In vitro veritas 2010; 11:103-109 ISSN: 1697-5421

Revisió

Secció de Genètica molecular de l'Associació Catalana de Ciències de Laboratori Clínic

Descripció de les variacions de seqüència en genètica molecular

Ariadna Padró Miquel, Beatriz Candás Estébanez

Laboratori Clínic, Hospital Universitari de Bellvitge, L'Hospitalet de Llobregat

Introducció

És imprescindible que la comunitat científica arribi a acords internacionals per tal de promoure una comunicació clara i plenament informativa que no estigui subjecte a males interpretacions degudes a falta de claredat o a traduccions lingüístiques poc apropiades. Aquesta problemàtica encara es fa més accentuada pel que fa al camp de la genètica molecular, donada la ingent quantitat de variacions de seqüència descrites fins al moment en el DNA, i el ritme vertiginós de nous descobriments.

En una revisió bibliogràfica de genètica molecular, la dificultat rau en la diversitat de descripcions utilitzades per a una mateixa variació. Per exemple, la nomenclatura recomanada internacionalment de la variació del gen de l'apolipoproteïna C-III que es caracteritza per un canvi puntual a la zona 3' UTR (zona no traduïda) del gen, 40 posicions a continuació del darrer nucleòtid codificant, és "NM_000040.1:c.*40C>G". Tot i això, a la literatura científica, es pot trobar descrita com a "Sst1, 3238C>G" o "c.386C>G", entre d'altres. I és que sovint els científics han optat per nomenclatures diferents i ambigües, basades en el nom de l'enzim de restricció emprat —evidentment no exclusiu de la variació en estudi—, en un identificatiu arbitrari o bé en la posició de la variació dins d'una seqüència de DNA, de RNA missatger o de proteïna que no és de referència ni està degudament identificada. A més a més, els articles publicats generalment no contenen descripcions "sinònimes" de les variacions, amb la conseqüent impossibilitat d'identificar-les si no es coneixen amb antelació.

La Unió Internacional de Química Pura i Aplicada i la Federació Internacional de Química Clínica han impulsat diverses iniciatives encaminades a unificar

criteris pel que fa a la descripció sistemàtica de les propietats biològiques, entre elles les pertanyents a la genètica molecular (1), per tal de facilitar el traspàs d'informació inequívoca entre els professionals de la salut. Soares de Araujo *et al.* (1) recomanen seguir la guia proposada per Antonarakis i Den Dunnen (2) pel que fa a la descripció de les variacions de seqüència per part dels laboratoris que examinen propietats de genètica molecular.

L'objectiu del present article és recollir i acostar als professionals del laboratori clínic les recomanacions d'Antonarakis i Den Dunnen per descriure les variacions de seqüència de les propietats de genètica molecular.

Descripció de la variació

Les recomanacions d'Antonarakis i Den Dunnen per la descripció de variacions de seqüència estan pensades per tal que siguin estables, amb significat i inequívoques; amb especial atenció a eliminar inconsistències i clarificar convencions confuses.

En primer lloc es recomana l'ús del terme neutre "variació de seqüència", en detriment de "mutació" o bé "polimorfisme", ja que aquests darrers no tenen una definició inequívoca. El terme "mutació" s'ha emprat per indicar un canvi de seqüència mentre que en altres disciplines indica un canvi que causa una malaltia. De forma similar, el terme "polimorfisme" s'usa tant per indicar un canvi que no causa malaltia, com un canvi que es troba present en més de l'1 % de la població.

Nomenclatura del gen

La descripció de qualsevol variació de seqüència comença amb el nom del gen estudiat. Seguint les recomanacions de la Unió Internacional de Química Pura i Aplicada i la Federació Internacional de Química Clínica, s'ha d'emprar el símbol oficial del gen (excepcionalment en lloc del nom complet), però amb lletres rodones (no cursives), disponible a la *Human Genome Nomenclature Database, HUGO* (3), precedit per la paraula "Gen". També es pot consultar a l'entrada per gen de la base de dades del *National Center for Biotechnology Information* (4).

Per exemple:

Malaltia	Denominació recomanada del gen estudiat	Altres noms de gen no recomanats
Hemocromatosi	Gen HFE	HH; HFE1; HLA-H
Atàxia espinocerebel·losa de tipus 1	Gen ATXN1	SCA1
Malaltia de Huntington	Gen HTT	HD, IT15

Descripció de la seqüència de referència

Totes les variacions haurien d'estar descrites al nivell més bàsic, és a dir, sobre el DNA. Les descripcions han d'estar sempre en relació a una seqüència de referència, ja sigui genòmica o de DNA codificant, preferentment aquesta darrera, donat que, tal com veurem més endavant, la pròpia descripció informa de si la variació és intrònica, exònica o en zona no codificant. La seqüència de referència sol trobar-se abreujada com a *RefSeq* a les bases de dades. Va seguida de la codificació "NC_" per DNA genòmic, "NM_" per DNA codificant i "NP_" per la proteïna, i a continuació un número característic de la seqüència, i un segon número separat per un punt, característic de la versió emprada.
Per exemple "NM_000410.3", caracteritza la tercera versió de la seqüència de referència del DNA codificant del gen de l'hemocromatosi.

Quan s'informa del nom abreujat del gen i el número de seqüència, s'han d'emprar els símbols { }. Seguint amb l'exemple anterior, es descriuria: Gen HFE{NM_000410.3}

Posició de la variació a la seqüència de referencia

La descripció de la posició de la variació ha d'anar precedida d'una lletra que indica el tipus de seqüència de referència emprada. Així, s'empra "c." per al DNA codificant, "g." per al DNA genòmic, "m." per al DNA mitocondrial, "r." per l'RNA i "p." per a la proteïna.

Pel que fa al DNA codificant, s'atorga la posició +1 a la adenina (A) del primer ATG que codifica per la primera metionina de la proteïna. La seqüència del DNA codificant es pot obtenir fàcilment des de la base de dades del gen d'interès, seleccionant l'opció CDS (de l'anglès *CoDing Sequence*). La nova finestra mostra la seqüència del DNA codificant juntament amb els aminoàcids de la proteïna que codifica cada trinucleòtid, així com la versió i número de seqüència.

El següent quadre, adaptat de la pàgina web <http://www.hgvs.org/mutnomen>, és un exemple molt il·lustratiu de numeració de la posició de la variació en funció de la seqüència emprada:

		Numeració dels nucleòtids a la seqüència de referència		
Part del gen		**DNA genòmic**	**DNA codificant**	**proteïna**
Regió flanquejant 5' del gen		1 a 270	(-300 a -31)	-
exó 1	5' UTR	271 a 300	-30 a -1	-
	regió codificant	301 a 312	1 a 12	1 a 4
intró 1		313 a 412	*12+1 ... 12+50, 13-50 ... 13-1*	-
exó 2		413 a 488	13 a 88	5 a 29 (30)
intró 2		489 a 689	*88+1 ... 88+100, 89-100 ... 89-1*	-
exó 3		689 a 723	89 a 123	30 a 41
intró 3	conté "splicing" alternatiu poc freqüent des de la posició 800 a 859 (DNA codificant 123+77 a 123+136)	724 a 1023	*123+1 ... 123+150, 124-150 ... 124-1*	-
exó 4		1024 a 1200	124 a 300	42 a 100
intró 4		1201 a 1600	*300+1 ... 300+200, 301-200 ... 301-1*	-
exó 5	**regió codificant**	1601 a 1630	301 a 330	101 a 109
	3' UTR conté una extensió $(CA)_7$-des dels nucleòtids 1700 al 1713 (DNA codificant *70 a *83); addició d'una cua de poli-A a la posició 1825 (DNA codificant *195)	1631 a 1850	*1 a *220	-
Regió flanquejant 3' del gen		1851 a 2000	(*221 a *370)	-

És a dir, a la seqüència codificant del DNA, el signe negatiu és indicatiu de la zona 5'UTR (per exemple "c.-5G>T"), l'absència de signe és indicativa de la zona codificant (per exemple "c.5G>T"), qualsevol signe després d'un número indica una zona intrònica (per exemple "c.256+1G>T" a l'extrem 5' d'un intró o bé "c.257-1G>T" a la zona 3' d'un intró), i finalment l'asterisc és indicatiu de la zona 3'UTR (per exemple "c.*5G>T").

Tipus de variació de seqüència

Els símbols per informar del tipus de variació de seqüència trobada també estan definits:

- Les substitucions es designen mitjançant ">" després del número del nucleòtid afectat. Per exemple "c.845G>A" o "c.123+89C>T".

- Les delecions es designen mitjançant les lletres "del". Per exemple "g.413delG" per a una deleció puntual o "c.92_94del3" —o "c.92_94delGAC"— per a una deleció de tres nucleòtids que comença a la posició 92 i acaba a la 94.

- Les duplicacions s'indiquen amb les lletres "dup". Per exemple "c.92_94dupGAC" —o "c.92_94dup3"— per a una duplicació dels nucleòtids situats a les posicions 92, 93 i 94.

- Les insercions s'indiquen amb les lletres "ins". Per exemple "c.51_52insT" per a una inserció puntual o "c.51_52insGAGA" per a una inserció de tres nucleòtids. No s'han d'incloure les duplicacions.

- Les repeticions de petites seqüències s'indiquen entre claudàtors. Per exemple "c.*70CA[6]+[11]" identifica 6 repeticions del dinucleòtid CA en un al·lel i 11 en l'altre, a la posició 70 de la zona 3'UTR del gen.

Les inversions, conversions, translocacions i altres canvis més complexos s'indiquen igualment amb els seus símbols corresponents.

Seguint amb l'exemple de l'hemocromatosi, la variació "p.C282Y" es troba a la posició "c.845G>A". Per tant, la forma recomanada d'indicar aquesta variació és:

Gen HFE{NM_000410.3}:c.845G>A

De la mateixa manera, per la malaltia de Huntington, la forma adequada per indicar la repetició de trinucleòtids —si, per exemple, s'observessin 17 repeticions en un al·lel i 24 a l'altre— seria:

Gen HTT{NM_002111.6}:c.51CAG[17]+[24]

Variacions en el DNA mitocondrial

El DNA mitocondrial és petit i completament conegut. En aquest cas concret, es recomana descriure les variacions en relació a la seqüència completa d'aquest DNA (seqüència genòmica de referència: NC_012920.1). Les descripcions haurien d'anar precedides per "m.". Per exemple, les tres variacions de seqüència més prevalents en la malaltia de Leber s'haurien de descriure de la següent manera:

Gen MT-ND1{NC_012920.1}:m.3460G>A

Gen MT-ND4{NC_012920.1}:m.11778G>A

Gen MT-ND6{NC_012920.1}:m.14484T>C

Els canvis en la seqüència de proteïna, per tal de distingir que es tracta de producció mitocondrial, haurien de ser descrits incloent-hi una referència a aquesta proteïna, per exemple p.ATP6:Leu156Pro (NP_536848.1).

Casos particulars

Els diferents al·lels pertanyents a gens del sistema dels antígens leucocitaris humans (HLA) segueixen una nomenclatura especial (5). Cada al·lel HLA té assignat un únic número que correspon a un màxim de quatre subgrups de números diferents. Generalment s'empren els primers dos subgrups de números, i es reserva l'ús del tercer i quart per quan és estrictament necessari.

El gen es denomina segons la nomenclatura oficial (Gen HLA-A, Gen HLA-B, etc.) seguit d'un asterisc i del grup al·lèlic, que generalment es correspon amb l'antigen que se'n deriva (per exemple, HLA-B*57). El segon subgrup numèric, separat per dos punts,

identifica l'al·lel específic (seguint amb l'exemple anterior: HLA-B*57:01) el qual segueix el número d'ordre amb el qual les seqüències de DNA s'han anat determinant. Els al·lels que difereixen en els dos subgrups de nombres han de diferir en un o més nucleòtids que deriven en un canvi de seqüència de la proteïna codificada. Els al·lels que difereixen només en substitucions de nucleòtids sinònimes (substitucions no codificants) es distingeixen pel tercer subgrup de nombres (per exemple, HLA-B*57:01:07). I finalment, els al·lels que únicament difereixen en la presència de variacions de seqüència dels introns o de les regions 5' o 3' UTR (de l'anglès *untranslated regions*), es distingeixen per l'ús del quart subgrup de nombres (per exemple, HLA-A*02:101:01:02).

Els al·lels del sistema HLA descrits fins al moment es troben detallats a la base de dades IMGT/HLA, que inclou les seqüències oficials definides per la *WHO Nomenclature Committee For Factors of the HLA System* (6).

També s'han construït d'altres bases de dades molt extenses amb gens pertanyents a grups concrets com els del Citocrom P450 (7), o els de l'aldehid deshidrogenasa (8). Cal tenir en compte, però, que sempre que sigui possible, s'hauria de seguir la identificació de variacions suggerida per Den Dunnen i Antonarakis.

Altres formes de descriure variacions

Finalment, cal mencionar que Den Dunnen també dóna per vàlides dues altres formes alternatives per descriure les variacions de seqüència, donat que també compleixen les característiques de ser inequívoques, estables i amb significat, encara que és preferible la descripció que identifica la posició a la seqüència de referència, és a dir, la que s'ha descrit exhaustivament en el present document.

Es tracta del "número rs" (de l'anglès *reference SNP*) que és únic per cada variació (per exemple rs1800652) que es pot consultar a la base de dades SNP del *National Center for Biotechnology Information* (4), i del "número MIM" consultable a la base de dades *Online Mendelian Inheritance in Man* (OMIM) que identifica cada variació descrita a la literatura científica amb un codi numèric característic del gen i de la variació (per exemple MIM *613609.0001).

Els esforços que s'estan fent actualment des del *National Center for Biotechnology Information* (4) per tal de relacionar tota la informació d'aquestes diverses bases de dades entre sí, són imprescindibles per poder unificar internacionalment la descripció de les variacions de seqüència en genètica molecular.

Conclusions i reflexió final

És necessari descriure adequadament les variacions de seqüència i així s'hauria d'exigir a totes les revistes científiques.

Den Dunnen i Antonarakis han publicat una sèrie de recomanacions, fruit del consens científic, amb l'objectiu d'unificar els criteris de descripció de les variacions. Val la pena conèixer-les i aplicar-les tant en els textos científics com als propis informes del laboratori clínic. En aquest sentit, no ens hauríem de paralitzar per la por de perdre la denominació habitual, ja que la major part de vegades no és la recomanada internacionalment i fins i tot pot ser errònia. En tot cas, l'ús simultani d'alguns sinònims sempre pot resultar d'ajuda en aquesta època d'impàs, fins que s'uniformitzi definitivament la descripció de les variacions de seqüència en genètica molecular.

Bibliografia

1. Federació Internacional de Química Clínica, Unió Internacional de Química Pura i Aplicada. Propietats i unitats en les ciències de laboratori clínic. Part XVIII. Propietats i unitats en biologia molecular clínica [Preparat per Soares de Araujo P, Zingales B, Alía-Ramos P, Blanco A, Fuentes-Arderiu X, Mannhalter C, et al.]. In vitro veritas 2004;5, art. 69: <http://www.acclc.cat/> (Accés: 2010-12-4).

2. den Dunnen JT and Antonarakis SE. Nomenclature for the description of sequence variants. Hum Mutat 2000;15:7-12. <http://www.hgvs.org/mutnomen> (Accés: 2010-12-4).

3. Human Genome Nomenclature Database. <http://www.genenames.org/> (Accés: 2010-12-4).

4. National Center for Biotechnology Information. >http://www.ncbi.nlm.nih.gov> (Accés: 2010-12-4).

5. HLA Nomenclature. <http://hla.alleles.org/nomenclature/> (Accés: 2010-12-4).

6. IMGT/HLA Database: International ImMunoGeneTics project for human major histocompatibility complex. <http://www.ebi.ac.uk/imgt/hla/> (Accés: 2010-12-4).

7. Sim SC, Ingelman-Sundberg M. The Human Cytochrome P450 (CYP) Allele Nomenclature website: a peer-reviewed database of CYP variants and their associated effects. Hum Genomics. 2010;4:278-81.

8. Aldehid Dehydrogenase Gene Superfamily Database: <http://www.aldh.org.> (Accés: 2010-12-4).