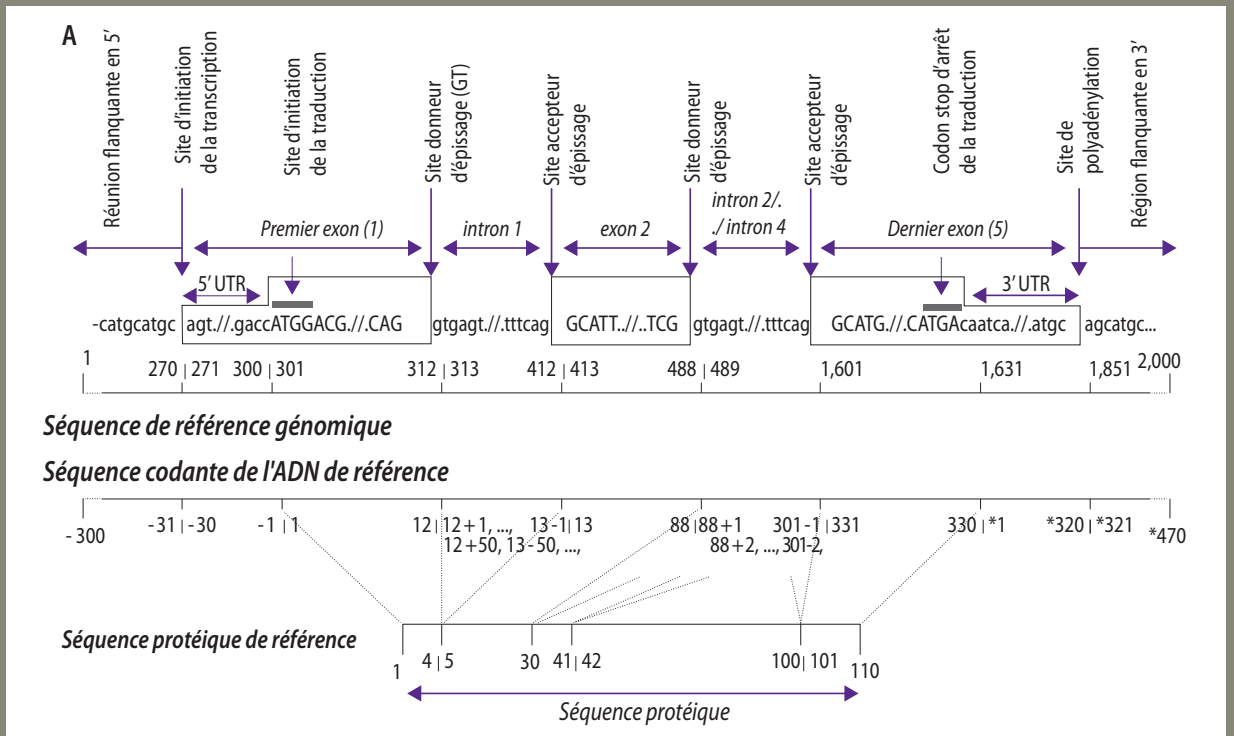
| | | |
|---|---|---|
| ADN | codant | c. |
| | génomique | g. |
| | mitochondrial | m. |
| ARN | | r. |
| Protéine | | p. |

- soit une séquence génomique (position dans le chromosome) ; le préfixe "g." précédera alors la description du variant ;
- soit la séquence codante du transcrit du gène étudié ; dans ce cas, le préfixe "c." précédera la description *(figure 1, p. 120)*.

Les séquences génomiques de référence sont plus précises. Elles font référence au chromosome et à la version du génome étudié (GRCh37/hg19 en 2009, GRCh38/hg38 en 2013). Cette nomenclature est souvent utilisée par la bio-informatique dans les résultats de séquençage à haut débit. En pratique, les séquences de référence de l'ADNc (ADN complémentaire, correspondant au transcrit) sont plus largement utilisées. Elles proviennent usuellement de la base de données RefSeqGene (Reference Sequences Gene), du National Center for Biotechnology Information (NCBI) *[7]*, par exemple : *EGFR* NM_005228.3. Plusieurs transcrits peuvent exister pour un même gène, c'est pourquoi il est important de toujours associer la description du variant à la séquence de référence choisie. Le choix de la séquence de référence se portera sur celle du transcrit principal, qui est souvent le plus abondant dans les cellules. À défaut, c'est le transcrit le plus long ou celui rapporté dans les publications qui sera choisi. Par exemple, pour *EGFR*, il existe 5 transcrits avec un nombre d'exons variables. La séquence de référence utilisée en pratique est celle du transcrit qui contient 28 exons : NM_005228.3 *(figure 2, page 121)*.

** Service de biochimie et biologie moléculaire, unité d'oncologie et génétique moléculaires, pôle de biologie-pathologie-génétique, CHRU de Lille.*
*** Service de génétique, unités de génétique constitutionnelle et de pharmacogénomique, institut Curie, Paris.*

**Figure 1. A.** Séquences de référence (d'après la HGVS). **B.** Exemple de nomenclature d'une mutation activatrice sur le gène *KRAS*.



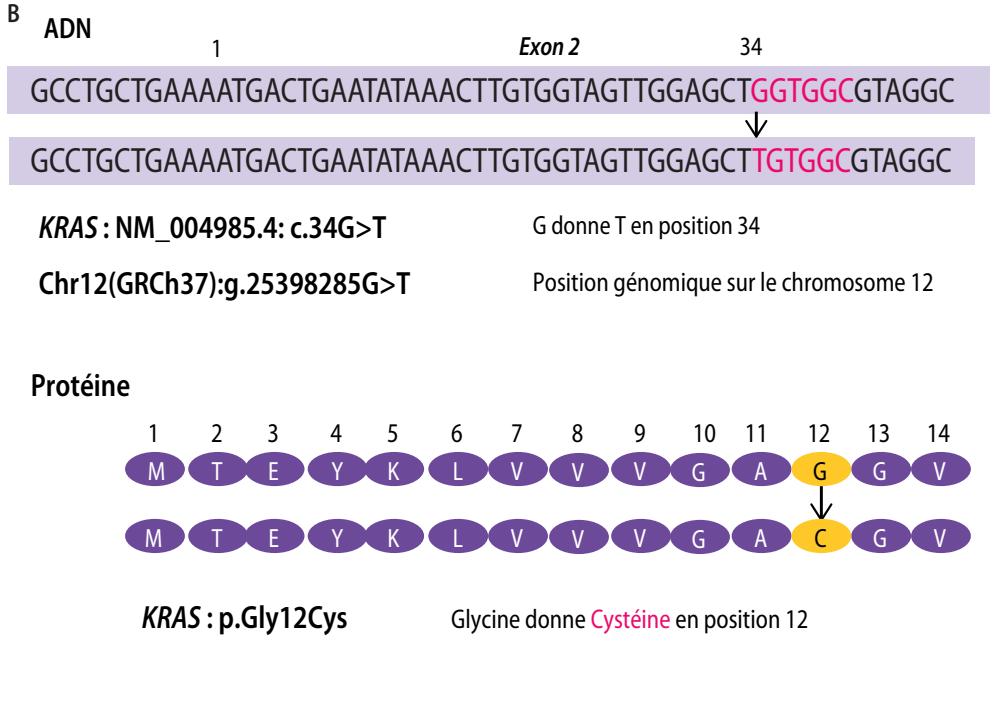



**Figure 2.** Séquences de référence pour le gène *EGFR* (d'après le logiciel Alamut). La séquence NM_005228.3 est la séquence de référence utilisée en routine diagnostique.

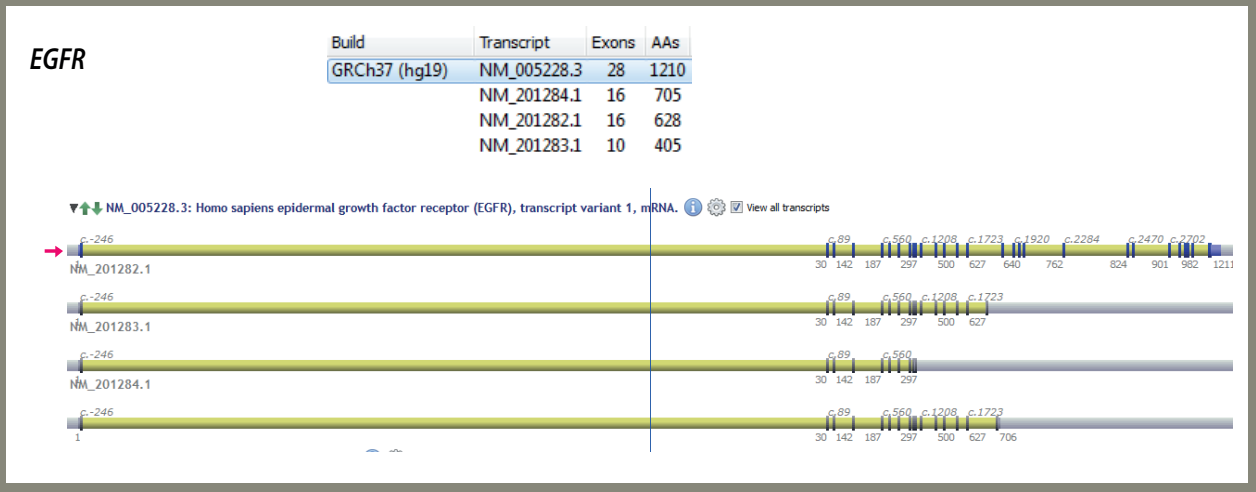
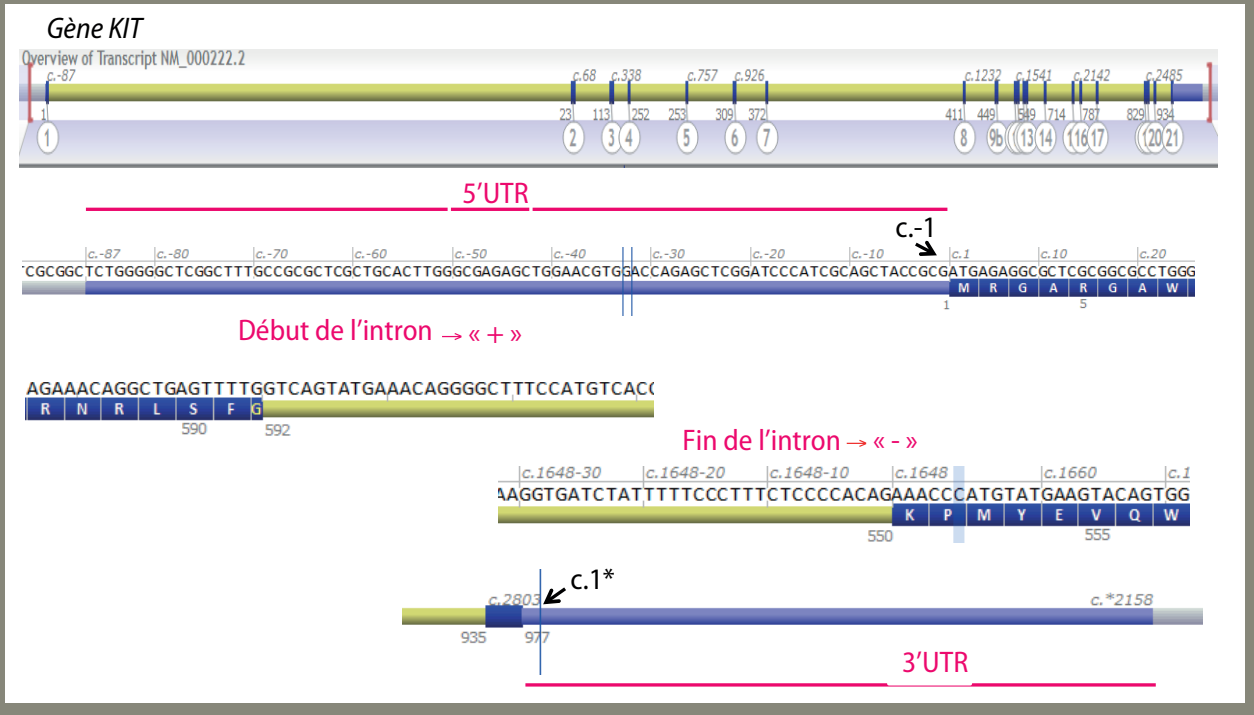
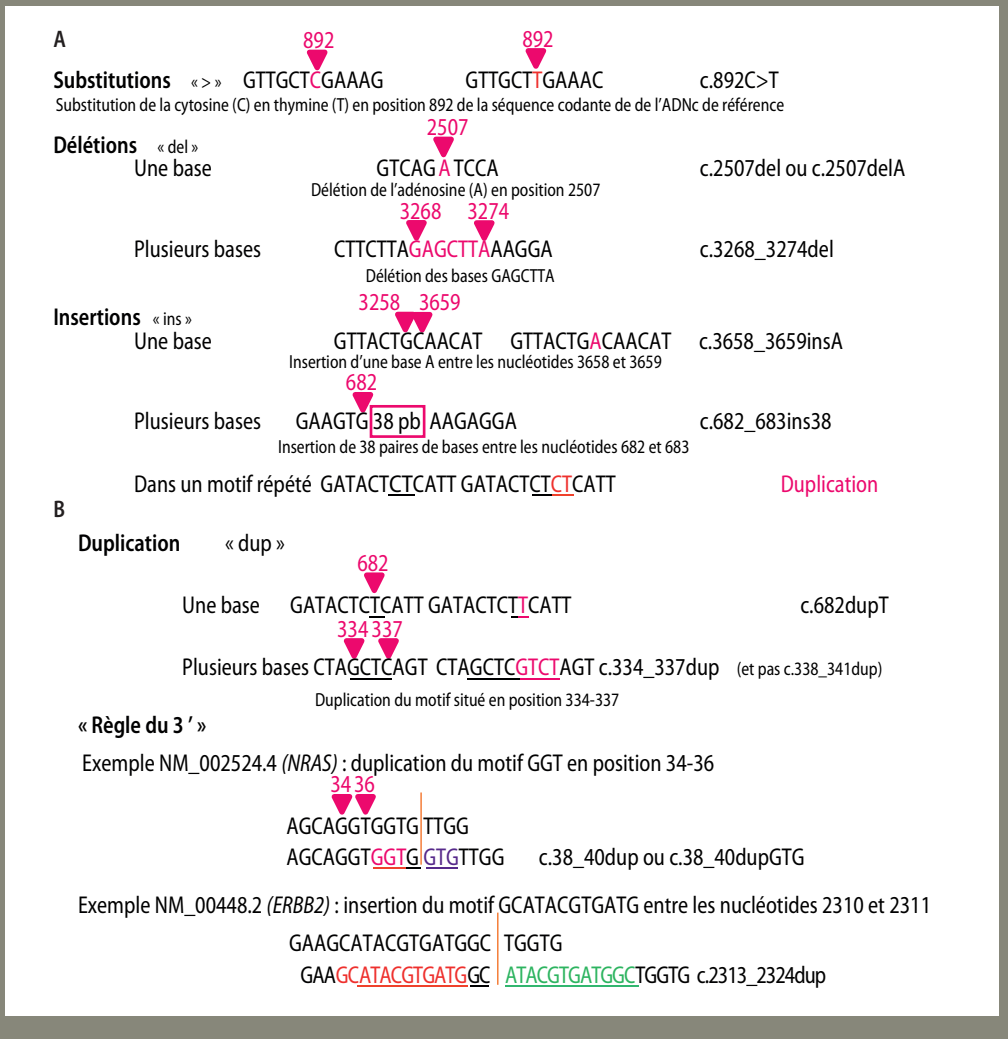



**Figure 3.** Numérotation nucléotidique (d'après le logiciel Alamut). Exemple du gène KIT *NM_000222.2* : 21 exons, séquence codante de 2 485 paires de bases, 934 acides aminés.



La recommandation de la HGVS est maintenant d'utiliser comme séquence de référence les séquences LRG (Locus Reference Genomic) *[8]*, par exemple, LRG_304 pour *EGFR*.
La numérotation des exons n'est pas encouragée par la nomenclature de la HGVS, car celle-ci peut varier en fonction de la séquence de référence et de l'histoire de la découverte du gène. Néanmoins, elle est couramment utilisée en pratique clinique. Classiquement, les exons sont notés de 1 à X à partir du premier exon transcrit, même si celui-ci n'est pas traduit. Il peut néanmoins y avoir des nomenclatures usuelles spécifiques pour certains gènes. Ainsi, la description usuelle du gène *BRCA1* ne comporte pas d'exon numéroté "4". Pour le gène *PIK3CA*, la mutation NM_006218.2(*PIK3CA*) : c.1624G>A, p.Glu542Lys est localisée dans l'exon 10 selon les règles de nomenclature, alors que, dans la littérature, elle est décrite dans l'exon 9. Ce décalage est dû à la présence d'un exon 1 non traduit.
Les variants décrits à partir d'une séquence de référence protéique seront associés au préfixe "p.". Il est important de noter que, pour des raisons de précision, les variants doivent toujours être décrits au niveau de l'ADN. En effet, une même variation sur la protéine peut résulter de variants nucléotidiques différents. Par exemple, le variant BRAF p.Val600Glu correspond aux variants nucléotidiques c.1799T>A et c.1799_1800delinsAA.

**Figures 4.** Règles de nomenclature des variants – nomenclature officielle de la HGVS. A. Partie 1. B. Partie 2.



### Description des variants en nomenclature nucléotidique

Nous nous concentrerons sur la description des variants à partir de la séquence codante du transcrit de référence.

La numérotation des nucléotides sur la séquence de référence de l'ADNc débute au nucléotide 1, A de l'ATG, codon d'initiation de la traduction (les nucléotides situés en 5' de l'ATG seront notés – 1, – 2, etc.) *[figure 1, p. 120, et figure 3, p. 121]*. Pour les nucléotides situés au début d'un intron (après la fin d'un exon), la numérotation du dernier nucléotide de l'exon précédent sera associée au signe "+" et à la position du nucléotide dans l'intron ("c.77+1G"…). Pour les nucléotides situés en fin d'intron (avant le début d'un exon), un signe "–" sera associé au premier nucléotide de l'exon suivant ("c.78–2A"). L'utilisation de l'abréviation IVS *(InterVening Sequence)* n'est pas recommandée pour la numérotation des nucléotides introniques. Cette nomenclature (par exemple, "IVS5+1A") utilisait le numéro de l'exon à proximité duquel intervenait l'anomalie (ici, exon 5 position 1 après la fin de l'exon).

Les changements de nucléotides sont décrits par le signe ">" pour une substitution, "del" pour une délétion et "ins" pour une insertion. L'insertion d'un motif identique à celui de la séquence de référence sera rapportée comme une duplication, et "dup" sera utilisé. Dans le cas d'une combinaison de délétion et d'insertion, "delins" sera utilisé. Il n'y a jamais d'espace dans la dénomination des variants.

La description des principales variations de séquence est résumée dans la ***figure 4*** et le ***tableau II***. Pour les **substitutions**, la variation de séquence est notée en commençant par le numéro du nucléotide de référence suivi du nucléotide de référence, puis du signe ">", suivi du nucléotide modifié ou alternatif ("c.892C>T" signifie que le C en position 892 de la séquence de référence est remplacé par T). La même règle s'applique aux **délétions** : "c.2507delA" signifie la délétion d'un A en position 2507 de la séquence de référence. Pour les délétions, il n'est pas nécessaire de préciser les nucléotides touchés (ainsi "c.2507del" est aussi correct). Si plusieurs nucléotides sont délétés, le variant s'écrira "c.3268_3274delGAGCTTA" ou "c.3268_3274del", le signe "_" indiquant que plusieurs nucléotides sont touchés. Les insertions s'écrivent sous la forme suivante : "c.3658_3659insA", pour l'insertion d'un A entre les nucléotides 3658 et 3659. Pour les **insertions-délétions**, la règle est la même : la position des nucléotides délétés est celle indiquée, et les nucléotides insérés doivent être précisés : "c.34_35delinsTTC" signifie la délétion de 2 nucléotides en positions 34 et 35 et leur remplacement par les nucléotides TTC. Pour les délétions, les insertions et les duplications, c'est toujours la position la plus en 3' de la séquence de référence qui porte le changement (la plus à droite de la séquence). L'environnement des séquences peut ainsi produire des dénominations inattendues en fonction des répétitions de motifs *(figure 4)*.

Deux variations de séquence situées sur un même allèle doivent s'écrire entre crochets : "c.[2573T>G;2369C>T]". Si, par contre, elles sont situées sur 2 allèles différents d'un même gène, il faut écrire "c.[2573T>G];[2369C>T]". Dans la plupart des cas, l'information n'est pas connue ; l'écriture "c.[2573T>G(;)2369C>T]" est alors recommandée.

**Tableau II.** Exemple de nomenclature.

| | |
|---|---|
| **Substitution** | |
| *KRAS* : c.34G>T | Sur le cDNA codant, G, en position 34, est remplacé par T |
| **Duplication** | |
| *STK11* : c.179dup | Duplication du nucléotide 179 |
| *ERBB2* : c.2314_2325dup | Duplication du segment 2314 à 2325 |
| EGFR : p.Tyr772_Ala775dup | Duplication du segment à partir de la tyrosine en 772 jusqu'à l'alanine en 775 |
| **Insertion** | |
| *BRCA1* : c.2745_2746insTT | Insertion de 2 T entre les nucléotides 2745 et 2746 |
| ERBB2 : p.Ala775_Gly776insLeu | insertion d'une leucine entre l'Alanine en 775 et la Glycine en 776 |
| **Délétion-insertion** | |
| *BRAF* : c.1798_1799delinsAA | Délétion des nucléotides 1798 et 1799 remplacés par 2 adénosines AA |
| MSH6 : p.Gly881delinsGluSer | Délétion de la glycine en position 881 et remplacement par 2 acides aminés Glu et Ser |
| **Décalage cadre de lecture** | |
| STK11 :p.Arg28Serfs*23 | Arginine est le premier acide aminé changé, en position 28, en sérine, la taille du segment en aval est de 23, codon stop (*) inclus |

### Description des variants en nomenclature protéique

Le codon d'initiation de la traduction est noté "+ 1". Pour la description des variants protéiques, l'acide aminé de référence est noté avant le numéro du

codon, suivi de l'acide aminé muté (par exemple, "BRAF p.Val600Glu" signifie que la valine du codon 600 est substituée par la glutamine). Le signe ">" n'est pas utilisé pour décrire les modifications au niveau des acides aminés. Les codons stop sont désignés par "*" (par exemple, "TP53 p.Trp91*" signifie que le tryptophane codé par le codon 91 est remplacé par un codon stop). Ils peuvent aussi être nommés "ter". Les délétions sont désignées par l'abréviation "del" ("EGFR p.Leu747_Ser752del" signifie que les codons 747 à 752 sont délétés). De même, "ins" est utilisé pour les insertions et "delins" pour les insertions-délétions ("KRAS p.Gly12_Gly13delinsValCys"). En cas de décalage du cadre de lecture entraînant l'apparition d'un codon stop prématuré, "fs*" est utilisé, suivi de la position du codon stop : par exemple, "STK11 p.Arg28Serfs*23" signifie que l'acide aminé Arginine en position 28 est substitué par une Sérine avec un décalage de lecture sur 23 acides aminés (codon stop à la 23e position).
Le code des acides aminés à 3 lettres est à préférer, car il ne permet aucune ambiguïté dans l'interprétation du résultat. En effet, l'écriture "G12C", correspondant à un changement de glycine en cystéine sur le 12e acide aminé de la protéine, pourrait être confondue avec une mutation en position 12, d'un nucléotide guanine en cytosine.

## Conclusion

Pour compléter cette description de la nomenclature, on pourra se reporter au site de la HGVS *(5)*, en particulier pour le nom des variants plus complexes. Les règles sont parfois arbitraires, et surtout sujettes à des mises à jour régulières. Par exemple, le codon stop a été nommé directement "stop", puis "*" (astérisque) et maintenant "Ter". Néanmoins, l'intérêt de la nomenclature internationale est d'éviter toute ambiguïté dans la description des variants lors de la poursuite des analyses ou du suivi des patients. L'homogénéité de ces résultats est en constante amélioration dans les comptes-rendus de génétique somatique. Parallèlement à cette nomenclature officielle, il est évidemment important de joindre une interprétation précise de l'effet du variant. Pour aider le clinicien, elle peut aussi être associée à une nomenclature usuelle, en particulier si elle est utilisée dans les essais cliniques.

*Les auteurs déclarent ne pas avoir de liens d'intérêts en rapport avec la fiche réalisée.*

## Références


**1.** *Beaudet AL, Tsui LC. A suggested nomenclature for designating mutations. Hum Mutat 1993;2:245-8.*

**2.** *Beutler E. The designation of mutations. Am J Hum Genet 1993;53:783-5.*

**3.** *Beutler E, McKusick VA, Motulsky AG, Scriver CR, Hutchinson F. Mutation nomenclature: nicknames, systematic names, and unique identifiers. Hum Mutat 1996;8:203-6.*

**4.** *Antonarakis SE. Nomenclature Working Group. Recommendations for a nomenclature system for human gene mutations. Hum Mutat 1998;11:1-3.*

**5.** *http://www.hgvs.org/mutnome*

**6.** *http://www.genenames.org/about/guidelines*

**7.** *http://www.ncbi.nlm.nih.gov/refseq/rsg/*

**8.** *http://www.lrg-sequence.org/*