

# Trabajo práctico especial

*Introducción a la bioinformática*



*Buireo, Juan Martín* 51061

*Menzella, Facundo* 51533

*Noriega, José* 51231

*Purita, Martín* 51187

**25/11/2015**

## Introducción

Para la realización de este trabajo práctico se decidió usar Perl. A pesar de no tener conocimiento del mismo, notamos que era un lenguaje potente pero sencillo de aprender y además vimos que el módulo BioPerl era mucho más completo y simple de utilizar que BioJava. La misma había sido nuestra otra alternativa ya que tenemos mucho conocimiento de Java, pero vimos más compleja su utilización para este trabajo práctico.

Para poder ejecutar el proyecto es necesario tener instalado perl (para poder compilar los scripts) y BioPerl ya que los scripts usan funciones de esta librería.

El proyecto se encuentra en <https://github.com/bahui80/bioinformatica>. Al descargar el proyecto puede observarse la siguiente estructura:

- src: se encuentran todos los scripts para ejecutar las funcionalidades pedidas en la consigna.
- inputs: se encuentran archivos de ejemplo para utilizar como input en los scripts.
- outputs: todas las salidas que generen los scripts se depositarán en esta carpeta.
- docs: se encuentran todos los documentos relacionados al proyecto (consigna e informe).

## Ejercicio 1

Este script lee una secuencia de nucleótidos de un archivo en formato GenBank y los traduce a sus secuencias de aminoácidos posibles teniendo en cuenta los Reading Frames. Dado que una secuencia puede ser leída comenzando desde el primero, segundo o tercer nucleótido comenzando la lectura desde la izquierda o desde la derecha, se obtienen como salida 6 archivos en formato Fasta con las secuencias de aminoácidos posibles.

Para ejecutar el programa basta con situarse en la carpeta src y desde la consola ejecutar:

```
$ perl Ex1.pm INPUT_FILE
```

Donde INPUT\_FILE es un archivo en formato GenBank. Las 6 secuencias posibles de aminoácidos se encontrarán en la carpeta outputs.

## Ejercicio 2

Este script lee una o varias secuencias de aminoácidos de archivos en formato Fasta y realiza un blast de manera remota usando NCBI BLAST. Por eso para ejecutarlo se necesita de una conexión a Internet. Se obtiene como resultado un reporte Blast por cada secuencia de aminoácidos ingresada.

Para ejecutar el programa basta con situarse en la carpeta src y desde la consola ejecutar:

```
$ perl Ex2.pm INPUT_FILE1 [INPUT_FILE2] [INPUT_FILE3] ....
```

Donde INPUT\_FILE<sub>i</sub> es una secuencia de aminoácidos en formato Fasta. El resultado se encontrará en la carpeta outputs. Habrá un resultado distinto por cada archivo ingresado.

## Ejercicio 3

Este script lee un reporte blast y un patrón que se reciben como entrada e identifica aquellos hits donde aparezca dicho pattern y obtiene la secuencia completa del hit en formato Fasta.

Para ejecutar el programa basta con situarse en la carpeta src y desde la consola ejecutar:

```
$ perl Ex3.pm INPUT_FILE PATTERN
```

Donde INPUT\_FILE es un reporte Blast y pattern es un patrón (texto) que filtra aquellos hits en los cuales en su descripción aparece dicho patrón.

## Ejercicio 4

Para poder correr el ejercicio 4 primero hay que instalarse EMBOSS. Para ello desde una terminal en una máquina con Linux hacer:

```
$ sudo apt-get install emboss
```

Una vez instalado EMBOSS, al ejecutar cualquier comando (como puede ser getorf o patmatmotifs) el mismo solicitará la ubicación de los archivos: prosite.dat y prosite.doc para instalar una base de datos local. Los mismos se pueden descargar de: <http://prosite.expasy.org/prosuser.html>. Una vez instalado y configurado EMBOSS, se puede proceder a ejecutar el programa.

Este script lee una secuencia de nucleótidos en formato Fasta o Genbank calcula los ORF (mediante una llamada al sistema al comando getorf). Luego lee cada uno de los ORF y mediante una llamada al sistema al comando patmatmotifs lee la secuencia y realiza un análisis contra los motivos de la base de datos de Prosite.

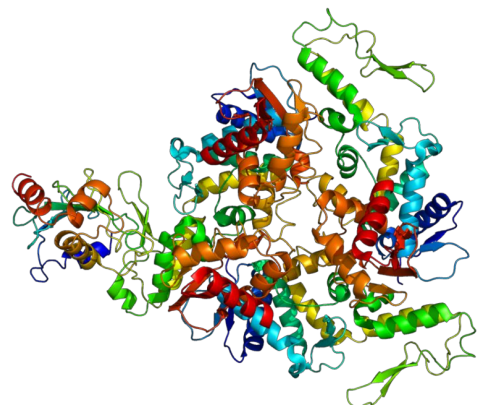
Para ejecutar el programa basta con situarse en la carpeta src y desde la consola ejecutar:

```
$ perl Ex4.pm INPUT_FILE
```

Donde INPUT\_FILE es una secuencia de nucleótidos en formato Fasta o Genbank.

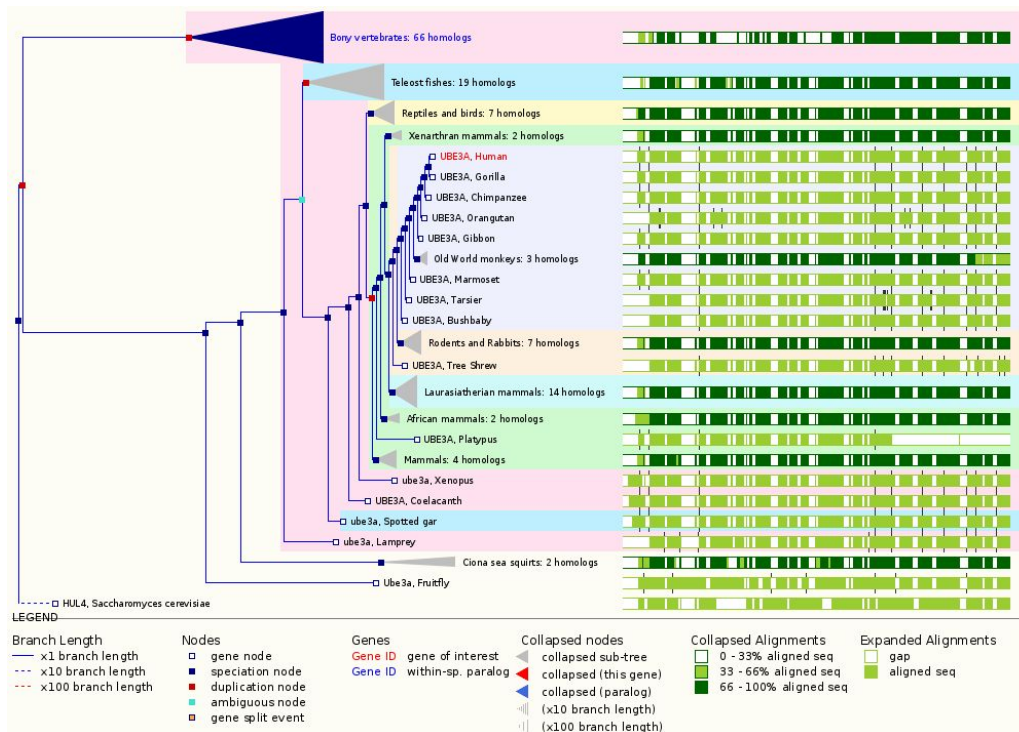
## Ejercicio 5

- a. El gen elegido es UBE3A (<http://www.ncbi.nlm.nih.gov/gene/7337>). Este gen es maternalmente expresado en el cerebro y otros tejidos. La eliminación de este gen causa el Síndrome de Angelman que es una enfermedad que se caracteriza por un retraso en el desarrollo, una capacidad lingüística reducida o nula, escasa receptividad comunicativa, escasa coordinación motriz, con problemas de equilibrio y movimiento, ataxia, estado aparente de alegría permanente, con risas y sonrisas en todo momento. También se pueden mostrar fácilmente excitables, con



hipermotricidad y dificultad de atención. Tiene una incidencia estimada de un caso cada 15.000 a 30.000 nacimientos.

b.















**Imagen 1: Genes homólogos al UBE3A obtenidos de Ensembl**

Analizando el árbol de Ensembl se puede observar que el gen se encuentra homólogo en muchos otros organismos (por ejemplo reptiles y pájaros, roedores y conejos, etc).

Observando la raíz del árbol pudimos ver que el gen UBE3A pertenece al grupo taxonómico de los Animales y Hongos.

## Genes

Genes identified as putative homologs of one another during the construction of HomoloGene.

	UBE3A, <i>H.sapiens</i> ubiquitin protein ligase E3A
	UBE3A, <i>P.troglodytes</i> ubiquitin protein ligase E3A
	UBE3A, <i>M.mulatta</i> ubiquitin protein ligase E3A
	UBE3A, <i>C.lupus</i> ubiquitin protein ligase E3A
	UBE3A, <i>B.taurus</i> ubiquitin protein ligase E3A
	Ube3a, <i>M.musculus</i> ubiquitin protein ligase E3A
	Ube3a, <i>R.norvegicus</i> ubiquitin protein ligase E3A
	UBE3A, <i>G.gallus</i> ubiquitin protein ligase E3A
	ube3a, <i>X.tropicalis</i> ubiquitin protein ligase E3A (human papilloma virus E6-associated protein, Angelman syndrome)
	ube3a, <i>D.rerio</i> ubiquitin protein ligase E3A
	Ube3a, <i>D.melanogaster</i> Ubiquitin protein ligase E3A
	AgaP_AGAP012366, <i>A.gambiae</i> AgaP_AGAP012366

**Imagen 2:** Genes homólogos al UBE3A obtenidos de HomoloGene

Analizando la respuesta de HomoloGene, podemos observar que el gen UBE3A está presente en el Homo Sapiens, los toros, gallos, entre otros.

La principal diferencia que encontramos entre ambas bases de datos (HomoloGene y Ensembl) es que la manera de presentar los resultados en Ensembl es más “usable” ya que muestra los resultados agrupados por taxonomía en forma de árbol el cual se puede ir colapsando o expandiendo. En cambio en HomoloGene se muestran en forma de lista. Otra diferencia que encontramos es que la cantidad de genes homólogos encontrados es mayor en Ensembl que en HomoloGene.

c.



Name	Transcript ID	bp	Protein	Biotype	CCDS	UniProt	RefSeq	Flags
UBE3A-001	<a href="#">ENST00000614096</a>	8741	<a href="#">872aa</a>	Protein coding	<a href="#">CCDS45191</a>	<a href="#">Q05086</a>	<a href="#">NM_130839</a> <a href="#">NP_570854</a>	TSL:1 GENCODE basic APPRIS ALT1
UBE3A-007	<a href="#">ENST00000630424</a>	5075	<a href="#">852aa</a>	Protein coding	<a href="#">CCDS32177</a>	<a href="#">Q9H2G0</a>	-	TSL:5 GENCODE basic APPRIS P3
UBE3A-008	<a href="#">ENST00000232165</a>	5051	<a href="#">852aa</a>	Protein coding	<a href="#">CCDS32177</a>	<a href="#">Q9H2G0</a>	<a href="#">NM_130838</a> <a href="#">NP_570853</a>	TSL:5 GENCODE basic APPRIS P3
UBE3A-002	<a href="#">ENST00000566215</a>	3176	<a href="#">852aa</a>	Protein coding	<a href="#">CCDS32177</a>	<a href="#">Q9H2G0</a>	-	TSL:1 GENCODE basic APPRIS P3
UBE3A-004	<a href="#">ENST00000428984</a>	2986	<a href="#">852aa</a>	Protein coding	<a href="#">CCDS32177</a>	<a href="#">Q9H2G0</a>	-	TSL:5 GENCODE basic APPRIS P3
UBE3A-003	<a href="#">ENST00000397954</a>	2873	<a href="#">875aa</a>	Protein coding	<a href="#">CCDS45192</a>	<a href="#">Q05086</a>	<a href="#">NM_000462</a> <a href="#">NP_000453</a>	TSL:5 GENCODE basic
UBE3A-201	<a href="#">ENST00000438097</a>	2658	<a href="#">852aa</a>	Protein coding	<a href="#">CCDS32177</a>	<a href="#">Q9H2G0</a>	-	TSL:1 GENCODE basic APPRIS P3
UBE3A-005	<a href="#">ENST00000625778</a>	3217	<a href="#">765aa</a>	Protein coding	-	<a href="#">A0A0D9SG77</a>	-	TSL:5 GENCODE basic
UBE3A-009	<a href="#">ENST00000630607</a>	1873	<a href="#">62aa</a>	Protein coding	-	<a href="#">A0A0D9SGJ0</a>	-	CDS 3' incomplete TSL:3
UBE3A-013	<a href="#">ENST00000629252</a>	1112	<a href="#">49aa</a>	Protein coding	-	<a href="#">A0A0D9SFU3</a>	-	CDS 3' incomplete TSL:5
UBE3A-014	<a href="#">ENST00000630907</a>	889	<a href="#">148aa</a>	Protein coding	-	<a href="#">A0A0D9SES7</a>	-	CDS 3' incomplete TSL:5
UBE3A-010	<a href="#">ENST00000628267</a>	779	<a href="#">50aa</a>	Protein coding	-	<a href="#">A0A0D9SGJ1</a>	-	CDS 3' incomplete TSL:5
UBE3A-019	<a href="#">ENST00000626068</a>	772	<a href="#">237aa</a>	Protein coding	-	<a href="#">A0A0D9SGI3</a>	-	CDS 3' incomplete TSL:2
UBE3A-012	<a href="#">ENST00000628733</a>	706	<a href="#">187aa</a>	Protein coding	-	<a href="#">A0A0D9SF91</a>	-	CDS 3' incomplete TSL:5
UBE3A-011	<a href="#">ENST00000625681</a>	654	<a href="#">14aa</a>	Protein coding	-	<a href="#">A0A0D9SFH3</a>	-	CDS 3' incomplete TSL:5
UBE3A-022	<a href="#">ENST00000604860</a>	571	<a href="#">189aa</a>	Protein coding	-	<a href="#">S4R306</a>	-	CDS 5' incomplete TSL:5
UBE3A-023	<a href="#">ENST00000626176</a>	445	<a href="#">143aa</a>	Protein coding	-	<a href="#">A0A0D9SG63</a>	-	CDS 5' incomplete TSL:3
UBE3A-016	<a href="#">ENST00000629886</a>	436	<a href="#">77aa</a>	Protein coding	-	<a href="#">A0A0D9SG54</a>	-	CDS 3' incomplete TSL:3
UBE3A-018	<a href="#">ENST00000628890</a>	424	<a href="#">40aa</a>	Protein coding	-	<a href="#">A0A0D9SEJ2</a>	-	CDS 3' incomplete TSL:3
UBE3A-017	<a href="#">ENST00000626589</a>	394	<a href="#">8aa</a>	Protein coding	-	<a href="#">A0A0G2JQQ5</a>	-	CDS 3' incomplete TSL:5
UBE3A-021	<a href="#">ENST00000631247</a>	789	No protein	Processed transcript	-	-	-	TSL:5
UBE3A-020	<a href="#">ENST00000626793</a>	733	No protein	Processed transcript	-	-	-	TSL:5
UBE3A-015	<a href="#">ENST00000627018</a>	474	No protein	Processed transcript	-	-	-	TSL:3

**Imagen 3: Splice variants del gen UBE3A obtenidos de Ensembl**

Se pudo determinar que en Ensembl existen **23** splice variants para el gen **UBE3A**.

Se pudo determinar que en NCBI existen **98** formas alternativas de splicing para el gen **UBE3A**

#### Search results

Items: 1 to 20 of 98

Filters activated: Alternatively spliced. [Clear all](#) to show 179 items.

Showing Current items.

<< First < Prev Page 1 of 5 Next > Last >>

Name/Gene ID	Description	Location	Aliases	MIM
<input type="checkbox"/> <a href="#">UBE3A</a> ID: 7337	ubiquitin protein ligase E3A [ <i>Homo sapiens</i> (human)]	Chromosome 15, NC_000015.10 (25337234..25439043, complement)	ANCR, AS, E6-AP, EPVE6AP, HPVE6A	601623
<input type="checkbox"/> <a href="#">Ube3a</a> ID: 22215	ubiquitin protein ligase E3A [ <i>Mus musculus</i> (house mouse)]	Chromosome 7, NC_000073.6 (59227814..59311536)	4732496B02, 5830462N02Rik, A130086L21Rik, Hpve6a, mKIAA4216	
<input type="checkbox"/> <a href="#">Ube3a</a> ID: 361585	ubiquitin protein ligase E3A [ <i>Rattus norvegicus</i> (Norway rat)]	Chromosome 1, NC_005100.4 (116586906..116678161)		
<input type="checkbox"/> <a href="#">UBE3A</a> ID: 418686	ubiquitin protein ligase E3A [ <i>Gallus gallus</i> (chicken)]	Chromosome 1, NC_006088.3 (131234215..131283481)		
<input type="checkbox"/> <a href="#">ube3a</a> ID: 407881	ubiquitin protein ligase E3A (human papilloma virus E6-associated protein, Angelman syndrome) [ <i>Xenopus (Silurana) tropicalis</i> (western clawed frog)]			
<input type="checkbox"/> <a href="#">UBE3A</a> ID: 533136	ubiquitin protein ligase E3A [ <i>Bos taurus</i> (cattle)]	Chromosome 21, AC_000178.1 (2346808..2444194, complement)		
<input type="checkbox"/> <a href="#">UBE3A</a> ID: 711270	ubiquitin protein ligase E3A [ <i>Macaca mulatta</i> (Rhesus monkey)]	Chromosome 7, NC_007864.1 (4272934..4373144, complement)		
<input type="checkbox"/> <a href="#">UBE3A</a> ID: 479010	ubiquitin protein ligase E3A [ <i>Canis lupus familiaris</i> (dog)]	Chromosome 3, NC_006585.3 (35346640..35442170)		

**Imagen 4: Splice variants del gen UBE3A obtenidos de NCBI**

Suponemos que el número correcto de splicing es el que proporciona NCBI dado que ENSEMBL maneja un grupo acotado de especies es su base de datos.

- d. Utilizando la información proporcionada por UniProt podemos ver que el producto génico de nuestro gen interactúa con 147 proteínas de la siguiente manera:

#### Protein-protein interaction databases

BioGrid <sup>i</sup>	<a href="#">113185</a> . 147 interactions.
DIP <sup>i</sup>	<a href="#">DIP-6002N</a> .
IntAct <sup>i</sup>	<a href="#">Q05086</a> . 48 interactions.
MINT <sup>i</sup>	<a href="#">MINT-147444</a> .
STRING <sup>i</sup>	<a href="#">9606.ENSPP00000381045</a> .

**Imagen 5:** Interacciones con otras proteínas del gen UBE3A obtenidos de UniProt

Utilizando la información proporcionada por NCBI podemos ver que el producto génico de nuestro gen interactúa con 163 proteínas de la siguiente manera:

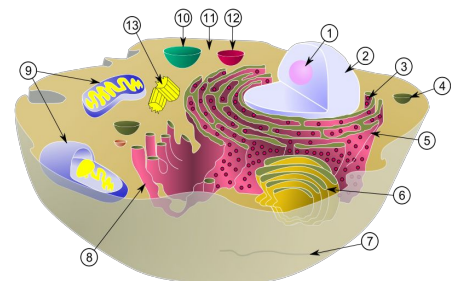
Interactions						
Products	Interactant	Other Gene	Complex	Source	Items 1 - 25 of 163	
NP_000453.2	<a href="#">NP_003338.1</a>	<a href="#">UBE2L3</a>		<a href="#">BIND</a>	<a href="#">PubMed</a>	UbcH7 interacts with E6AP.
NP_000453.2	<a href="#">NP_004214.1</a>	<a href="#">UBE2L6</a>		<a href="#">BIND</a>	<a href="#">PubMed</a>	UbcH8 interacts with E6AP.
Q05086	<a href="#">P10275</a>	<a href="#">AR</a>		<a href="#">HPRD</a>	<a href="#">PubMed</a>	
Q05086	<a href="#">P51451</a>	<a href="#">BLK</a>		<a href="#">HPRD</a>	<a href="#">PubMed</a>	
Q05086	Basic charge Y linked 2B	<a href="#">BPY2B</a>		<a href="#">HPRD</a>	<a href="#">PubMed</a>	
Q05086	Basic charge, Y-linked, 2C	<a href="#">BPY2C</a>		<a href="#">HPRD</a>	<a href="#">PubMed</a>	
Q05086	<a href="#">O14599</a>	<a href="#">BPY2C</a>		<a href="#">HPRD</a>	<a href="#">PubMed</a>	
Q05086	<a href="#">Q56P03</a>	<a href="#">EAPP</a>		<a href="#">HPRD</a>	<a href="#">PubMed</a>	
Q05086	<a href="#">Q6IE81</a>	<a href="#">JADE1</a>		<a href="#">HPRD</a>	<a href="#">PubMed</a>	
Q05086	<a href="#">P33993</a>	<a href="#">MCM7</a>		<a href="#">HPRD</a>	<a href="#">PubMed</a>	
Q05086	<a href="#">P08401</a>	<a href="#">PGR</a>		<a href="#">HPRD</a>	<a href="#">PubMed</a>	
Q05086	<a href="#">P10276</a>	<a href="#">RARA</a>		<a href="#">HPRD</a>	<a href="#">PubMed</a>	
Q05086	<a href="#">O15126</a>	<a href="#">SCAMP1</a>		<a href="#">HPRD</a>	<a href="#">PubMed</a>	
Q05086	<a href="#">P04278</a>	<a href="#">SHBG</a>		<a href="#">HPRD</a>	<a href="#">PubMed</a>	
Q05086	<a href="#">O15198</a>	<a href="#">SMAD9</a>		<a href="#">HPRD</a>	<a href="#">PubMed</a>	
Q05086	<a href="#">P04637</a>	<a href="#">TP53</a>		<a href="#">HPRD</a>	<a href="#">PubMed</a>	
Q05086	<a href="#">P51668</a>	<a href="#">UBE2D1</a>		<a href="#">HPRD</a>	<a href="#">PubMed</a>	
Q05086	<a href="#">P61077</a>	<a href="#">UBE2D3</a>		<a href="#">HPRD</a>	<a href="#">PubMed</a>	
Q05086	<a href="#">P51965</a>	<a href="#">UBE2E1</a>		<a href="#">HPRD</a>	<a href="#">PubMed</a>	
Q05086	<a href="#">P62253</a>	<a href="#">UBE2G1</a>		<a href="#">HPRD</a>	<a href="#">PubMed</a>	
Q05086	<a href="#">P60604</a>	<a href="#">UBE2G2</a>		<a href="#">HPRD</a>	<a href="#">PubMed</a>	
Q05086	<a href="#">P68036</a>	<a href="#">UBE2L3</a>		<a href="#">HPRD</a>	<a href="#">PubMed</a>	
Q05086	<a href="#">O14933</a>	<a href="#">UBE2L6</a>		<a href="#">HPRD</a>	<a href="#">PubMed</a>	
Q05086	<a href="#">Q05086</a>	<a href="#">UBE3A</a>		<a href="#">HPRD</a>	<a href="#">PubMed</a>	
Q05086	<a href="#">Q9UHD9</a>	<a href="#">UBQLN2</a>		<a href="#">HPRD</a>	<a href="#">PubMed</a>	

**Imagen 6:** Interacciones con otras proteínas del gen UBE3A obtenidos de NCBI

- e. Componente celular del cual forma parte nuestra proteína:

- Matriz Citoplasmática (Cytosol): disolución gelatinosa, rica en agua. En el caso de ser eucariota, se encuentra delimitada por la membrana celular y membrana nuclear. De lo contrario (procariota), no tendría membrana nuclear.

Más precisamente, lo podemos encontrar dentro del Proteasoma, presente en las células eucariotas y Archaeas



(bacterias). Este complejo proteico se encarga de realizar la degradación de las proteínas denominada proteólisis.

Procesos biológicos a los que pertenece:

- Androgen Receptor Signaling Pathway
- Brain Development
- Positive Regulation of Phosphatidylinositol 3-kinase Cascade
- Positive Regulation of Transcription From RNA Polymerase II Promoter
- Protein K48-linked Ubiquitination
- Protein Modification Process
- Ubiquitin-dependent Protein Catabolic Process

Función molecular en que trabaja esta proteína:

- Síntesis de péptidos (Amino Acid Ligase Activity)
- Unión proteica
- Actividad de coactivador de transcripción
- Ubiquitin-protein Ligase Activity

f. Pathways:

- Ubiquitin mediated proteolysis  
(ver [http://www.kegg.jp/kegg-bin/show\\_pathway?hsa04120](http://www.kegg.jp/kegg-bin/show_pathway?hsa04120))
- Viral carcinogénesis  
(ver [http://www.kegg.jp/kegg-bin/show\\_pathway?hsa05203](http://www.kegg.jp/kegg-bin/show_pathway?hsa05203))

- g. La variación RS111033595, es un polimorfismo de un solo nucleótido que genera la aparición de un codón de stop en el gen **UBE3A**, causando el *Síndrome de Angelman*. Se genera una mutación sin sentido en el exón 15, en el cual se sustituye una **G** por una **A** en el nucleótido 2304:

TATTTTTTCTCATTAGGGAGTTCTG**[A/G]**GAAATCGTTCATTCATTTACAGATG

Lo cual resulta un codón de stop prematuro, generando una transcripción más corta. Solo el 10% de los casos de Angelman Syndrome son causadas por mutaciones en el gen **UBE3A**. Sin embargo no se encontraron frecuencias poblacionales para esta variante en particular. El paper "*Prenatal Diagnosis and Carrier Detection for a Point Mutation in UBE3A Causing Angelman Syndrome*" de Tsai TF, menciona que se realizaron estudios prenatales en una mujer embarazada de ascendencia Judía Asquenazi e Iraquí, la cual ya tenía dos hijos con Síndrome de Angelman. En dicho trabajo se demuestra el SNP mencionado.