

## Exercici 1. Estratègies d'alineament.

### 1. Alineament global de les seqüències CDS del gen TMEM106B (humà) i el gen tmem106b (ratolí) amb CLUSTAL.

Obtenim les dues seqüències CDS amb el navegador GenomeBrowser de UCSC (veure seqüències obtingudes a [annex1a](#)). Observem que existeixen dues variants del gen TMEM106B humà (isoformes):

```
>hg38_refGene_NM_018374 range=chr7:12214811-12231975 5'pad=0 3'pad=0 strand=+
repeatMasking=none
>hg38_refGene_NM_001134232 range=chr7:12214811-12231975 5'pad=0 3'pad=0 strand=+
repeatMasking=none
```

#### Sequence Retrieval Region Options:

- ☐ Promoter/Upstream by 5000 bases
  - ☐ 5' UTR Exons
  - ☒ CDS Exons
  - ☐ 3' UTR Exons
  - ☐ Introns
  - ☐ Downstream by 1000 bases
  - ☒ One FASTA record per gene.
  - ☐ One FASTA record per region (exon, intron, etc.) with
  - ☐ Split UTR and CDS parts of an exon into separate records
- Note: if a feature is close to the beginning or end of a chromosome, the region may be truncated.

CLUSTAL permet fer l'alineament global de dos o més seqüències. Efectua la correspondència entre les seqüències completes, maximitzant el número total de caràcters coincidents. Aquest alineament global està indicat per efectuar comparacions entre seqüències que a priori codifiquen elements amb una funció biològica semblant i que, possiblement, tenen una mida i estructura similars. Per exemple, com en el nostre cas, el mateix gen en dues espècies diferents, o proteïnes homòlogues. En funció de la distància evolutiva entre les dues espècies comparades, trobarem més o menys similitud.

Si comparem les 3 seqüències:

Percent Identity Matrix - created by Clustal2.1

1: hg38_refGene_NM_018374	100.00	100.00	88.85
2: hg38_refGene_NM_001134232	100.00	100.00	88.85
3: mm10_refGene_NM_027992	88.85	88.85	100.00

S'observa un 100% d'identitat entre les dues isoformes del gen humà, i un 88.85% d'identitat entre qualsevol de les dues humanes amb la de ratolí (Veure resultat alineament complet en [annex 1b](#)):

Percent Identity Matrix - created by Clustal2.1

1: hg38_refGene_NM_018374	100.00	88.85
2: mm10_refGene_NM_027992	88.85	100.00

### 2. Mateix alineament global de les proteïnes amb CLUSTAL. Grau d'homologia entre seqüències.

Obtenim les seqüències de la proteïna amb "Predicted Protein" a GenomeBrowser d'UCSC (veure seqüències obtingudes en [annex2](#)):

#### Links to sequence:

- [Predicted Protein](#)
- [mRNA Sequence](#) (may be different from the genomic sequence)
- [Genomic Sequence](#) from assembly
- [CDS FASTA alignment](#) from multiple alignment

I tornem a realitzar l'alineament amb CLUSTAL:

CLUSTAL O(1.2.4) multiple sequence alignment

```
NP_060844      MGKSLSHLPLHSSKEDAYDGVTS - ENMRNGLVNSEVHNEDGRNGDVSQFPYVEFTGRDSV      59
NP_082268      MGKSLSHLPLHSNKEDGYDGVTS TDNMRNGLVSSEVHNEDGRNGDVSQFPYVEFTGRDSV      60
                ***** ,*** ,***** :***** ,*****

NP_060844      TCPTCQGTGRIPRGQENQLVALIPYDQRLRPRRTKLYVMASVFVCLLLSGLAVFFLFPR      119
NP_082268      TCPTCQGTGRIPRGQENQLVALIPYDQRLRPRRTKLYVMASVFVCLLLSGLAVFFLFPR      120
                *****

NP_060844      SIDVKYIGVKSAYVSYDVQKRTIYLNITNTLNITNNNYSVEVENITAQVQFSKTVIGKA      179
NP_082268      SIEVKYIGVKSAYVSYDAEKRTIYLNITNTLNITNNNYSVEVENITAQVQFSKTVIGKA      180
                ** ,***** ,*****

NP_060844      RLNNITIIGPLDMKQIDYTVPTVIAEEMSYMYDFCTLISIKVHNIVLMMQVTVTTTYFGH      239
NP_082268      RLNNITNIGPLDMKQIDYTVPTVIAEEMSYMYDFCTLISIKVHNIVLMMQVTVTTAYFGH      240
                ***** ,***** ,*****

NP_060844      SEQISQERYQYVDCGRNTTYQLGQSEYLNVLQPQQ      274
NP_082268      SEQISQERYQYVDCGRNTTYQLAQSEYLNVLQPQQ      275
                ***** ,*****
```

El grau d'homologia entre les dues seqüències de la proteïna es d'un 95.99%.

Percent Identity Matrix - created by Clustal2.1

```
1: NP_060844      100.00   95.99
2: NP_082268      95.99  100.00
```

Com hem dit abans, el gen TMEM106B humà té dues isoformes, però les proteïnes resultants són iguals:

Percent Identity Matrix - created by Clustal2.1

```
1: NP_001127704   100.00  100.00
2: NP_060844      100.00  100.00
```

Per tant, la homologia amb tmem106b és la mateixa:

Percent Identity Matrix - created by Clustal2.1

```
1: NP_001127704   100.00   95.99
2: NP_082268      95.99  100.00
```

L'alt percentatge d'identitat ens indica que existeix una gran homologia entre les dues proteïnes, com és d'esperar quan es fa un alineament global entre proteïnes que resulten del mateix gen de dues espècies diferents. La semblança serà més evident si existeix menor distància evolutiva, per exemple en aquest cas, on tots dos organismes són mamífers. L'augment d'identitat entre proteïnes respecte a la seqüència CDS és deguda a que hi ha canvis evolutius als gens (mutacions puntuals) que no són rellevants per la funció de la proteïna. Un exemple serien les mutacions silencioses, al tercer nucleòtid del codó. A més, existeixen aa amb propietats similars que es poden intercanviar dins de la proteïna sense alterar el producte final.

### 3. Alineament local amb BLAST de les dues seqüències CDS.

BLAST fa servir un “diccionari” de paraules per identificar quines seqüències de la seva base de dades seran més similars a la nostra. En el problema que se ns planteja, triarem l’opció *blastn*, que serveix per alinear seqüències de nucleòtids curtes i fer comparacions entre espècies, com és el nostre cas.

Table 1. Key features of the BLAST search pages in the “Basic BLAST” category

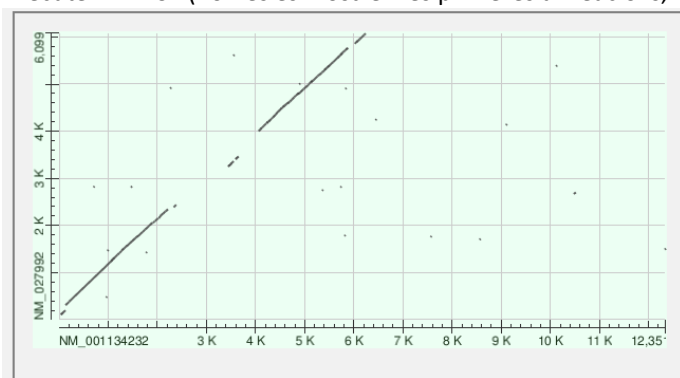
Search page	Query & database	Alignment	Programs & functions (default program in bold)
nucleotide blast	nucleotide vs nucleotide	Nucleotide	<b>megablast</b> : for sequence identification, intra-species comparison <b>discontiguous megablast</b> : for cross-species comparison, searching with coding sequences <b>blastn</b> : for searching with shorter queries, cross-species comparison
protein blast	Protein vs protein	protein	<b>blastp</b> : general sequence identification and similarity searches <b>Quick BLASTP</b> : with a kmer match to accelerate search speed for very similar proteins <b>DELTA-BLAST</b> [3] : protein similarity search with higher sensitivity than blastp <b>PSI-BLAST</b> : iterative search for position-specific score matrix (PSSM) to identify distant relatives for a protein family <b>PHI-BLAST</b> : protein alignment with input pattern as anchor/constraint
blastx	nucleotide (tr) vs protein	protein	<b>blastx</b> : for identifying potential protein products encoded by a nucleotide query
tblastn	protein vs nucleotide (tr)	protein	<b>tblastn</b> : for identifying database sequences encoding proteins similar to the query
tblastx	nucleotide (tr) vs nucleotide (tr)	protein	<b>tblastx</b> : for identifying nucleotide sequences similar to the query based on their coding potential

Marquem “Align two or more sequences” per fer la comparació. Cambiem el “Program Selection” per defecte a *blastn*, i deixem els “Algorithm parameters” que hi ha per defecte:

The screenshot shows the BLAST search interface with the following details:

- blastn** tab is selected.
- Enter Query Sequence**:
  - Input: NM\_001134232
  - Buttons: Clear, Query subrange (From, To)
- Or, upload file**:
  - Seleccionar archivo: Ningún archivo seleccionado
- Job Title**:
  - Input: NM\_001134232: Homo sapiens transmembrane protein...
  - Input: Enter a descriptive title for your BLAST search
- Align two or more sequences**: ☒
- Enter Subject Sequence**:
  - Input: NM\_027992
  - Buttons: Clear, Subject subrange (From, To)
- Or, upload file**:
  - Seleccionar archivo: Ningún archivo seleccionado
- Program Selection**:
  - Optimize for:
    - ☐ Highly similar sequences (megablast)
    - ☐ More dissimilar sequences (discontiguous megablast)
    - ☒ Somewhat similar sequences (blastn)

Executem BLAST (només es mostren les primeres alineacions, veure arxiu complet en [annex3](#)):



RID: B5G0BE69114  
 Job Title: NM\_001134232: Homo sapiens transmembrane protein...  
 Program: BLASTN  
 Subject: Mus musculus transmembrane protein 106B (Tmem106b), mRNA ID: NM\_027992.3 (nucleic acid) Length: 6099  
 Query #1: Homo sapiens transmembrane protein 106B (TMEM106B), transcript variant 2, mRNA Query ID: ref|NM\_001134232.2 Length: 12351

Sequences producing significant alignments:

Description	Max Score	Total Score	Query cover	E Value	Per. Ident	Accession
Mus musculus transmembrane protein 106B (Tmem106b), mRNA	1501	2771	37%	0.0	76.41	NM_027992.3

Alignments:

>Mus musculus transmembrane protein 106B (Tmem106b), mRNA  
 Sequence ID: NM\_027992.3 Length: 6099  
 Range 1: 310 to 2346

Score: 1501 bits (1664), Expect: 0.0,  
 Identities: 1623/2124 (76%), Gaps: 126/2124 (5%), Strand: Plus/Plus

```

Query 133  ACATGGGAAAGTCTCTTTCTCATTGCTTTGCATTCAAGCAAAGAAGTCTTATGATG 192
             |||||
Sbjct 310  ACATGGGAAAGTCTCTTTCTCATTGCTTTGCATTCAAAATAAGAAGTGGCTATGATG 369

Query 193  GAGTCACATCT---GAAACATGAGGAATGGACTGGTTAATAGTGAAGTCCATAATGAAG 249
             || ||||| || ||||| ||||| ||||| ||||| || |||||
Sbjct 370  GCGTTACATCGACAGACAATATGAGAAATGGATTGGTTAGCAGTGAAGTGCACAACGAAG 429
  
```

Observem que hi ha un 76% d'identitat. Si realitzem la mateixa alineació però amb l'altre variant del gen, obtenim resultats semblants (veure annex 2):

>Mus musculus transmembrane protein 106B (Tmem106b), mRNA

Sequence ID: NM\_027992.3 Length: 6099

Score: 1508 bits (1671), Expect: 0.0,

Identities: 1628/2130 (76%), Gaps: 126/2130 (5%), Strand: Plus/Plus

[Download](#)
[GenBank](#)
[Graphics](#)
Sort by: E value

**Mus musculus transmembrane protein 106B (Tmem106b), mRNA**

Sequence ID: [NM\\_027992.3](#) Length: 6099 Number of Matches: 27

Range 1: 304 to 2346 [GenBank](#) [Graphics](#) [Next Match](#) [Previous Match](#)

Score	Expect	Identities	Gaps	Strand
1508 bits (1671)	0.0	1628/2130 (76%)	126/2130 (5%)	Plus/Plus

```

CDS:transmembrane pr 1      M G K S L S H L P L H S S K E D A
Query 321  CCTCAGACATGGGAAAGTCTCTTTCTCATTGCTTTGCATTCAAGCAAAGAAGTCTT 380
             |||||
Sbjct 304  CCTCAAAACATGGGAAAGTCTCTTTCTCATTGCTTTGCATTCAAAATAAGAAGTGGCT 363
CDS:transmembrane pr 1      M G K S L S H L P L H S S K E D G
             |||||

CDS:transmembrane pr 18     Y D G V T S E N M R N G L V N S E V H
Query 381  ATGATGGAGTCACATCT---GAAACATGAGGAATGGACTGGTTAATAGTGAAGTCCATA 437
             |||||
Sbjct 364  ATGATGGCGTTACATCGACAGACAATATGAGAAATGGATTGGTTAGCAGTGAAGTGCACA 423
CDS:transmembrane pr 18     Y D G V T S T D N M R N G L V S S E V H
             |||||

CDS:transmembrane pr 37     N E D G R N G D V S Q F P Y V E F T G R
Query 438  ATGAAGATGGAAGAAATGGAGATGTCTCTCAGTTTCCATATGTGGAAATTTACAGGAAGAG 497
             |||||
Sbjct 424  ACGAAGACGGAAGAAATGGAGATGTCTCTCAGTTCCCATATGTGGAAATTTACTGGAAGAG 483
CDS:transmembrane pr 38     N E D G R N G D V S Q F P Y V E F T G R
  
```

Blastn ens proporciona diferents alineaments, amb diferents puntuacions o *scores*. És un alineament local que realitza exclusivament la correspondència entre aquells fragments de les seqüències que tenen coincidència màxima de caràcters, descartant la resta de regions que no presenten mínima similitud. Així com sempre és possible obtenir un alineament global, és poc freqüent identificar alineaments locals si no existeix certa conservació funcional. Com hem dit abans, per comparar un mateix gen o proteïnes homòlogues entre diferents espècies, és millor utilitzar un programa d'alineament global com, per exemple, CLUSTAL.

#### 4. Alineament global amb CLUSTAL de genomicA.txt i genomicB.txt.

En aquestes dues seqüències les lletres minúscules representen la regió no traduïble, i les majúscules ens indiquen els fragments codificants, que comencen per ATG. Alineem globalment totes dues seqüències amb CLUSTALO, i obtenim el següent resultat

Percent Identity Matrix - created by Clustal2.1

1: genomicA	100.00	44.09
2: genomicB	44.09	100.00

(veure alineament complet en [annex4](#)).

Observem que el percentatge d'identitat és baix, i que hi ha coincidència de regions codificants d'una seqüència amb no codificants de l'altre.

```

genomicA      aagctcttggccgggtgcagtggtcatgcctgtaatcccATGGGAAAGTCTCTTTCTCA 420
genomicB      GCATCat--gggaggagctgtctctaagatctctaagtgactttgaggccttttgctca 303
              *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *

genomicA      TTTGCCTTTGCATTCAAGCAAAGAAGATGCagttccccatttctgtcgccacaccttga 480
genomicB      ttgtcttggatattagccctt-----ggcaccccttttagtcacgctaatacccccta 354
              ** * *  ***  *  *  ***  *  *  ***  *  *  *  *  *  *  *

```

## 5. Alineament local amb BLAST de genomicA.txt i genomicB.txt.

En aquest cas, farem servir la funció *megablast*, que està recomanada per comparacions intraespecífiques. Es tracta d'un alineament múltiple local que efectua únicament la correspondència entre aquells fragments que tenen una coincidència relevant, descartant la resta de regions.

Com és un alineament local, el BLAST troba dues coincidències, amb una identitat del 100% entre A i B (veure [annex5](#)):

<a href="#">Download</a>	<a href="#">Graphics</a>	Sort by: <a href="#">E value</a>
<b>genomicB</b>		
Sequence ID: Query_63835 Length: 800 Number of Matches: 2		
Range 1: 201 to 251 <a href="#">Graphics</a>		<a href="#">Next Match</a> <a href="#">Previous Match</a>
Score	Expect	Identities
95.3 bits(51)	2e-23	51/51(100%)
Gaps	Strand	
0/51(0%)	Plus/Plus	
Query 751	ATGAGTGTGGATCCAGCTTGTCCCAAGCTTGCCTTGCTTTGAAGCATCA	801
Sbjct 201	ATGAGTGTGGATCCAGCTTGTCCCAAGCTTGCCTTGCTTTGAAGCATCA	251
Range 2: 701 to 750 <a href="#">Graphics</a>		<a href="#">Next Match</a> <a href="#">Previous Match</a> <a href="#">First Match</a>
Score	Expect	Identities
93.5 bits(50)	6e-23	50/50(100%)
Gaps	Strand	
0/50(0%)	Plus/Plus	
Query 401	ATGGGAAAGTCTCTTTCTCATTGCTTTGCATTCAAGCAAAGAAGATGC	450
Sbjct 701	ATGGGAAAGTCTCTTTCTCATTGCTTTGCATTCAAGCAAAGAAGATGC	750

## 6. Quin programa és més adequat per cada comparació?

El programa CLUSTAL fa alineaments globals, per tant és més addient per fer comparacions de seqüències que codifiquen elements que a priori tenen una funció biològica semblant i que, per tant, tenen una mida i estructura similars. Per exemple, les comparacions inter-específiques d'un mateix gen, o seqüències d'aa. de dues proteïnes homòlogues.

En canvi, BLAST fa alineaments locals, per tant estaria indicat per comparar seqüències on el seu contingut, en general, no està altament conservat, com les regions reguladores o dominis funcionals específics en proteïnes. En canvi, l'alineament local realitza exclusivament correspondències entre aquells fragments de les seqüències que tenen coincidència màxima, descartant les altres regions. L'alineament local es addient per identificar elements de petit tamany conservats dins un context general divers. Aquests elements possiblement estan relacionats amb la configuració de dominis de proteïnes o llocs d'unió a factors de transcripció (TFs).

## 7. Alineament d'una seqüència de nucleòtids de pollastre, *genomeC*, amb una seqüència de proteïna de TMEM106B humana.

En aquest cas, fem servir blastx, que identifica proteïnes potencials, codificades dins d'una seqüència "consulta" de nucleòtids. Processa totes les possibles proteïnes codificades dintre de la seqüència problema, fent servir una matriu de substitució per recuperar aquells aa que produeixen un canvi sinònim (veure [annex6](#), per resultat complet de l'alineament. Només s'han inclòs els *scores* més alts):

Align Sequences Translated BLAST: blastx

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

Clear Query subrange

From To

Or, upload file

Seleccionar archivo Ningún archivo seleccionado

Genetic code

Standard (1)

Job Title

genomeC

Enter a descriptive title for your BLAST search

☒ Align two or more sequences

Enter Subject Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

Clear Subject subrange

From To

Or, upload file

Seleccionar archivo Ningún archivo seleccionado

**BLAST** Search protein sequence using Blastx (search protein subjects using a translated nucleotide query)

Trobem que genomeC té un 85.14% de coincidència amb NP\_001127704 (amb un E value significatiu), que és una de les isoformes de TMEM106B en humans:

Sequences producing significant alignments							Download	Manage Columns	Show	100	?
<input checked="" type="checkbox"/> select all	1 sequences selected										Graphics
	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession				
<input checked="" type="checkbox"/>	NP_001127704 length=274	135	606	5%	9e-39	85.14%	Query_64215				

Les regions que tenen un alt percentatge d'identitat, estarien implicades en funcions biològiques comuns entre les dues espècies, que s'han conservat al llarg de l'evolució. Aquesta identitat correspon a una possible seqüència ancestral comú. Aquesta simil.litud és més gran que si haguèssim comparat les dues seqüències de nucleòtids, ja que, com hem dit anteriorment, blastx té en compte els canvis sinònims(que no impliquen canvi d'aa).

## 8. Programa MEME. Estudi de la regulació gènica mitjançant factors de transcripció.

MEME descobreix *motius* o patrons de seqüència que es produeixen repetidament en un grup de seqüències. Representa una alternativa a altres aplicacions que busquen llocs d'unió a factors de transcripció. Es basa en la identificació de motius conservats en grups de seqüències: busca iteracions dins de la seqüència i després les compara amb catàlegs d'informació coneguda. MEME representa aquests motius com a matrius que descriuen la probabilitat de cada lletra possible en cada posició del patró. Els motius MEME individuals no contenen *gaps*. MEME divideix els patrons de longitud variable en dos o més motius separats, i tracta tots els caràcters ambigus en seqüències com comodins que coincideixen amb totes les lletres.

Es prenen com a *input* un grup de seqüències que hipotèticament comparteixen algun mecanisme de regulació i com a *output* tants motius com se sol·liciti. Utilitza tècniques estadístiques per triar automàticament l'amplada, el nombre de casos, i la millor descripció per a cada motiu. També podem afegir un segon *input*(control) de seqüències i descobrir motius del conjunt primari en relació al conjunt de control. Això s'anomena *descobrimet de motius discriminatoris*.

Bailey i Elkan, a "*Fitting a mixture model by expectation maximization to discover motifs in biopolymers*" (California, 1994) expliquen que MEME ha implementat l'algoritme MM per descobrir múltiples motius en seqüències de DNA. A partir d'una sèrie de seqüències de biopolímers no alineats i possiblement relacionats, estima els paràmetres d'un model probabilístic mixte de dos components. Un component descriu una sèrie de subseqüències similars de determinada amplitud (el "*motiu*"), i l'altre component descriu totes les altres posicions en les seqüències (el "*background*"). Ajustar el model inclou estimar la freqüència relativa de motius, i aquesta freqüència determina el llindar per a una òptima classificació bayesiana que podrà ser emprada per trobar motius en altres dades. L'algoritme només requereix una sèrie de seqüències no alineades i especificar l'amplada del

motius. Aleshores, estima quin nombre de cops apareix cada motiu en cada seqüència i ens retorna un alineament d'aquestes repeticions. És capaç de descobrir diferents motius amb diferents freqüències d'aparició en una única base de dades.

## 9. Regulació transcripcional de TMEM106B al llarg de l'evolució. Regió promotora del gen.

Les regions promotores d'un gen són "interruptors" que s'accionen per diferents TFs. Els TFs són proteïnes amb un domini per detectar llocs d'unió al DNA. Aquests llocs d'unió són seqüències de 5-15 nucleòtids, que són reconeguts normalment per més d'un TF. Les regions reguladores poden estar ubicades immediatament abans del gen (promotor basal), o a uns pocs milers de bases (promotor proximal). Obtenim la regió promotora de TMEM106B humà (hg38) emprant les anotacions de RefSeq en el navegador d'UCSC (5000 nucleòtids upstream):

**Sequence Retrieval Region Options:**

- ☒ Promoter/Upstream by 5000 bases
- ☐ 5' UTR Exons
- ☐ CDS Exons
- ☐ 3' UTR Exons
- ☐ Introns
- ☐ Downstream by 1000 bases
- ☐ One FASTA record per gene.
- ☒ One FASTA record per region (exon, intron, etc.) w
- ☐ Split UTR and CDS parts of an exon into separ

Note: if a feature is close to the beginning or end of a

(Veure regions promotores que hem obtingut a [annex7](#))

```
>hg38_refGene_NM_018374 range=chr7:12206294-12211293 5'pad=0 3'pad=0 strand=+ repeatMasking=none

tttttttatttgcgtttccctaggtttagtgttctttaaatttgttttt
tgttttgttttgttttaggattcactgaattcttgaatatatggattt

Fem el mateix amb ratolí mm10:

>mm10_refGene_NM_027992 range=chr6:13064759-13069758 5'pad=0 3'pad=0 strand=+ repeatMasking=none
tgtgtgtatgtgtttgtatgtgtgtatgtgtgtatgtgtgtgtgtatg
tgtgtgtatgtgtgtatgtgtgtgtgtatgtgtgtttatgtgtgtgtg

Amb rata rn6:


>rn6_refGene_NM_001004267 range=chr4:39512679-39517678 5'pad=0 3'pad=0 strand=+ repeatMasking=none
gccaggaacaccagaagcagagtttgagcaggcaggggcagagagctga
ggcaaaagacaccagaatgaaagcaagaaaggaactgggtgaaagtatgt

i amb pollastre galgal6:

>galGal6_refGene_NM_001012558 range=chr2:26565971-26570970 5'pad=0 3'pad=0 strand=+ repeatMasking=none
cagctgcaattgctgctatgtttcttaggtaaagacaagtggtttggaat
gtgatttgaataaccgtgttctgaaagaagctattcagaagcattaaac
```

## 10. MEME. Busquem en les 4 seqüències ortòlogues anteriors els 10 millors motius de 5-15pb.

Seleccionem el mode clàssic, ja que el altres dos mètodes necessiten un control del què nosaltres no disposem:

Select the motif discovery mode 

**Classic mode**  
You provide one set of sequences and MEME discovers motifs enriched in this set. Enrichment is measured relative to a (higher order) random model based on frequencies of the letters in your sequences, or relative to the frequencies given in a "Custom background model" that you may provide (see Advanced options).

**Discriminative mode**  
You provide two sets of sequences and MEME discovers motifs that are enriched in the first (primary) set relative to the second (control) set. In Discriminative mode, we first calculate a position-specific prior from the two sets of sequences. MEME then searches the first set of sequences for motifs using the position-specific prior to inform the search. This approach is based on the simple discriminative prior "D" described in Section 3.5 of Narlikar et al. We modify their approach to search for the "best" initial motif width, and to handle protein sequences using spaced triples. Refer to the [psp-gen](#) documentation and to our paper for more details.

**Differential Enrichment mode**  
You provide two sets of sequences and MEME discovers motifs that are enriched in the first (primary) set relative to the second (control) set. In Differential Enrichment mode, MEME optimizes an objective function based on the hypergeometric distribution to determine the relative enrichment of sites in the primary sequences compared to the control sequences.

Seleccionem 10 motius amb una amplada de 5-15 pb i executem.



**Data Submission Form**

Perform motif discovery on DNA, RNA, protein or custom alphabet datasets.

**Select the motif discovery mode**

☒ Classic mode
 ☐ Discriminative mode
 ☐ Differential Enrichment mode

**Select the sequence alphabet**

Use sequences with a standard alphabet or specify a custom alphabet.

☒ DNA, RNA or Protein
 ☐ Custom

**Input the primary sequences**

Enter sequences in which you want to find motifs.

Type in sequences

```
catgagcaataactgcccagggctgcacctgcagcagctctggactgg
tgtagagggtgctgctccagagcagcctggagccatgcccctgctc
atgtcaccatgctgtcataatgctatttcttcttctctctctga
agtctgcttattgttgccttggattcctgacatggctcctcag
ccctccctccacagcccccaggtctgatcatttctgctgtggctgag
gctcccaagctgggtgagttctctcctgaccccaactcctctaccc
ccagtgtgccccactgcatgctcctgctgctgctgccccacactc
ctgagctctctcctggagcagctattcacagaccagctgcaggcac
aggagagcaacaactcaagttcattcctcctcagtgctttatttg
ttcggtaaatgcttggaaatcatctcactgctcagcctgagtacac
gtgggtgtgcatgctcttgaataacgcatgaggtacagaaacaa
tgccactcccaacagaggactgtgctgttctctactgtctttatgc
ggtctgtcacagctctcttgcctcagttctcacaggagcagtgagg
```

**Select the number of motifs**

How many motifs should MEME find?

**Input job details**

(Optional) Enter your email address.

(Optional) Enter a job description.

**Advanced options**

What should be used as the background model?

0-order model of sequences

How wide can motifs be?

Minimum width:  Maximum width:

How many sites must each motif have?

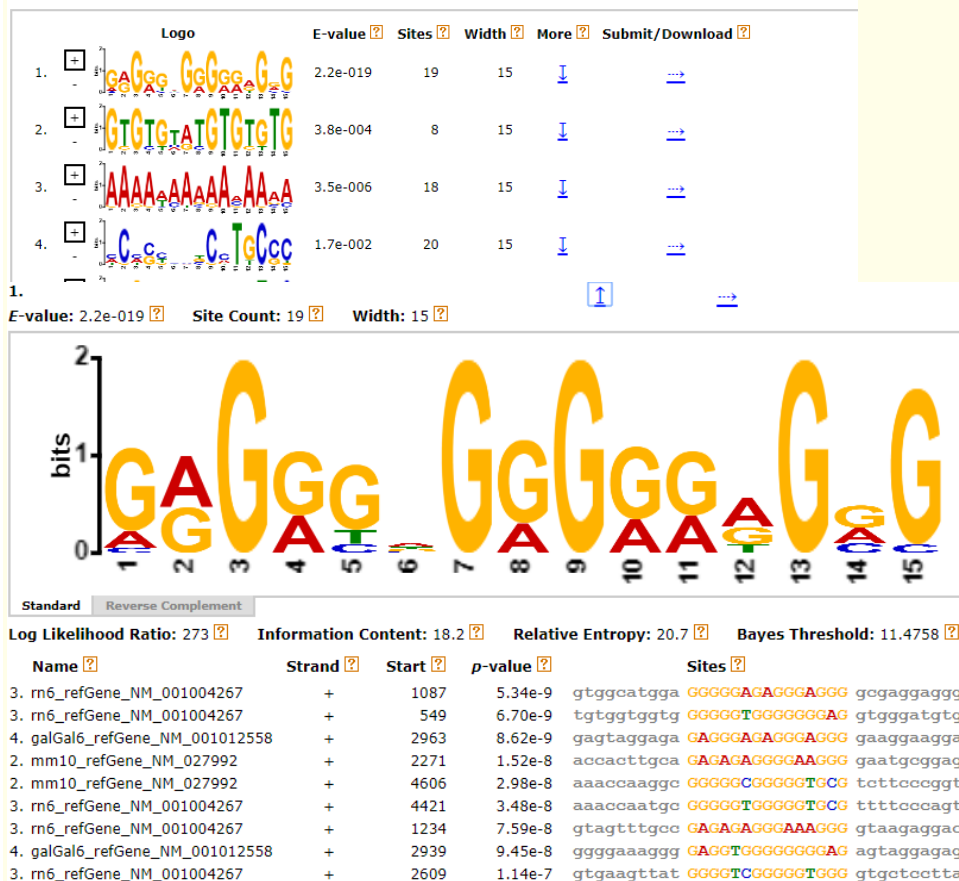
☐ Minimum sites: 
☐ Maximum sites:

Can motif sites be on both strands? (DNA/RNA only)

☐ search given strand only
 ☐ Should MEME restrict the search to palindromes? (DNA only)
 ☐ look for palindromes only

A continuació mostrem els primers resultats obtinguts (veure [annex8](#)). Com es demanen els 10 millors i MEME mostra els 10 primers que troba, sense tenir en compte l'E-value, farem la cerca de 30 i seleccionarem els 10 millors (veure [annex8b](#)).

#### DISCOVERED MOTIFS



El primer resultat té un E-value significatiu, així que el farem servir per fer la comparació amb la base de dades de Jaspár, mitjançant l'opció TOMTOM.

#### 11. TOMTOM dins de la suite de programes MEME. Exemple amb algun del motius anteriors.

JASPAR és un catàleg de models de predicció pel reconeixement del llocs d'unió a factors de transcripció en regions reguladores. Trobem diferents col·leccions de matrius de pesos classificades per espècies. La llista de motius reconeguts per un determinat factor de transcripció (TF) es representen gràficament amb un logotip de seqüència, de manera que podem avaluar la variabilitat d'aquests motius. Aquests mateixos logotips són els que obtenim en el outputs de MEME. Desde MEME, la opció *Submit to program* – TOMTOM, permet comparar el motiu/s obtinguts amb múltiples catàlegs per identificar a quina família de TFs podria pertànyer.



## QUERY MOTIFS

Detailed information of matrix profile **MA0528.1**

## Exercici 2. Anotació computacional del gens.

GENEID és un dels programes pioners de predicció gènica. Identifica els exons codificants d'una seqüència problema *ab initio* (de novo o intrínseca), és a dir, es basa en la construcció de models estadístics a partir de l'estudi d'abundants col·leccions d'exons reals per capturar la composició dels senyals i les regions codificants internes sense fer servir informació externa derivada de l'ús de bases de dades d'homologies. GENEID posseeix, a més, un potent algorisme d'acoblament de gens basat en l'ús de programació dinàmica. Els programes de predicció *ab initio* inclouen també fitxers de paràmetres per cada grup taxonòmic, basats en la distribució estadística dels senyals i regions codificants dins de cada espècie. Es calculen també els valors òptims de la distribució de nucleòtids al voltant de cada senyal, i s'identifiquen els falsos positius prenent com a referència una àmplia mostra de regions no exòniques (model negatiu de contrast). Executem la cerca per la nostra seqüència problema *anonima.fa* amb format de sortida GFF:

```
## gff-version 2
## date Thu May 7 18:44:05 2020
## source-version: geneid v 1.2 -- geneid@imim.es
# Sequence human - Length = 37571 bps
# Original Gene Structure. 2 genes. Score = 31.87
# Gene 1 (Forward). 11 exons. 622 aa. Score = 31.58
human geneid_v1.2 First 157 286 9.81 + 0 human_1
human geneid_v1.2 Internal 10376 10458 1.45 + 2 human_1
human geneid_v1.2 Internal 12800 12857 0.89 + 0 human_1
human geneid_v1.2 Internal 15504 15655 -0.00 + 2 human_1
human geneid_v1.2 Internal 16764 16924 1.03 + 0 human_1
human geneid_v1.2 Internal 17225 17406 5.73 + 1 human_1
human geneid_v1.2 Internal 23771 23865 -1.35 + 2 human_1
human geneid_v1.2 Internal 25045 25142 2.96 + 0 human_1
human geneid_v1.2 Internal 26262 26281 2.17 + 1 human_1
human geneid_v1.2 Internal 27296 27427 2.70 + 1 human_1
human geneid_v1.2 Terminal 28008 28858 6.20 + 2 human_1
# Gene 2 (Forward). 4 exons. 141 aa. Score = 0.28
human geneid_v1.2 First 30518 30529 -2.92 + 0 human_2
human geneid_v1.2 Internal 30780 30932 0.68 + 0 human_2
human geneid_v1.2 Internal 31931 31994 2.93 + 0 human_2
human geneid_v1.2 Terminal 33682 33875 -0.40 + 2 human_2
```

Si canviem el format de sortida a *geneid* obtenim la proteïna resultant de la traducció d'aquests 2 gens:

```
# Optimai Gene Structure. 2 genes. score = 31.57
# Gene 1 (Forward). 11 exons. 622 aa. Score = 31.58
First      157      286      9.81 + 0 1      8.07      2.83      20.67      0.00      AA      1: 44 human_1
Internal    10376    10458    1.45 + 2 0      5.58      2.65      3.77      0.00      AA      44: 71 human_1
Internal    12800    12857    0.89 + 0 1      3.87      3.09      5.54      0.00      AA      72: 91 human_1
Internal    15504    15655    -0.00 + 2 0      0.91      4.65      5.41      0.00      AA      91:141 human_1
Internal    16764    16828    1.03 + 0 2      4.32      1.67      6.10      0.00      AA     142:163 human_1
Internal    17225    17406    5.73 + 1 1      3.69      3.72     15.71      0.00      AA     163:224 human_1
Internal    23771    23865    -1.35 + 2 0     -0.44      3.68      4.25      0.00      AA     224:255 human_1
Internal    25045    25142    2.96 + 0 2      3.54      0.05     14.52      0.00      AA     256:288 human_1
Internal    26262    26281    2.17 + 1 1      6.90      4.53      0.77      0.00      AA     288:295 human_1
Internal    27296    27427    2.70 + 2 1      0.45      5.26     10.70      0.00      AA     295:339 human_1
Terminal    28008    28858     6.20 + 2 0      4.56      0.00     21.14      0.00      AA     339:622 human_1

>human_1|geneid_v1.2_predicted_protein_1|622_AA
MAPAMQPAEIQFAQRLASSEKGIKRAVKKLRQYISVKTQRETTGGFSQEELLKIWKGLFY
CMWVQDEPLLQEEELANTIAQLVHAVNNSAAQACVWFSSRIKVFLLVLMKEVLCPESQSPN
GVRFFHFDIYLDLSEKVGKGLADQLNKFDPFCKIAAKTKDHTLVQTIARGVFEAIVD
QSPFVPEETMEEQKTKVGDGDLAEEIIPENVSLLRAVSKKKKTALGKNHRSKDGLSDERG
RDDCGTFEDTGPLLQFDYKAVADRLLMETSRKNTPHFNKRRLSKLKKFQDLSEGSSISQ
LSFAEDISADEDDQILSQGKHKKKGNKLEKTNLEKEKGRVFCVEEEDSESLQKRRRK
KKKKHHLQPENPGPGGAAPSLQNRGREPEASGLKALKARVAEPGAETSSSTGEESGSEH
PFAVPMHNKRKRPRKKSRAHREMLESAVLPPEDMSQSGPSGSHFQGPGRGSPPTGAQLLK
RKRKLGVVFNGLSTPAWPPLOQEGPPTGPAEGANSHTTLPQRRRLQKKKAGPGSLEL
CGLPSQKTASLKKRKKMRVMSNLVEHNGVLESEAGQPQALVRWEHPQASSPQRHSLASMG
LHCLLRGRVGGGQASGLSSS*

# Gene 2 (Forward). 4 exons. 141 aa. Score = 0.28
First      30518    30529    -2.92 + 0 0      1.42      1.23      1.21      0.00      AA      1: 4 human_2
Internal    30780    30932     0.68 + 0 0      2.81      3.31      5.01      0.00      AA      5: 55 human_2
Internal    31931    31994     2.93 + 0 1      4.88      4.80      5.31      0.00      AA     56: 77 human_2
Terminal    33682    33875    -0.40 + 2 0     -0.70      0.00     12.53      0.00      AA     77:141 human_2

>human_2|geneid_v1.2_predicted_protein_2|141_AA
MKIKGSSGTCSSILKKQKRAESDFVKFDTPLFKPLFFRAKSSSTATHPPGPAVQLNKTP
SSSKVIFGLNRNWTAEFKKIDKSILVSTPGSRVAFDPEQKPLHGVLTPTSTSSPASP
VAKKFLITTPRRRPRAMDFF*
```

Si seleccionem com a format de sortida geneid i l'opció “All exons” obtenim tots els possibles exons:

## Output options

Output format:

geneid

Elements to be included in the output:

Signals	Exons
<input type="checkbox"/> Acceptor sites.	<input type="checkbox"/> First exons.
<input type="checkbox"/> Donor sites.	<input type="checkbox"/> Internal exons.
<input type="checkbox"/> Start codons.	<input type="checkbox"/> Terminal exons.
<input type="checkbox"/> Stop codons.	<input type="checkbox"/> Single genes.
	<input type="checkbox"/> Open reading frames
	<input checked="" type="checkbox"/> All exons

Només mostrarem els primers ja que amb una seqüència de més de 37mil bases s'obtenen també milers de possibles exons:

```
# Sequence human - Length = 37571 bps
# Exons(x) predicted in sequence human: [0,37570]
Begin      0      0      0.00 + 0 0      0.00      0.00      0.00      0.00      1      X
Begin      0      0      0.00 - 0 0      0.00      0.00      0.00      0.00      1      X
Internal    2      77      -6.25 + 0 1     -0.43     -0.72     -1.40      0.00     26      PRRLCDAISPRAAAAAFCVAARADg
Internal    2      77      -5.75 + 1 0     -0.43     -0.72     -0.16      0.00     26      cRGAFVTPSARAPPPPPSVQSRPGRT
Internal    2     237     -10.74 + 0 2     -0.43     -0.04     -13.65      0.00     79      PRRLCDAISPRAAAAAFCVAARADGGWLLRSARLAAAAPRVAR
Internal    2     237     -9.10 + 1 1     -0.43     -0.04     -9.55      0.00     80      cRGAFVTPSARAPPPPPSVQSRPGRTVAGCSAALGWLQRHRLR
Internal    2     286     -7.53 + 1 2     -0.43     2.83     -9.92      0.00     96      cRGAFVTPSARAPPPPPSVQSRPGRTVAGCSAALGWLQRHRLR
Internal    2     303     -9.80 + 1 1     -0.43     -0.01     -11.35      0.00    102      cRGAFVTPSARAPPPPPSVQSRPGRTVAGCSAALGWLQRHRLR
Terminal    2      18     -5.78 + 2 0     -0.43      0.00     -1.31      0.00      6      ccAAPL*
Terminal    2     241     -10.69 + 0 0     -0.43      0.00     -13.57      0.00     80      PRRLCDAISPRAAAAAFCVAARADGGWLLRSARLAAAAPRVAR
First      20     24     -2.76 - 0 2     -4.78      8.52      0.00      0.00      2      Mgc
Internal    20     246     -10.82 - 0 2     -4.78      0.38     -7.94      0.00     76      LLHRSVPDALLAGRQPLGKLDLGRHLGGGHRARWSIPGRATRGA
Internal    20     246     -11.43 - 2 0     -4.78      0.38     -9.48      0.00     76      ctSSPLGPGCPSRWTPAAGQIGSRPAAWRGPSRPLEHPRPRNPR
Internal    20     201     -10.78 - 0 2     -4.78     -0.90     -5.91      0.00     61      PLGKLDLGRHLGGGHRARWSIPGRATRGAAAAARALRSSQPPSA
Internal    20     201     -10.33 - 2 0     -4.78     -0.90     -4.80      0.00     61      ccAGQIGSRPAAWRGPSRPLEHPRPRNPRCRCQSQAEEQPATV
Internal    20     107     -8.33 - 1 0     -4.78     -0.52     -0.36      0.00     30      cQPSAAEQPATVRPGRDCTEGGGGARADG
Internal    20     107     -9.83 - 2 2     -4.78     -0.52     -4.11      0.00     30      ccSRALRSSQPPSARAATAQKAAAAAARGLMgc
Internal    20     103     -10.74 - 0 0     -4.78     -4.44     -0.52      0.00     28      PSAAEQPATVRPGRDCTEGGGGARADG
Internal    20     103     -12.10 - 1 2     -4.78     -4.44     -3.92      0.00     29      cRALRSSQPPSARAATAQKAAAAAARGLMgc
Internal    20     51      -8.69 - 0 2     -4.78     -1.39      0.05      0.00     11      KAAAAARGLMgc
Internal    20     51      -8.37 - 2 0     -4.78     -1.39      0.83      0.00     11      aaGGGGGARADG
Terminal    24     344     -2.71 - 0 0      0.00     -0.40      6.32      0.00    107      POPRPPARHVARLTAAVRPFVSLCVFTILMYWRSFFTARSMPFSS
Terminal    24     338     -4.75 - 0 0      0.00     -4.25      7.01      0.00    105      PRPPARHVARLTAAVRPFVSLCVFTILMYWRSFFTARSMPFSLD
Terminal    24     246     -4.14 - 1 0      0.00      0.38      1.58      0.00     75      cFFTARSMPFSLDASRWANWISAGCMAGAIAPAGASPAAPQAVPLQPAERCGAASHRP
Terminal    24     201     -6.84 - 1 0      0.00     -0.90     -3.23      0.00     60      cRWANWISAGCMAGAIAPAGASPAAPQAVPLQPAERCGAASHRP
```

Podem estudiar el patró de nucleòtids si seleccionem visualitzar els senyals que delimiten el inici i final d'aquests exons (donada la seva extensió, també es mostraran només els primers):

Signals	Exons
<input checked="" type="checkbox"/> Acceptor sites.	<input type="checkbox"/> First exons.
<input checked="" type="checkbox"/> Donor sites.	<input type="checkbox"/> Internal exons.
<input checked="" type="checkbox"/> Start codons.	<input type="checkbox"/> Terminal exons.
<input checked="" type="checkbox"/> Stop codons.	<input type="checkbox"/> Single genes.
	<input type="checkbox"/> Open reading frames
	<input checked="" type="checkbox"/> All exons

Sequence human - Length = 37571 bps

```
# Starts(+) predicted in sequence human: [0,37570]
Start      140      142      3.00      +      TTGCGCGCGCGGGGATGCTC
Start      157      159      8.07      +      CTCCAGCGGGCGCGATGGCC
Start      169      171      3.77      +      CGATGGCCCCCGCCATGCAG
Start      318      320      4.97      +      GTCAGCGCGCCACATGGCG
Start      369      371      4.31      +      GGCCCCGGCACGGAATGCGG
Start      496      498      -0.81     +      GGGTTCCGTTACTTATGAAA
Start      537      539      -0.77     +      GAGAAAAGACCCACATGGGT
Start      757      759      -2.32     +      GTGTAGTAAAGGAATGAAA
```

És recomanable fer servir diferents programes de predicció gènica per estudiar quines són les prediccions en comú. També és interessant comparar aquestes prediccions amb bases de dades de proteïnes per enriquir aquestes prediccions.

## 2. GENSCAN. Predicció de gens en la seqüència *anonima.fa*

GENSCAN és un programa que prediu la localització de gens i les estructures exó/intró en seqüències gèniques de diferents organismes.

Deixarem com a “suboptimal exon cutoff” el que ve per defecte, ja que de moment ens interessen els exons més plausibles. Seleccionem “Vertebrate”. Els programes de predicció gènica treballen amb bases de dades d'exons pròpies de cada grup taxonòmic, per capturar amb precisió la distribució estadística de senyals i regions codificants, i fer les parametrizables.

Organism:  Suboptimal exon cutoff (optional):

Sequence name (optional):

Print options:

Upload your DNA sequence file (upper or lower case, spaces/numbers ignored):

Or paste your DNA sequence here (upper or lower case, spaces/numbers ignored):

La quantitat de possibles exons que obtenim aquí és menor. Observem una llista resultant d'aquests exons on també s'indica el tipus d'exó, l'inici, el final, la longitud, *score*, ORF,...i a continuació el pèptid predit i la possible regió CDS (només mostrem una part):

Gn.Ex	Type	S	.Begin	...End	.Len	Fr	Ph	I/Ac	Do/T	CodRg	P....	Tscr..	
1.01	Init	+	162	291	130	2	1	107	80	324	0.752	33.81	MAMAMQPAETQFAQLASSEKGIKRAVKKLRQVTSVKTKRETGGFSQEELKIKKGLFY
1.02	Intr	+	10381	10463	83	2	2	94	92	26	0.829	2.96	CMNQDEPLLEELANTIAQLVHVNNSAAQHLFIQTFTQPMREWGIDRLRLKYYML
1.03	Intr	+	12805	12862	58	0	1	97	99	62	0.963	6.66	IRLVLRQSFVFLKRWMEESRIKVFLDVLKKEVLCPESSQSPNGVRFHIDIIYLDLSKVG
1.04	Intr	+	14367	14452	86	1	2	47	95	49	0.678	1.04	GKELLADQNLKFIDPFCKIAAKTDHILVQTIARGVFAIVDQSPFVPEETHEEQTKVG
1.05	Intr	+	15133	15194	62	0	2	53	86	51	0.694	-1.07	DGOLSAAEIPENEVSRRRAVSKKKTALGKNYSBKDGLSDERGRDCCGTFEDTGPLLQDY
1.06	Intr	+	15531	15660	130	0	1	27	99	108	0.642	6.50	KAVADRLLEMTSRKNTPHFNKRLSKLKKFQQLSEGSSISQLSFAEDISADEDDQILSQ
1.07	Intr	+	16769	16833	65	1	2	78	83	73	0.995	3.32	GIOBKXGKLLKTNLEKEKGKQELQGALGGGCLMTTDLNHLPLSPKISGNGTISVPYV
1.08	Intr	+	17230	17411	182	1	2	77	91	192	0.962	17.91	FINGQKEGFSQQLGHEEVGPDOKGSRVFCVEEDESLSQKRRRXXXXXHLQENPGPG
1.09	Intr	+	23776	23870	95	2	2	37	94	55	0.688	0.68	GAAPSLQNRIGREPEASGLKALKARVAEPGAEATSSTGEESGSEHPPAVPMNKRRPRK
1.10	Intr	+	25050	25147	98	2	2	64	26	129	0.640	3.11	KSPRAHREMLESAVLPPEDMSQSGPSGHPQGRGSPGTGGAQLLKRKLGVPVNGSGL
1.11	Intr	+	26267	26286	20	2	2	91	100	-1	0.600	-2.35	STPAWPLQDEGPTGAEGANSHITLPQRRRLQKKKAGPSLELCGLPSQKTASLKKR
1.12	Intr	+	27301	27432	132	2	0	41	121	120	0.872	11.22	KMRVMSNLVEHNGVLESEAGQQAALHNLPEPPVCRQRHAAHTSESQVRDPVSLWA
1.13	Intr	+	27668	27856	189	0	0	51	67	92	0.625	2.96	VSCCTRNECPGASVVLCKPELCRMEGLSASAVRKTAGRRGSSGTCSSLKKQLRAESD
1.14	Intr	+	28013	28737	725	0	2	85	95	470	0.762	38.55	FWKFDTPFLPKPLFFRRAKSSTATHPPGPAVLNKTPTSSSKXVTFGLNRMHTAEFKTKD
													SLLVSPGTGPSRVAFDPEQKPLHGVLTPTSSPASSPLVAKKPLTTTPRRRPRAMDF

## 3. FEGENESH. Predicció de l'estructura gènica basada en HMM (*multiples genes, both chains*).

FEGENESH és un software de predicció gènica *ab initio* (de novo o intrínseca) basat en HMM.

Introduïm de nou la nostra seqüència *anonima.fa*

Local file name:

Select organism specific gene-finding parameters:

Obtenim (veure [annex9](#)) dins de la nostra seqüència problema, un possible gen i 16 possibles exons. L'aplicació ens retorna, a més de la llista d'exons (amb especificacions similars a les de les altres aplicacions), la seqüència del mRNA (perquè és una opció marcada per defecte) i de la proteïna resultant del gen predit.

```

FGENESH 2.6 Prediction of potential genes in Homo_sapiens genomic DNA
Time : Thu May 7 16:26:14 2020
Seq name: human
Length of sequence: 37571
Number of predicted genes 1: in +chain 1, in -chain 0.
Number of predicted exons 16: in +chain 16, in -chain 0.
Positions of predicted genes and exons: Variant 1 from 1, Score:115.993872
G Str Feature Start End Score ORF Len
1 + 1 CDSf 157 - 286 28.30 157 - 285 129
1 + 2 CDSi 10376 - 10458 7.43 10378 - 10458 81
1 + 3 CDSi 12800 - 12857 6.33 12800 - 12856 57
1 + 4 CDSi 14362 - 14447 3.34 14364 - 14447 84
1 + 5 CDSi 15128 - 15189 2.40 15128 - 15187 60
1 + 6 CDSi 15526 - 15655 6.39 15527 - 15655 129
1 + 7 CDSi 16764 - 16828 6.62 16764 - 16826 63
1 + 8 CDSi 17225 - 17406 12.24 17226 - 17405 180
1 + 9 CDSi 23771 - 23865 2.10 23773 - 23865 93
1 + 10 CDSi 25045 - 25142 0.60 25045 - 25140 96
1 + 11 CDSi 26262 - 26281 -1.21 26263 - 26280 18

```

#### Alguns exons i la possible proteïna:

```

>FGENESH:[exon] Gene: 1 Exon: 14 Pos: 30780 - 30932 153 bp., chain +
GGAAGCAGTGGGACTTGCAGTTCCCTGAAGAAGCAGAAGCTGAGGGCAGAGAGCGACTTT
GTGAAGTTTGACACCCCTCTTACCAAGCCCTGTCTCTCAGAAGAGCCAAAGCAGCAGC
ACTGCCACCCACCTCCAGGCCCTGCCGTCCAG
>FGENESH:[exon] Gene: 1 Exon: 15 Pos: 31931 - 31994 64 bp., chain +
CTAAACAAGACACCATCCAGCTCCAAGAAAGTCACCTTGGGGTGAACAGAAACATGACT
GCCG
>FGENESH:[exon] Gene: 1 Exon: 16 Pos: 33682 - 33875 194 bp., chain +
AATTCAAGAGACAGACAGAGATATCTTGGTCAGTCCACGGGCCCTTCTCGAGTGGCCT
TCGACCTGAACAGAGCCCTCCACGGGGTGTGAAGACCCCAACAGCTCACCTGCCA
GCTCACCCCTGGTGGCCAGAGCCCTGACCACCAACCAAGAGAGAGCCAGGCTA
TGATTTCTCTGA
>FGENESH: 1 16 exon (s) 157 - 33875 758 aa, chain +
MAPAMQPAEIQPAQLASSEKGIKRAVKKLQYISVKTQRETGGFSQEELLKIWKGLFY
CMWVQDEPLQLLELANTIAQLVHAVNNSAAQHLFIQTFWQTMNREWKIDRLRLDKYYML
IRLVLRQSEVLKRNWEEESRIKVFLLDVLKVEVLCPEQSPNGVRFHFDIYLDLQSVG
KRELLADQNLKFIKIDPFCKIAAKTKDHTLVQTIARGVFEIVDQSPFVPEETMEEQKTKV
DGDLSAEIEIPENEVSLRAVSKKKKALGNHNRKDKLSDEGRDDCGTFEDTGFLQLQDY
KAVADRLLLEMTSRKNTPHFNRRRLSKLIKPFQDLSESSISQLSFAEDISADEDDQLSQ
GKHKKKGNKLEKTNLEKGRVFCVEEDESSESLQKRRRKKKKHHLQEPENPGPGGAA
PSLEQNRGREPEASGLKALKARVAEPGAETSSSTGEESGSEHPFPAVPMHNRRKRPRKSP
RAHREMLESAVLPPEDMSQSGSPSGHPQGRGPTGGAQLLKRKRKLGVVNVNGSLSTP
AWPFLQOEGPPTGPAEGANSHTTLPQRRRLQKKRAGPGSLELGLPSQKTASLKKRKKMR
VMSNLVHNGVLESEAGQPQALGSSGTCSSSLKKQLRAESDFVKFDTFPLPKPLFFRRAK
SSTATHPGPAVQLNKTPSSSKKVTFLNLRNMTAEFKKTDKLSILVSPGSRVAFDPEQK
PLHGVLTPTSSPASSPLVAKKPLTTTPRRRPRAMDFF

```

#### 4. Alineament global amb CLUSTAL de les proteïnes obtingudes anteriorment en GENEID, GENSCAN i FGENESH.

Realitzem l'alineament global dels tres pèptids obtinguts en els apartats 1, 2 i 3 amb CLUSTAL. En primer lloc, en el cas de GENEID prenem només el gen 1 (veure seqüències alineades en [annex10](#)). Obtenim uns percentatges alts d'identitat, que és màxima entre els resultats de FGENESH i GENSCAN

Percent Identity Matrix - created by Clustal2.1

```

1: human_1|geneid_v1.2_predicted_protein_1|622_AA      100.00  97.80  94.04
2: FGENESH_                                           97.80  100.00  100.00
3: /tmp/05_08_20-11:44:37.fasta|GENSCAN_predicted_peptide_1|897_aa  94.04  100.00  100.00

```

La diferència més notable la trobem a la proteïna obtinguda per GENEID, observem d'entrada que el número d'aa és diferent i, si observem l'alineament obtingut a CLUSTAL, veiem que és principalment a l'extrem terminal on es perd aquesta identitat amb les altres dues seqüències.

Això és degut a que, el que GENSCAN i FGENESH identifiquen com un únic gen, en GENEID resulten 2, ja que reconeix un codó de parada a l'exó terminal del primer gen. Cada programa implementa els seus propis algorismes i estableix uns llindars (*cutoff o threshold*) de puntuació, de manera que els resultats no sempre coincideixen.

Canviem a CLUSTAL la seqüència de la proteïna de GENEID: li afegim el pèptid del gen (veure [annex11](#)). Observem que la identitat augmenta en l'extrem terminal:

```

human_1|geneid_v1.2_predicted_protein_1|622_AA      KMRVMSNLVEHNGVLESEAGQPQALVRWEHPQA-----SSPQRHSL-ASMG      600
FGENESH:      KMRVMSNLVEHNGVLESEAGQPQAL-----SSPQRHSL-ASMG      622
/tmp/05_08_20-11:44:37.fasta|GENSCAN_predicted_peptide_1|897_aa  KMRVMSNLVEHNGVLESEAGQPQALAAHLNLPEPPVCRQRHWAHTSESQVRDPVSLWVA  720
*****

human_1|geneid_v1.2_predicted_protein_1|622_AA      LHCLLRGRV-----GAGGQASGLSSMKIKGSSGTCSSLKKQKLAESD      644
FGENESH:      -----GSSGTCSSLKKQKLAESD      641
/tmp/05_08_20-11:44:37.fasta|GENSCAN_predicted_peptide_1|897_aa  VSCCTRNECPGASVVLCKVPELCRMEGLSASAVRKTAGRRGSSGTCSSLKKQKLAESD  780
*****

human_1|geneid_v1.2_predicted_protein_1|622_AA      FVKFDTPLPKPLFFRRAKSSSTATHPGPAVQLNKTPSSSKKVTFLNLRNMTAEFKKTDK  704
FGENESH:      FVKFDTPLPKPLFFRRAKSSSTATHPGPAVQLNKTPSSSKKVTFLNLRNMTAEFKKTDK  701
/tmp/05_08_20-11:44:37.fasta|GENSCAN_predicted_peptide_1|897_aa  FVKFDTPLPKPLFFRRAKSSSTATHPGPAVQLNKTPSSSKKVTFLNLRNMTAEFKKTDK  840
*****

human_1|geneid_v1.2_predicted_protein_1|622_AA      SILVSPGTPSRVAFDPEQKPLHGVLTPTSSPASSPLVAKKPLTTTPRRRPRAMDFF*  761
FGENESH:      SILVSPGTPSRVAFDPEQKPLHGVLTPTSSPASSPLVAKKPLTTTPRRRPRAMDFF-  758
/tmp/05_08_20-11:44:37.fasta|GENSCAN_predicted_peptide_1|897_aa  SILVSPGTPSRVAFDPEQKPLHGVLTPTSSPASSPLVAKKPLTTTPRRRPRAMDFF-  897
*****

```

Percent Identity Matrix - created by Clustal2.1

```
1: human_1|geneid_v1.2_predicted_protein_1|622_AA      100.00   99.30   94.35
2: FGENESH_                                           99.30   100.00   100.00
3: /tmp/05_08_20-11_44_37.fasta|GENSCAN_predicted_peptide_1|897_aa  94.35   100.00   100.00
```

## 5. Coordenades dels exons.

	GENEID		GENSCAN	FGENESH	Exó
Gen 1	inicial	157-286**	162-291***	157-286***	inicial
	intern	10376-10458	10381-10463	10376-10458	intern
	intern	12800-12857	12805-12862	12800-12857	intern
	intern		14367-14452	14362-14447	intern
	intern		15133-15194	15128-15189	intern
	intern	15504-15655	15531-15660	15526-15655	intern
	intern	16764-16828	16769-16833	16764-16828	intern
	intern	17225-17406*	17230-17411**	17225-17406*	intern
	intern	23771-23865	23776-23870	23771-23865	intern
	intern	25045-25142	25050-25147	25045-25142	intern
	intern	26262-26281	26267-26286	26262-26281	intern
	intern	27296-27427*	27301-27432*	27296-27427*	intern
	intern		27668-27856		intern
	terminal	28008-28858**	28013-28737***	28008-28732***	intern
Gen 2			30241-30385		intern
	inicial	30518-30529	30594-30676		intern
	intern	30780-30932	30785-30937*	30780-30932*	intern
	intern	31931-31994	31936-31999*	31931-31994*	intern
	terminal	33682-33875	33687-33880*	33682-33875	terminal

Observem a la taula que hi han moltes coincidències. Alguns exons han estat simultàneament detectats pels 3 programes. A la taula s'han marcat també els exons amb un *score* més alt (\*\*\*) i també hi ha coincidències en aquesta puntuació.

Amb BLASTP podem reforçar aquestes prediccions obtingudes amb estratègies *de novo*, fent una validació per homologia amb proteïnes conegudes. Prenem com a exemple l'exó terminal del Gen1 que obtingut amb GENEID que, a més, té una alta puntuació i coincideix bastant amb el altres dos resultats.

Utilitzem com a base de dades la refseq\_protein. Obtenim el següent resultat:

Descriptions

Graphic Summary

Alignments

Taxonomy

Sequences producing significant alignments

Download

Manage Columns

Show

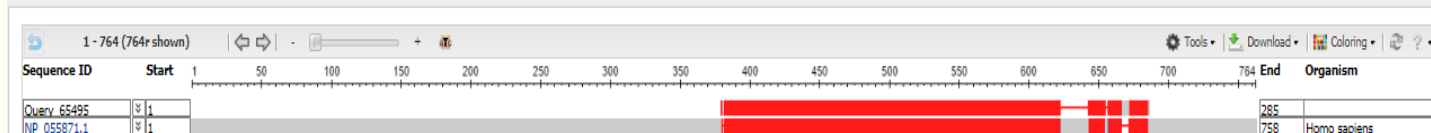
100

?

☒ select all
 1 sequences selected

[GenPept](#)
[Graphics](#)
[Distance tree of results](#)
[Multiple alignment](#)

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
<input checked="" type="checkbox"/>	ribosomal RNA processing protein 1 homolog B [Homo sapiens]	483	483	84%	3e-167	100.00%	NP_055871.1





ribosomal RNA processing protein 1 homolog B [Homo sapiens]

Sequence ID: [NP\\_055871.1](#) Length: 758 Number of Matches: 1

Range 1: 382 to 622 [GenPept](#) [Graphics](#) [Next Match](#) [Previous Match](#)

Score	Expect	Method	Identities	Positives	Gaps
483 bits(1243)	3e-167	Compositional matrix adjust.	241/241(100%)	241/241(100%)	0/241(0%)
Query	3	SRVFCVEEEDSESSLOKRRRKKKKHHLQENPGGGAAPSLQNRGREPEASGLKALKA			62
Sbjct	382	SRVFCVEEEDSESSLOKRRRKKKKHHLQENPGGGAAPSLQNRGREPEASGLKALKA			441
Query	63	RVAEPGAEATSSSTGEESGSEHPPAVPMHNKRKRPRKKSAPRAHREMLES AVLPPEDMSQSG			122
Sbjct	442	RVAEPGAEATSSSTGEESGSEHPPAVPMHNKRKRPRKKSAPRAHREMLES AVLPPEDMSQSG			501
Query	123	PSGSHPOGPRGSPGTGGAQLLKRKRKLGVVPVNGSLSTPAWPPLQOEGPPTGPAEGANSH			182
Sbjct	502	PSGSHPOGPRGSPGTGGAQLLKRKRKLGVVPVNGSLSTPAWPPLQOEGPPTGPAEGANSH			561
Query	183	TTLPQRRRLQKKKAGPGSLELCGLPSQKTASLKKRKKMRVMSNLVEHNGVLESEAGQPQA			242
Sbjct	562	TTLPQRRRLQKKKAGPGSLELCGLPSQKTASLKKRKKMRVMSNLVEHNGVLESEAGQPQA			621
Query	243	L 243			
Sbjct	622	L 622			

Ara fem el mateix amb l'exó intern 30780-30932 que, segons GENEID, pertany a un altre gen:

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#) [Query subrange](#)

From

To

Or, upload file [Seleccionar archivo](#) Ningún archivo seleccionado

Job Title  Enter a descriptive title for your BLAST search

☐ Align two or more sequences

Choose Search Set

Database [Reference proteins \(refseq\\_protein\)](#)

Organism  ☐ exclude ☐

Exclude ☐ Models (XM/XP) ☐ Non-redundant RefSeq proteins (WP) ☐ Uncultured/environmental

Program Selection

Algorithm ☒ blastp (protein-protein BLAST)

Obtenim la mateixa proteïna:

☒ select all 1 sequences selected [GenPept](#) [Graphics](#) [Distance tree of results](#) [Multiple alignment](#)

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
<input checked="" type="checkbox"/>	<a href="#">ribosomal RNA processing protein 1 homolog B [Homo sapiens]</a>	106	106	100%	2e-28	100.00%	<a href="#">NP_055871.1</a>

ribosomal RNA processing protein 1 homolog B [Homo sapiens]

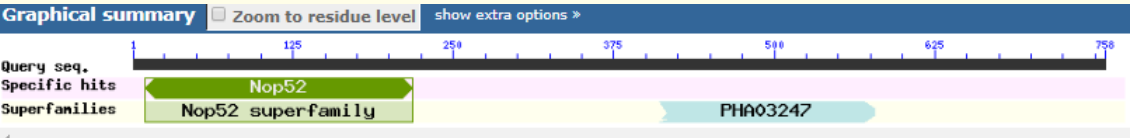
Sequence ID: [NP\\_055871.1](#) Length: 758 Number of Matches: 1

Range 1: 623 to 673 [GenPept](#) [Graphics](#) [Next Match](#) [Previous Match](#)

Score	Expect	Method	Identities	Positives	Gaps
106 bits(264)	2e-28	Compositional matrix adjust.	51/51(100%)	51/51(100%)	0/51(0%)
Query	1	GSSGTCSSLLKKQLRAESDFVKFDTPLPKPLFFRRAKSSSTATHPPGPAVQ			51
Sbjct	623	GSSGTCSSLLKKQLRAESDFVKFDTPLPKPLFFRRAKSSSTATHPPGPAVQ			673

Es tracta d'una proteïna amb 2 dominis:

- Nop52: proteïna nucleolar que es creu està involucrada en la generació de la 28S rRNA .
- PHA03247: proteïna del tégument .



La seqüència completa d'aa és:



### Translation (758 aa):

MAPAMQPAEIQFAQRLASSEKGI~~DR~~AVKKLRQYISVKTQRETGGFSQEELLKI~~W~~KGLFYCMVQDEPLL  
QEELANTIAQLVHAVNMISAAQH~~L~~FIQTFWQTMNREWKGI~~DR~~LRLDKYMYLIRLVLRSFEVLKRNWEEES  
RIKVF~~L~~DLVLMKEVLCPE~~S~~QSPNGVRFHFID~~I~~YLDLSKVGGKELLADQNLKFIDPFCKIAAKTKDHTLVQ  
TIARGVF~~E~~AIVDQSPFVPEETMEEQTKVGDGDL~~S~~AAEIPENEVSLRAVSKKKTA~~L~~GKNHSRKDGLSDE  
RGRDDCGTFEDTGPLQFDYKAVADRLLEMTSRKNTPHFNRKRLSKLIK~~F~~QDLSE~~G~~SSISQLSFAEDIS  
ADEDDQILSQGKHKKKGNKLL~~E~~KTNLEKEK~~G~~SRVFCVEEEDSESSLQKRRRKKKKHHLQENPGPGGAA  
PSLEQNRGREPEASGLKALKARVAEPGA~~E~~ATSS~~T~~GEESGSEHPPAVPMHNKRKRPRKKS~~P~~RAHREMLESA  
VLPPEDMSQSGPSGSHPGQGRGSP~~T~~GGAQLLKRKRKLGVVPVNGSGLSTPAWPLQ~~Q~~EGPPTGPAEGANS  
HTTLPQRRRLQKKKAGPGSLELCGLPSQKTASLKKRKKMRVMSNLVEHMGVLESEAGQPQALGSSGTCS  
LKKQKLRAESDFVKFDTPLPKPLFFRRAKSS~~T~~ATHPPGPAVQLNKTPSSSKKVTFGLNRNMTA~~E~~FKKTD  
KSILVSP~~T~~GPSRVAFDPEQKPLHGLVKTP~~T~~SSPASSPLVAKPLTTTPRRRRPRAMDF

En blau es representen els exons i en vermell els aa implicats en l'splicing. Comprobem que, efectivament, que els dos exons hi són.

## 6. Localització en el genoma (BLAT).

### RefSeq Gene RRP1B

**RefSeq:** [NM\\_015056.3](#) **Status:** Validated  
**Description:** ribosomal RNA processing 1B  
**Molecule type:** mRNA  
**Source:** BestRefSeq  
**Biotype:** protein\_coding  
**Synonyms:** KIAA0179, Nnp1, NNP1L, PPP1R136, RRP1  
**OMIM:** [610654](#)  
**Protein:** [NP\\_055871.1](#)  
**HGNC:** [23818](#)  
**Entrez Gene:** [23076](#)  
**GeneCards:** [RRP1B](#)  
**AceView:** [RRP1B](#)

#### mRNA/Genomic Alignments (NM\_015056.3)

BROWSER	SIZE	IDENTITY	CHROMOSOME	STRAND	START	END	QUERY	START	END	TOTAL
<a href="#">browser</a>	5078	100.0%	21	+	43659560	43696079	NM_015056.3	1	5078	5078

[View details of parts of alignment within browser window.](#)

**Position:** [chr21:43659560-43696079](#)

**Band:** 21q22.3

**Genomic Size:** 36520

**Strand:** +

**Gene Symbol:** RRP1B

**CDS Start:** complete

**CDS End:** complete

Un cop tenim la seqüència localitzada, afegim anonima.fa:

### BLAT Search Genome

Genome: ☐ Search all Assembly: ☐ Dec. 2013 (GRCh38/hg38) Query type: ☐ BLAT's guess Sort output: ☐ query,score Output type: ☐ hyperlink

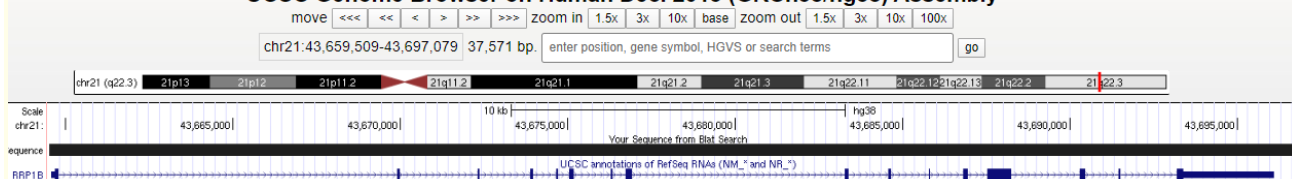
submit I'm feeling lucky clear

Paste in a query sequence to find its location in the the genome. Multiple sequences may be searched if separated by lines starting with '>' followed by the sequence name.

**File Upload:** Rather than pasting a sequence, you can choose to upload a text file containing the sequence.

Upload sequence:  anonima.fa

### UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly



## BLAT Search Results

ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHROM	STRAND	START	END	SPAN
<a href="#">browser details</a>	human	37571	1	37571	37571	100.0%	chr21	+	43659509	43697079	37571
<a href="#">browser details</a>	human	886	18232	20611	37571	88.9%	chr20	+	21734186	22174431	440246
<a href="#">browser details</a>	human	601	18696	19955	37571	87.4%	chr15	-	42655108	42656320	1213
<a href="#">browser details</a>	human	595	18635	19617	37571	88.1%	chr3	+	75534740	75535779	1040
<a href="#">browser details</a>	human	586	18556	19964	37571	84.9%	chr10	+	62391808	62392831	1024
<a href="#">browser details</a>	human	566	18554	20161	37571	85.5%	chr2	-	39200465	39201425	961
<a href="#">browser details</a>	human	565	18875	20117	37571	87.7%	chr3	+	15744540	15745728	1189
<a href="#">browser details</a>	human	554	18557	20147	37571	84.9%	chr10	+	59721602	59722542	941
<a href="#">browser details</a>	human	536	18819	20165	37571	85.7%	chr2	+	73585562	73586788	1227
<a href="#">browser details</a>	human	535	18953	20168	37571	85.6%	chr6	+	63194136	63195017	882
<a href="#">browser details</a>	human	531	18554	19659	37571	84.5%	chr16_KZ559113v1_fix	+	123422	124250	829
<a href="#">browser details</a>	human	519	18555	19631	37571	86.4%	chr3	+	141132858	141133675	818
<a href="#">browser details</a>	human	519	18234	19073	37571	86.3%	chr1	+	232826713	232827646	934
<a href="#">browser details</a>	human	497	18235	19145	37571	93.8%	chr21	-	44588159	45275677	687519

Les coordenades del gen RRP1B són chr21: 43659560-43696079(+), mentre que la localització de anonima.fa és chr21:43659509-43697079 (+) en la versió hg38 del genoma humà (a la imatge la veiem com una línia densa i negra).

### 7. Conversió de les prediccions a format GFF. Crear un *custom track*.

Calculem manualment les posicions:

```
browser position chr21: 43659509-43697079
browser hide all
track name=Genscan description="AMS genscan local prediction" visibility=2
chr21    Genscan    first    43659670    43659799    33.81    +    2    gen1
chr21    Genscan    internal  43669889    43669971    2.96    +    2    gen1
chr21    Genscan    internal  43672313    43672370    6.66    +    0    gen1
chr21    Genscan    internal  43673875    43673960    1.04    +    1    gen1
chr21    Genscan    internal  43674641    43674702    1    +    0    gen1
chr21    Genscan    internal  43675039    43675168    6.5    +    0    gen1
chr21    Genscan    internal  43676277    43676341    3.32    +    1    gen1
chr21    Genscan    internal  43676738    43676919    17.9    +    1    gen1
chr21    Genscan    internal  43683284    43683378    0.68    +    2    gen1
chr21    Genscan    internal  43684558    43684655    3.11    +    2    gen1
chr21    Genscan    internal  43685775    43685794    1    +    2    gen1
chr21    Genscan    internal  43686809    43686940    11.22    +    2    gen1
chr21    Genscan    internal  43687176    43687364    2.96    +    0    gen1
chr21    Genscan    internal  43687521    43688245    38.55    +    0    gen1
chr21    Genscan    internal  43689749    43689893    1.26    +    0    gen1
chr21    Genscan    internal  43690102    43690184    1    +    1    gen1
chr21    Genscan    internal  43690293    43690445    13.67    +    1    gen1
chr21    Genscan    internal  43691444    43691507    10.39    +    0    gen1
chr21    Genscan    terminal  43693195    43693388    9.38    +    1    gen1
```

```
browser position chr21: 43659509-43697079
browser hide all
track name=Geneid description="AMS geneid local prediction" visibility=2
chr21    Geneid    first    43659665    43659794    9.81    +    0    gen1
chr21    Geneid    internal  43669884    43669966    1.45    +    2    gen1
chr21    Geneid    internal  43672308    43672365    1    +    0    gen1
chr21    Geneid    internal  43675012    43675163    1    +    2    gen1
chr21    Geneid    internal  43676272    43676336    1.03    +    0    gen1
chr21    Geneid    internal  43676733    43676914    5.73    +    1    gen1
chr21    Geneid    internal  43683279    43683373    1    +    2    gen1
chr21    Geneid    internal  43684553    43684650    2.96    +    0    gen1
chr21    Geneid    internal  43685770    43685789    2.17    +    1    gen1
chr21    Geneid    internal  43686804    43686935    2.7    +    2    gen1
chr21    Geneid    terminal  43687516    43688366    6.2    +    2    gen1
chr21    Geneid    first    43690026    43690037    1    +    0    gen2
chr21    Geneid    internal  43690288    43690440    1    +    0    gen2
chr21    Geneid    internal  43691439    43691502    2.93    +    0    gen2
chr21    Geneid    terminal  43693190    43693383    1    +    2    gen2
```

```
browser position chr21: 43659509-43697079
browser hide all
track name=fgenesh description="AMS fgenesh local prediction" visibility=2
chr21    fgenesh    first    43659665    43659794    28.3    +    0    gen1
chr21    fgenesh    internal  43669884    43669966    7.43    +    2    gen1
chr21    fgenesh    internal  43672308    43672365    6.33    +    0    gen1
chr21    fgenesh    internal  43673870    43673955    3.34    +    2    gen1
chr21    fgenesh    internal  43674636    43674697    2.4    +    0    gen1
chr21    fgenesh    internal  43675034    43675163    6.39    +    1    gen1
chr21    fgenesh    internal  43676272    43676336    6.62    +    0    gen1
chr21    fgenesh    internal  43676733    43676914    12.24    +    1    gen1
chr21    fgenesh    internal  43683279    43683373    2.1    +    2    gen1
chr21    fgenesh    internal  43684553    43684650    1    +    0    gen1
```

chr21	fgenesh	internal	43685770	43685789	1	+	1	gen1
chr21	fgenesh	internal	43686804	43686935	8.29	+	2	gen1
chr21	fgenesh	internal	43687516	43688366	33.29	+	2	gen1
chr21	fgenesh	internal	43690288	43690440	9.89	+	0	gen1
chr21	fgenesh	internal	43691439	43691502	13.04	+	0	gen1
chr21	fgenesh	terminal	43693190	43693383	1.9	+	2	gen1

Un cop tenim les nostres dades en format GFF creem les nostres *custom tracks* a genomeBrowser d'UCSC:

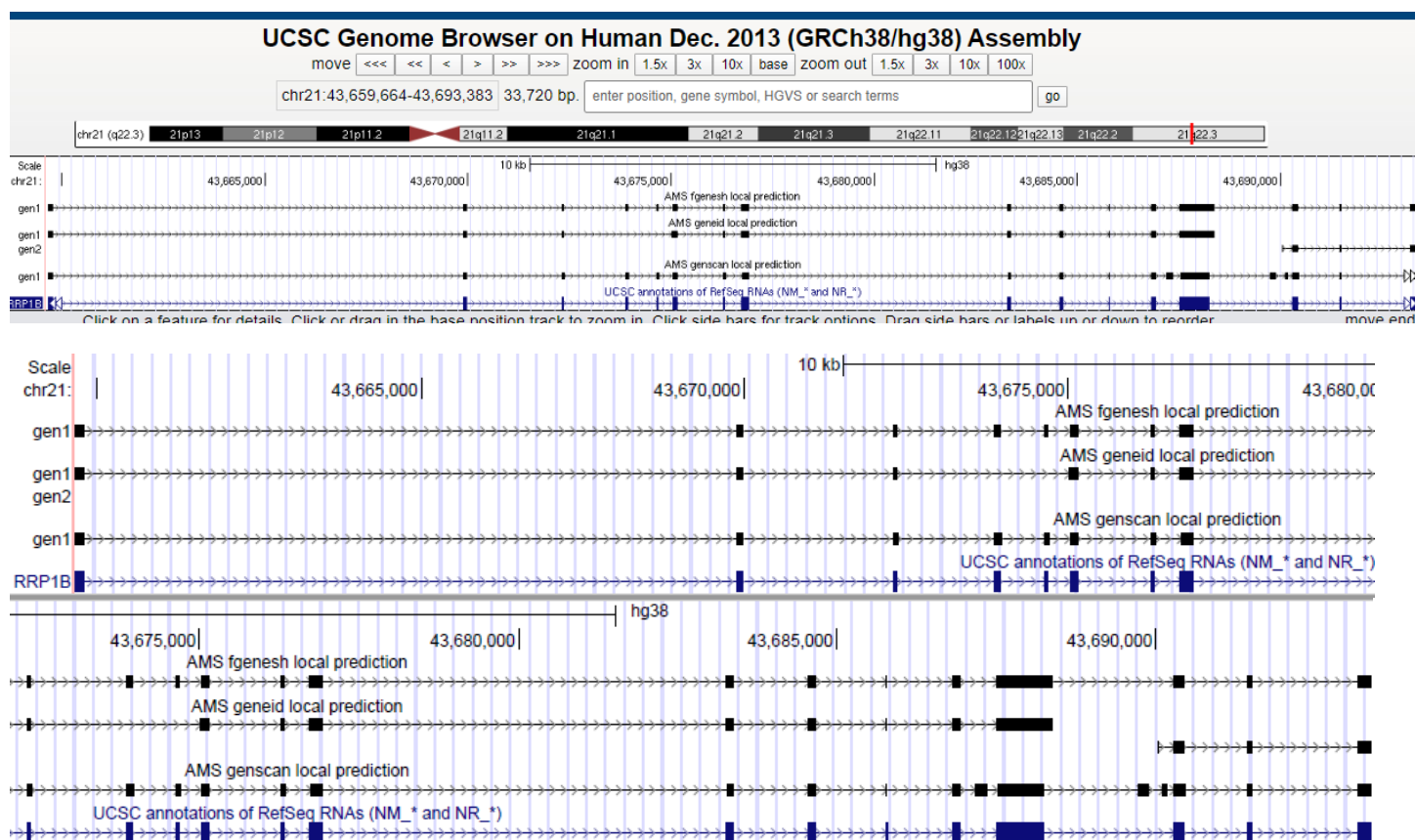
**Manage Custom Tracks**

genome: Human assembly: Dec. 2013 (GRCh38/hg38) [hg38]

Name	Description	Type	Doc	Items	Pos	delete
fgenesh	AMS fgenesh local prediction	gff		1	chr21:	<input type="checkbox"/>
Geneid	AMS geneid local prediction	gff		2	chr21:	<input type="checkbox"/>
Genscan	AMS genscan local prediction	gff		1	chr21:	<input type="checkbox"/>

view in:

I si les visualitzem quedarien així:



## 8. Correlació entre les prediccions amb Table Browser.

### a. GENEID-GENSCAN

Introduim les dades a TableBrowser de genomeBrowser d'UCSC. La regió genòmica que delimita *anonima.fa* és:

chr21:43659509-43697079. Busquem la correlació:

Correlate table 'Genscan' (ct\_Genscan\_4163) with table 'ct\_Geneid\_6038'

Select a group, track and table to correlate with:

group:  track:

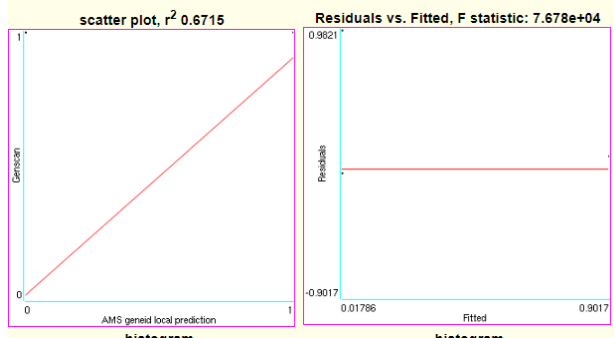
table:

Limit total data points in result:  Window data to:  bases

position: chr21:43,659,509-43,697,079 bases: 37,571

Position and # of data points in intersection	Correlation coefficient r	r <sup>2</sup>	Track	Minimum	Maximum	Mean	Variance	Standard deviation	Regression line y = m*x + b
chr21:43,659,509-43,697,079	0.8194	0.6715	Genscan	0	1	0.0717	0.06656	0.258	0.8838
37,571 data points			Geneid	0	1	0.06092	0.05721	0.2392	0.01786

Podem veure que els residus són constants :



## b. GENEID-FGENESH

Correlate table 'fgenes' (ct\_fgenes\_9515) with table 'ct\_Geneid\_6038'

Select a group, track and table to correlate with:

group: Custom Tracks track: Geneid

table: ct\_Geneid\_6038

Limit total data points in result: 40,000,000 Window data to: 1 bases

calculate clear selections return to table browser

position: chr21:43,659,509-43,697,079 bases: 37,571

Position and # of data points in intersection	Correlation coefficient r	r <sup>2</sup>	Track	Minimum	Maximum	Mean	Variance	Standard deviation	Regression line y = m*x + b
chr21:43,659,509-43,697,079 37,571 data points	0.959	0.9196	fgenes	0	1	0.06396	0.05987	0.2447	0.981 0.004195
			Geneid	0	1	0.06092	0.05721	0.2392	

scatter plot, r<sup>2</sup> 0.9196

Residuals vs. Fitted, F statistic: 4.296e+05

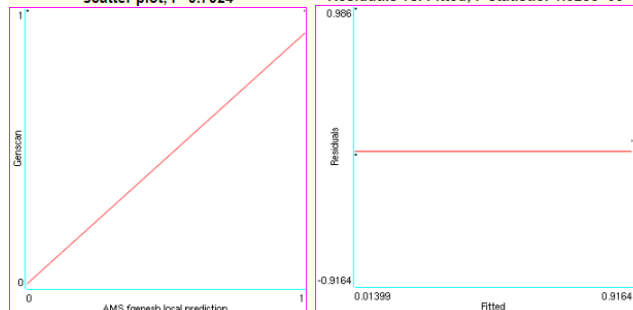
## c. GENSCAN-FGENESH

position: chr21:43,659,509-43,697,079 bases: 37,571

Position and # of data points in intersection	Correlation coefficient r	r <sup>2</sup>	Track	Minimum	Maximum	Mean	Variance	Standard deviation	Regression line y = m*x + b
chr21:43,659,509-43,697,079 37,571 data points	0.8558	0.7324	Genescan	0	1	0.0717	0.06656	0.258	0.9024 0.01399
			fgenes	0	1	0.06396	0.05987	0.2447	

scatter plot, r<sup>2</sup> 0.7324

Residuals vs. Fitted, F statistic: 1.028e+05



## d. GENEID-RRP1B (NCBI RefSeq)

### Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to calculate enrichment scores. For a description of the controls in this form, and the [User's Guide](#) for general information, see the [User's Guide](#). To examine the biological function of your set through annotation enrichments, send the data to the [Sequence and Annotation Downloader](#). All tables can be downloaded in their entirety from the [Sequence and Annotation Downloader](#).

clade: Mammal genome: Human assembly: Dec. 2013 (GRCh38/hg38)

group: Custom Tracks track: Geneid manage custom tracks track hubs

table: ct\_Geneid\_6038 describe table schema

region: genome position chr21:43,659,509-43,697,079 lookup define regions

identifiers (names/accessions): paste list upload list

filter: create

intersection: create

correlation: calculate clear (with: RefSeq All)

output format: all fields from selected table Send output to Galaxy GREAT

output file: (leave blank to keep output in browser)

file type returned: plain text gzip compressed

### Correlate table 'Geneid' (ct\_Geneid\_6038) with table 'ncbiRefSeq'

Select a group, track and table to correlate with:

group: Genes and Gene Predictions track: NCBI RefSeq

table: RefSeq All (ncbiRefSeq)

Limit total data points in result: 40,000,000 Window data to: 1 bases

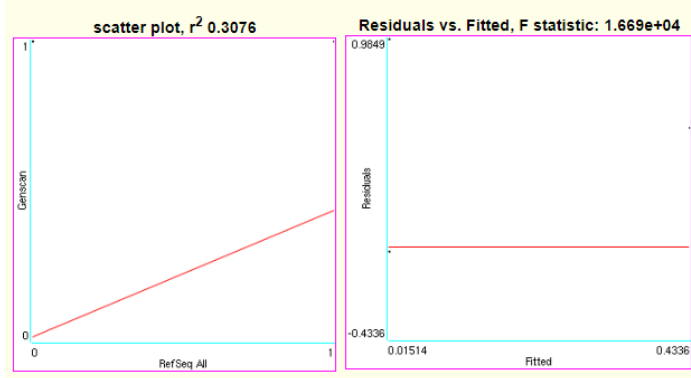
calculate clear selections return to table browser

position: chr21:43,659,509-43,697,079 bases: 37,571

Position and # of data points in intersection	Correlation coefficient r	r <sup>2</sup>	Track	Minimum	Maximum	Mean	Variance	Standard deviation	Regression line y = m*x + b
chr21:43,659,509-43,697,079	0.5922	0.3507	Geneid	0	1	0.06092	0.05721	0.2392	0.4143 0.004924
37,571 data points			RefSeq All	0	1	0.1352	0.1169	0.3419	

### e. GENSCAN-RRP1B (NCBI RefSeq)

Position and # of data points in intersection	Correlation coefficient r	r <sup>2</sup>	Track	Minimum	Maximum	Mean	Variance	Standard deviation	Regression line y = m*x + b
chr21:43,659,509-43,697,079	0.5546	0.3076	GenScan	0	1	0.0717	0.06656	0.258	0.4185 0.01514
37,571 data points			RefSeq All	0	1	0.1352	0.1169	0.3419	



### f. FGENESH-RRP1B

#### Correlate table 'fgenesh' (ct\_fgenesh\_9515) with table 'ncbiRefSeq'

Select a group, track and table to correlate with:

group: Genes and Gene Predictions track: NCBI RefSeq

table: RefSeq All (ncbiRefSeq)

Limit total data points in result: 40,000,000 Window data to: 1 bases

calculate clear selections return to table browser

position: chr21:43,659,509-43,697,079 bases: 37,571

Position and # of data points in intersection	Correlation coefficient r	r <sup>2</sup>	Track	Minimum	Maximum	Mean	Variance	Standard deviation	Regression line y = m*x + b
chr21:43,659,509-43,697,079	0.6211	0.3858	fgenesh	0	1	0.06396	0.05987	0.2447	0.4445 0.003878
37,571 data points			RefSeq All	0	1	0.1352	0.1169	0.3419	

### 9. Alineament múltiple global de les 3 proteïnes amb RRP1B (CLUSTAL). Quin programa de predicció és millor?

Introduïm ara a CLUSTAL les 3 prediccions obtingudes i la seqüència d'aa de RRP1B que obtenim del navegador Genome Browser. La proteïna obtinguda per GENEID serà la resultant de sumar el pèptid del gen 1 i el del gen 2, ja que ara sabem que es tracta d'un sol gen. D'el resultat de l'alineament, per qüestions d'espai, només mostrarem les regions on hi divergència.

#### Summary of RRP1B

#### mRNA/Genomic Alignments

BROWSER	SIZE	IDENTITY	CHROMOSOME	STRAND	START	END	QUERY	START	END	TOTAL
browser	5078	100.0%	21	+	43659560	43696079	NM_015056	1	5078	5078

View details of parts of alignment within browser window.

Position: chr21:43659560-43696079

Band: 21q22.3

Genomic Size: 36520

Strand: +

Gene Symbol: RRP1B

CDS Start: complete

CDS End: complete

#### Links to sequence:

- Predicted Protein



Observem a la matriu d'identitat que les proteïnes de FGENESH i GENSCAN tenen un 100% d'identitat amb RRP1B. En canvi, amb GENEID un 99.3%, que és un percentatge també molt alt.

Percent Identity Matrix - created by Clustal2.1

```

1: human_1|geneid_v1.2_predicted_protein_1|622_AA      100.00   99.30   99.30   94.35
2: FGENESH_                                           99.30   100.00  100.00  100.00
3: NP_055871                                           99.30   100.00  100.00  100.00
4: /tmp/05_08_20-11_44_37.fasta|GENSCAN_predicted_peptide_1|897_aa 94.35  100.00  100.00  100.00

```

Observem ara on són les diferències entre GENEID i RRP1B:

```

human_1|geneid_v1.2_predicted_protein_1|622_AA      CMWVQDEPLLQEEELANTIAQLVHAVNNSAAQAC----- 93
FGENESH:      CMWVQDEPLLQEEELANTIAQLVHAVNNSAAQHLLFQTFWQTMNREWKIDRLRLDKYYML 120
NP_055871     CMWVQDEPLLQEEELANTIAQLVHAVNNSAAQHLLFQTFWQTMNREWKIDRLRLDKYYML 120
/tmp/05_08_20-11:44:37.fasta|GENSCAN_predicted_peptide_1|897_aa  CMWVQDEPLLQEEELANTIAQLVHAVNNSAAQHLLFQTFWQTMNREWKIDRLRLDKYYML 120
*****

human_1|geneid_v1.2_predicted_protein_1|622_AA      -----VWFFSRIKVFLDVLMKEVLCPEQSPNGVRFHFIDYLDLSKVG 138
FGENESH:      IRLVLRQSFVLRKRWGEESRIKVFLDVLMKEVLCPEQSPNGVRFHFIDYLDLSKVG 180
NP_055871     IRLVLRQSFVLRKRWGEESRIKVFLDVLMKEVLCPEQSPNGVRFHFIDYLDLSKVG 180
/tmp/05_08_20-11:44:37.fasta|GENSCAN_predicted_peptide_1|897_aa  IRLVLRQSFVLRKRWGEESRIKVFLDVLMKEVLCPEQSPNGVRFHFIDYLDLSKVG 180
*****

```

Sembla ser que a GENEID i ha omès d'un exó o regió primer. A continuació observem, a la part final de la proteïna, que GENEID té un parell de regions extremes també.

```

human_1|geneid_v1.2_predicted_protein_1|622_AA      KMRVMSNLVEHNGVLESEAGQPQALVRWEHPQA-----SSPQRHSL-ASMG 600
FGENESH:      KMRVMSNLVEHNGVLESEAGQPQAL----- 622
NP_055871     KMRVMSNLVEHNGVLESEAGQPQAL----- 622
/tmp/05_08_20-11:44:37.fasta|GENSCAN_predicted_peptide_1|897_aa  KMRVMSNLVEHNGVLESEAGQPQALAAHLNLPPEPPVCRQRHWAHTSESQVRDPVSLMVA 720
*****

human_1|geneid_v1.2_predicted_protein_1|622_AA      LHCLLRGRV-----GAGGQASGLSSSMKIKGSSGTCSSLKQKLAESD 644
FGENESH:      -----GSSGTCSSLKQKLAESD 641
NP_055871     -----GSSGTCSSLKQKLAESD 641
/tmp/05_08_20-11:44:37.fasta|GENSCAN_predicted_peptide_1|897_aa  VSCCTRNECPGPASVVLCKVPELCRMGLSASAVRKTAGRRGSSGTCSSLKQKLAESD 780
*****

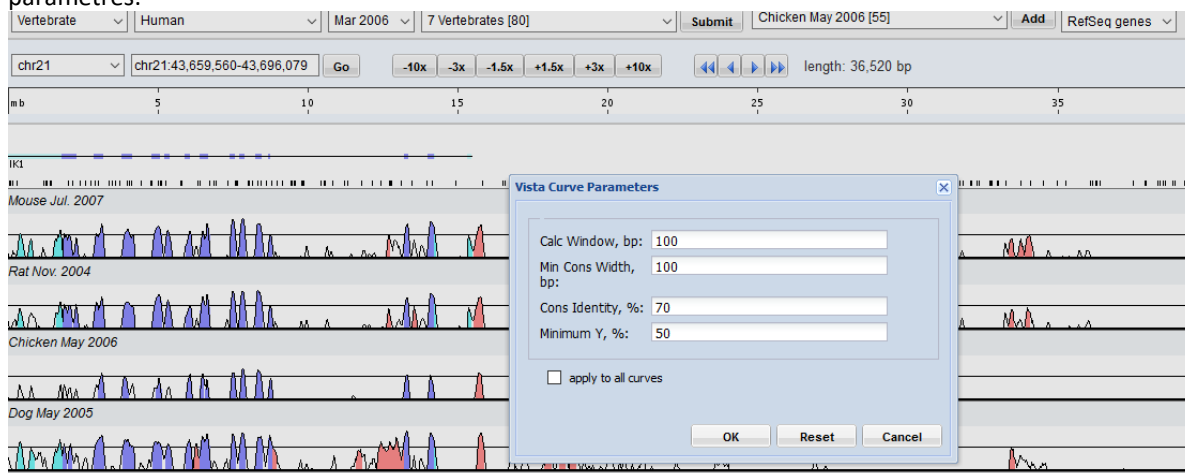
```

Com a conclusió podem dir que el que ha fet la millor predicció en aquest cas ha estat FGENESH, tant pels resultats de la matriu d'identitat, com pel valor de la correlació amb RRP1B.

## 10. VISTA: grau de conservació dels exons de RRP1B.

VISTA ens proporciona un alineament per homologia, reconstruint la estructura exònica d'un gen a partir de l'identificació de proteïnes homòlogues conservades en altres espècies. Això ens resultarà útil per guiar la predicció *ab initio*, identificant primer el conjunt de proteïnes conegudes en altres espècies que estan conservades. Així després podem informar a GENEID, GENSCAN o FGENESH sobre l'existència d'homologia amb proteïnes conegudes. Per exemple, podem millorar la predicció a GENEID augmentant la puntuació de aquells exons identificats de novo que presenten solapament amb les regions conservades en tots dos genomes. També ens serveix per reforçar les prediccions obtingudes per les estratègies *de novo*.

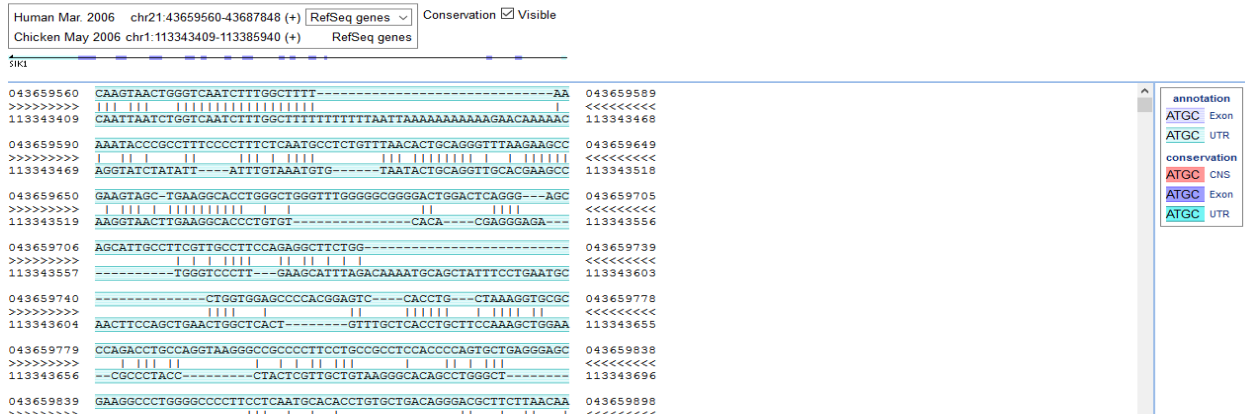
Els gràfics representen el percentatge de conservació entre seqüències alineades a una determinada coordenada de la seqüència base. La pista d'anotació gènica es mostra per sobre de les corbes de conservació, on les caixes de color blau fosc i clar representen exons i UTRs respectivament. Les regions d'alta conservació es coloreen segons l'anotació com a exons (blau fosc), UTR (blau clar) o sense codificació (rosa). Les regions es classifiquen en "conservades" mitjançant l'anàlisi de puntuacions de cada pb en l'interval genòmic, és a dir, "Amplada mínima conservada" (valor predeterminat 100 bp) i "Identitat de conservació" (valor predeterminat 70%). Es considera conservada una regió si la conservació sobre aquesta regió és superior o igual a la "Identitat de conservació" i té la longitud mínima de "Amplada mínima conservada". Podem configurar tots aquests paràmetres:





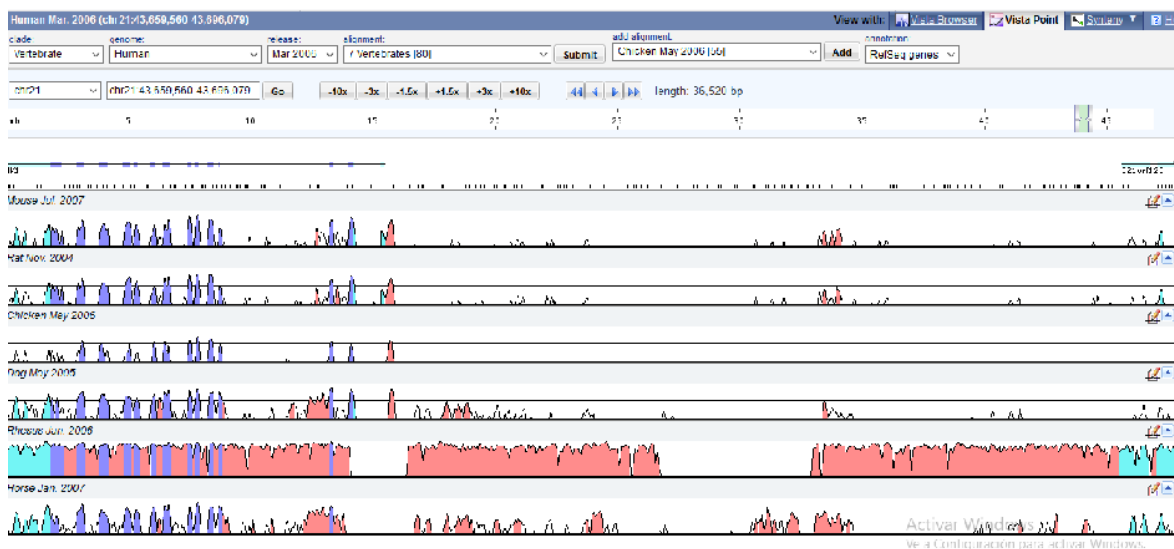
Seleccionem la versió del 2006 perquè conté més informació i, RefSeq genes a la llista d'anotacions. El programa ens permet descarregar un pdf amb la llista de tots els SNPs (veure [annex13](#)).

VISTA ens permet també fer alineaments entre espècies. Per exemple, amb pollastre (veure un altre alineament amb ratolí i *Rhesus* a [annex 14 i 15](#), respectivament).



Cal dir, que no es poden detectar elements funcionals curts com a estadísticament significatius en comparacions d'espècies molt properes. Un exemple extrem: ja que el genoma humà i el ximpanzé són un 98,7% idèntics fins i tot a les regions neutres, la gran majoria dels exons són massa breus per destacar com a estadísticament significatius. En general, la potència estadística per detectar elements funcionals de restricció curta augmenta a mesura que augmenta la divergència neutra total de l'espècie comparada.

En general, podem observar que les regions exòniques del gen RRP1B estan molt conservades a totes les espècies (veure [annex16](#))



Annex 1a. Seqüències CDS.

Humà (variant 1 i 2, respectivament):

```
>hg38_refGene_NM_018374 range=chr7:12214811-12231975 5'pad=0 3'pad=0 strand=+ repeatMasking=none
ATGGGAAAGTCTCTTTCTCATTTCGCTTTGCATTCAAGCAAAGAAGATGC
TTATGATGGAGTCACATCTGAAAACATGAGGAATGGACTGGTTAATAGTG
AAGTCCATAATGAAGATGGAAGAAATGGAGATGTCTCTCAGTTTCCATAT
GTGGAATTTACAGGAAGAGATAGTGTCACTGCCCTACTTGTCAAGGGAAC
AGGAAGAATTCCTAGGGGGCAAGAAAACCAACTGGTGGCATTGATTCCAT
ATAGTGATCAGAGATTAAGGCCAAGAAGAACAAAGCTGTATGTGATGGCT
TCTGTGTTTGTCTGTCTACTCCTTTCTGGATTGGCTGTGTTTTTCCTTTT
CCCTCGCTCTATCGACGTGAAATACATTGGTGTAATAATCAGCCTATGTCA
GTTATGATGTTTCAAGCGTACAATTTATTTAAATATCACAACACACTA
AATATAACAAACAATAACTATTACTCTGTGCGAAGTTGAAAACATCACTGC
CCAAGTTCATTTTCAAAAACAGTTATTGGAAAGGCACGCTTAAACAACA
TAACCATTTATTGGTCCACTTGATATGAAACAAATTGATTACACAGTACCT
ACCGTTATAGCAGAGGAAATGAGTTATATGTATGATTTCTGTACTCTGAT
ATCCATCAAAGTGCATAACATAGTACTCATGATGCAAGTTACTGTGACAA
CAACATACTTTGGCCACTCTGAACAGATATCCCAGGAGAGGTATCAGTAT
GTCGACTGTGGAAGAAACACAACCTTATCAGTTGGGGCAGTCTGAATATTT
AAATGTACTTCAGCCACAACAGTAA
```

```
>hg38_refGene_NM_001134232 range=chr7:12214811-12231975 5'pad=0 3'pad=0 strand=+ repeatMasking=none
ATGGGAAAGTCTCTTTCTCATTTCGCTTTGCATTCAAGCAAAGAAGATGC
TTATGATGGAGTCACATCTGAAAACATGAGGAATGGACTGGTTAATAGTG
AAGTCCATAATGAAGATGGAAGAAATGGAGATGTCTCTCAGTTTCCATAT
GTGGAATTTACAGGAAGAGATAGTGTCACTGCCCTACTTGTCAAGGGAAC
AGGAAGAATTCCTAGGGGGCAAGAAAACCAACTGGTGGCATTGATTCCAT
ATAGTGATCAGAGATTAAGGCCAAGAAGAACAAAGCTGTATGTGATGGCT
TCTGTGTTTGTCTGTCTACTCCTTTCTGGATTGGCTGTGTTTTTCCTTTT
CCCTCGCTCTATCGACGTGAAATACATTGGTGTAATAATCAGCCTATGTCA
GTTATGATGTTTCAAGCGTACAATTTATTTAAATATCACAACACACTA
AATATAACAAACAATAACTATTACTCTGTGCGAAGTTGAAAACATCACTGC
CCAAGTTCATTTTCAAAAACAGTTATTGGAAAGGCACGCTTAAACAACA
TAACCATTTATTGGTCCACTTGATATGAAACAAATTGATTACACAGTACCT
ACCGTTATAGCAGAGGAAATGAGTTATATGTATGATTTCTGTACTCTGAT
ATCCATCAAAGTGCATAACATAGTACTCATGATGCAAGTTACTGTGACAA
CAACATACTTTGGCCACTCTGAACAGATATCCCAGGAGAGGTATCAGTAT
GTCGACTGTGGAAGAAACACAACCTTATCAGTTGGGGCAGTCTGAATATTT
AAATGTACTTCAGCCACAACAGTAA
```

Ratolí:

```
>mm10_refGene_NM_027992 range=chr6:13071744-13084326 5'pad=0 3'pad=0 strand=+ repeatMasking=none
ATGGGAAAGTCTCTTTCTCACTTACCTTTGCATTCAAATAAAGAAGATGG
CTATGATGGCGTTACATCGACAGACAATATGAGAAATGGATTGGTTAGCA
GTGAAGTGCACAACGAAGACGGAAGAAATGGAGATGTCTCTCAGTTCCCA
TATGTGGAATTTACTGGAAGAGATAGTGTCACTTGTCCCACTTGCCAAGG
AACAGGAAGAATTCCTAGGGGACAAGAAAACCAACTGGTGGCATTGATTC
CATATAGTGATCAGCGGTTACGGCCAAGAAGAACAAAGCTGTATGTGATG
GCGTCTGTGTTTGTCTGCCTGCTCCTGTCTGGATTGGCTGTGTTTTTCT
TTTCCCTCGATCTATTGAGGTGAAGTACATTGGAGTAAATCAGCCTATG
TCAGCTACGACGCTGAAAAGCGAACCATATATTTAAATATCACGAACACA
CTAAATATAACAATAATAACTATTATTCTGTTGAAGTTGAAAACATCAC
TGCTCAAGTCCAGTTTTCAAAACCGTGATTGGAAGGCTCGTTTAAACA
ACATAACTAACATTGGCCCACTTGATATGAAGCAGATTGATTATACGGTA
CCCACAGTTATTGCAGAGGAAATGAGTTACATGTATGATTTCTGTACACT
GCTCTCCATCAAAGTGCACAACATAGTACTCATGATGCAAGTTACTGTAA
CAACAGCATACTTTGGACACTCTGAGCAGATATCTCAGGAAAGGTACCAG
TATGTGCGACTGTGGAAGGAACACGACTTACCAGTTGGCCCACTGAGTA
TCTAAATGTCCTTCAGCCACAACAATAA
```

Annex 1b. Resultat alineament en CLUSTAL.

CLUSTAL O(1.2.4) multiple sequence alignment

hg38_refGene_NM_018374	ATGGGAAAGTCTCTTTCTCATTTCGCTTTGCATTCAAGCAAAGAAGATGCTTATGATGGA	60
mm10_refGene_NM_027992	ATGGGAAAGTCTCTTTCTCACTTACCTTTGCATTCAAATAAAGAAGATGGCTATGATGGC	60
	*****	
hg38_refGene_NM_018374	GTCACATCT---GAAAACATGAGGAATGGACTGGTTAATAGTGAAGTCCATAATGAAGAT	117
mm10_refGene_NM_027992	GTTACATCGACAGACAATATGAGAAATGGATTGGTTAGCAGTGAAGTGCACACGAAGAC	120
	** ***** ** * ***** ***** ***** ***** ** * *****	
hg38_refGene_NM_018374	GGAAGAAATGGAGATGTCTCTCAGTTTCCATATGTGGAATTTACAGGAAGAGATAGTGTCT	177
mm10_refGene_NM_027992	GGAAGAAATGGAGATGTCTCTCAGTTCCCATATGTGGAATTTACTGGAAGAGATAGTGTCT	180

*****		
hg38_refGene_NM_018374	ACCTGCCCTACTTGTGTCAGGGAACAGGAAGAATTCCTAGGGGGCAAGAAAACCAACTGGTG	237
mm10_refGene_NM_027992	ACTTGTCCCACTTGCCAAGGAACAGGAAGAATTCCTAGGGGACAAGAAAACCAACTGGTG	240
** ** *		
hg38_refGene_NM_018374	GCATTGATTCCATATAGTGATCAGAGATTAAGGCCAAGAAGAACAAAGCTGTATGTGATG	297
mm10_refGene_NM_027992	GCATTGATTCCATATAGTGATCAGCGTTACGGCCAAGAAGAACAAAGCTGTATGTGATG	300
*****		
hg38_refGene_NM_018374	GCTTCTGTGTTTGTCTGTCTACTCCTTCTGGATTGGCTGTGTTTTTCTTTTCCCTCGC	357
mm10_refGene_NM_027992	GCGTCTGTGTTTGTCTGCGCTGCTCCTGTCTGGATTGGCTGTGTTTTTCTTTTCCCTCGA	360
** *****		
hg38_refGene_NM_018374	TCTATCGACGTGAAATACATTGGTGTAATAACAGCCTATGTCAGTTATGATGTTTCAAG	417
mm10_refGene_NM_027992	TCTATTGAGGTGAAGTACATTGGAGTAAATACAGCCTATGTCAGCTACGACGCTGAAAAG	420
***** ** *****		
hg38_refGene_NM_018374	CGTACAATTTATTTAAATATCACAAACACACTAAATATAACAACAATAACTATTACTCT	477
mm10_refGene_NM_027992	CGAACCATATATTTAAATATCACGAACACACTAAATATAACAATAATAACTATTATCT	480
** ** *		
hg38_refGene_NM_018374	GTCGAAGTTGAAAACATCAGTCCCAAGTTCAATTTTCAAAAACAGTTATTGGAAGGCA	537
mm10_refGene_NM_027992	GTTGAAGTTGAAAACATCAGTCTCAAGTCCAGTTTTCAAAAACCGTGATTGGAAAGGCT	540
** *****		
hg38_refGene_NM_018374	CGCTTAAACAACATAACCATTTATGGTCCACTTGATATGAAACAAATTGATTACACAGTA	597
mm10_refGene_NM_027992	CGTTTAAACAACATAACTAACATTGGCCCACTTGATATGAAGCAGATTGATTATACGGTA	600
** ***** *		
hg38_refGene_NM_018374	CCTACCGTTATAGCAGAGGAAATGAGTTATATGTATGATTTCTGTACTCTGATATCCATC	657
mm10_refGene_NM_027992	CCCACAGTTATGTCAGAGGAAATGAGTTACATGTATGATTTCTGTACTCTGCTCTCCATC	660
** ** *****		
hg38_refGene_NM_018374	AAAGTGCATAACATAGTACTCATGATGCAAGTTACTGTGACAACAACATACTTTGGCCAC	717
mm10_refGene_NM_027992	AAAGTGCACAACATAGTACTCATGATGCAAGTTACTGTAACAACAGCATACTTTGGACAC	720
***** *****		
hg38_refGene_NM_018374	TCTGAACAGATATCCCAGGAGAGGTATCAGTATGTCGACTGTGGAAGAAACACAACCTTAT	777
mm10_refGene_NM_027992	TCTGAGCAGATATCTCAGGAAAGGTACCAGTATGTCGACTGTGGAAGGAACACGACTTAC	780
***** *****		
hg38_refGene_NM_018374	CAGTTGGGGCAGTCTGAATATTTAAATGTACTTCAGCCACAACAGTAA 825	
mm10_refGene_NM_027992	CAGTTGGCCCCAGTCTGAGTATCTAAATGTCTTCAGCCACAACAATAA 828	
***** *****		

## Annex 2. Seqüències aa.

### Humà:

```
>NP_001127704 length=274
MGKSLSHLPLHSSKEDAYDGVTSENMRNGLVNSEVHNEDGRNGDVSQFPY
VEFTGRDSVTCPTCQGTGRIPRGQENQLVALIPYSDQRLRPRRTKLYVMA
SVFVCLLLSGLAVFFLFPRSIDVKYIGVKSAYVSYDVQKRTIYLNITNTL
NITNNNYSVEVENITAQVQFSKTVIGKARLNNITIIGPLDMKQIDYTV
TVIAEEMSYMYDFCTLISIKVHNIVLMMQVTVTTTYFGHSEQISQERYQY
VDCGRNTTYQLGQSEYLNVLQPQQ
>NP_060844 length=274
MGKSLSHLPLHSSKEDAYDGVTSENMRNGLVNSEVHNEDGRNGDVSQFPY
VEFTGRDSVTCPTCQGTGRIPRGQENQLVALIPYSDQRLRPRRTKLYVMA
SVFVCLLLSGLAVFFLFPRSIDVKYIGVKSAYVSYDVQKRTIYLNITNTL
NITNNNYSVEVENITAQVQFSKTVIGKARLNNITIIGPLDMKQIDYTV
TVIAEEMSYMYDFCTLISIKVHNIVLMMQVTVTTTYFGHSEQISQERYQY
VDCGRNTTYQLGQSEYLNVLQPQQ
```

### Ratolí:

```
>NP_082268 length=275
MGKSLSHLPLHSSKEDGYDGVTSNDNRNGLVSEVHNEDGRNGDVSQFP
YVEFTGRDSVTCPTCQGTGRIPRGQENQLVALIPYSDQRLRPRRTKLYVM
ASVFVCLLLSGLAVFFLFPRSIEVKYIGVKSAYVSYDAEKRTIYLNITNT
LNITNNNYSVEVENITAQVQFSKTVIGKARLNNITNIGPLDMKQIDYTV
PTVIAEEMSYMYDFCTLLSIKVNIVLMMQVTVTTAYFGHSEQISQERYQ
YVDCGRNTTYQLAQSEYLNVLQPQQ
```

### Annex 3. Resultat blastn.

RID: B5G0BE69114

Job Title:NM\_001134232:Homo sapiens transmembrane protein...

Program: BLASTN

Subject:Mus musculus transmembrane protein 106B (Tmem106b), mRNA ID: NM\_027992.3(nucleic acid) Length: 6099

Query #1: Homo sapiens transmembrane protein 106B (TMEM106B), transcript variant 2, mRNA Query ID: ref|NM\_001134232.2 Length: 12351

Sequences producing significant alignments:

	Max	Total Query	E	Per.		
Description	Score	Score	cover	Value	Ident	Accession
Mus musculus transmembrane protein 106B (Tmem106b), mRNA	1501	2771	37%	0.0	76.41	NM_027992.3

Alignments:

>Mus musculus transmembrane protein 106B (Tmem106b), mRNA

Sequence ID: NM\_027992.3 Length: 6099

Range 1: 310 to 2346

Score:1501 bits(1664), Expect:0.0,

Identities:1623/2124(76%), Gaps:126/2124(5%), Strand: Plus/Plus

Query 133 ACATGGGAAAGTCTCTTTCTCATTGCCTTTGCATTCAAGCAAAGAAGATGCTTATGATG 192

|||||

Sbjct 310 ACATGGGAAAGTCTCTTTCTCACTTACCTTTGCATTCAAATAAAGAAGATGGCTATGATG 369

Query 193 GAGTCACATCT---GAAAACATGAGGAATGGACTGGTTAATAGTGAAGTCCATAATGAAG 249

|||||

Sbjct 370 GCGTTACATCGACAGACAATATGAGAAATGGATTGGTTAGCAGTGAAGTGCACAACGAAG 429

Query 250 ATGGAAGAAATGGAGATGTCTCTCAGTTTCCATATGTGGAATTTACAGGAAGAGATAGTG 309

|||||

Sbjct 430 ACGGAAGAAATGGAGATGTCTCTCAGTTCCCATATGTGGAATTTACTGGAAGAGATAGTG 489

Query 310 TCACCTGCCCTACTTGTCTCAGGGAACAGGAAGAATTCCTAGGGGGCAAGAAAACCAACTGG 369

|||||

Sbjct 490 TCACTTGTCCCACTTGCCAAGGAACAGGAAGAATTCCTAGGGGACAAGAAAACCAACTGG 549

Query 370 TGGCATTGATTCCATATAGTGATCAGAGATTAAGGCCAAGAAGAACAAGCTGTATGTGA 429

|||||

Sbjct 550 TGGCATTGATTCCATATAGTGATCAGCGGTTACGGCCAAGAAGAACAAGCTGTATGTGA 609

Sbjct 610 TGGCGTCTGTGTTTGTCTGCCTGCTCCTGTCTGGATTGGCTGTGTTTTTCTTTCCCTC 669

Sbjct 670 GATCTATTGAGGTGAAGTACATTGGAGTAAAATCAGCCTATGTCAGCTACGACGCTGAAA 729

Sbjct 730 AGCGAACCATATATTTAAATATCACGAACACACTAAATATAACAAATAATAACTATTATT 789

Sbjct 790 CTGTTGAAGTTGAAAACATCACTGCTCAAGTCCAGTTTTCAAAAACCGTGATTGGAAAGG 849

Sbjct 850 CTCGTTTAAACAACATAACTAACATTGGCCCACTTGATATGAAGCAGATTGATTATACGG 909

Sbict 910 TACCCACAGTTATTGCAGAGGAAATGAGTTACATGTATGATTTCTGTACACTGCTCTCCA 969

Sbjct 970 TCAAAGTGCACAACATAGTACTCATGATGCAAGTTACTGTAACAACAGCATACTTTGGAC 1029

Sbjct 1030 ACTCTGAGCAGATATCTCAGGAAAGGTACCAGTATGTCGACTGTGGAAGGAACACGACTT 1089

Sbjct 1090 ACCAGTTGGCCCAGTCTGAGTATCTAAATGTCCTTCAGCCACAACAATAAACTCTAAAGG 1149

Sbict 1150 AGATGTGGTTAGAGCCGATGACACACCT-GAGACATCGTCAGAACTGTCAGTGAGGAGGT 1208

Sbjct 1729 CAACCATACCATTC--GGCCACATTCTGTTC--TACATTTTGTAGAACC-----TCTC 1778



Sbjct 1779 A-----GTATGTGTTTCATGTATAACTTTGTGAGCTTTCTGTGTGATCTTCAAACATATT 1833

|||||

Sbjct 1834 CCTTTAATGTACAATATTGTAAATAAA--GTGCATGGCTTTTATACAGCTTTGATAAAGG 1891

|||||

Sbjct 1892 TCAAATGAAGTAGTACAGATTAAGAATAAAGCAGGTTGTTCATAA---ACAATAAGGTA 1947

[illegible]

Sbjct 1948 GGAAATTGAAATTAAGAAACCTATAAAATATCAGTAGGTTCC-GCTACATGTTGGATCTG 2006

|||||      |||||

Sbjct 2007 ACATAACATAGATGCACAGTAAAACATTTTCGATGCTCTA-CTTTTTATATT----- 2056

|||||    |||||    |||||    |||||    |||||

Sbjct 2057 -----AATTATATAGATGCTTTTCCCAGCAATTAGGTA-----TCATCAATTTTATGA 2104

|||||

Sbjct 2105 ATGTTT TAGGGAAGAAAACCATTTCTTT-CTAGAAATTAGTCAAGGAAT-----AA 2154

| | | | | | | | | | | | | | | | | | | | | |

Sbjct 2155 TTTCATTGAACACTGGTGACTACATCAAATAAAACCTATAGTTTGCTTAAAAAATTTAA 2214

||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| |||||

Sbjct 2215 TCA--TTACCCATCATCCTGAGT-CGATGCTTCTAGCAGTTAC---AGTAAACATTTAT- 2267

[illegible]

Sbjct 2268 TTTTAAATTTGTCTAACATGGTTTTCAAAT----ATGATGGACGTATTCATC-TTGTTG 2322

Query 2194 ttttatattttcagtatcatttt 2217

|||||||

Sbjct 2323 CTTTATATTCCAGTGTCAATTTT 2346

RID: B5X08TP6114

Job Title:NM\_018374:Homo sapiens transmembrane protein...

Program: BLASTN

Subject:Mus musculus transmembrane protein 106B (Tmem106b), mRNA ID: NM\_027992.3(nucleic acid) Length: 6099

Query #1: Homo sapiens transmembrane protein 106B (TMEM106B), transcript variant 1, mRNA Query ID: ref|NM\_018374.4 Length: 12545

Sequences producing significant alignments:

	Max	Total Query	E	Per.	
Description	Score	Score	cover	Value	Ident Accession
Mus musculus transmembrane protein 106B (Tmem106b), mRNA	1508	2777	36%	0.0	76.43 NM_027992.3

Alignments:

>Mus musculus transmembrane protein 106B (Tmem106b), mRNA

Sequence ID: NM\_027992.3 Length: 6099

Range 1: 304 to 2346

Score:1508 bits(1671), Expect:0.0,

Identities:1628/2130(76%), Gaps:126/2130(5%), Strand: Plus/Plus

Query 321 CCTCAGACATGGGAAAGTCTCTTCTCATTGCGCTTGCATTCAAGCAAAGAAGATGCTT 380

|||||

Sbjct 304 CCTCAACATGGGAAAGTCTCTTCTCACTTACCTTGCATTCAAATAAAGAAGATGGCT 363

Query 381 ATGATGGAGTCACATCT---GAAAACATGAGGAATGGACTGGTTAATAGTGAAGTCCATA 437

|||||

Sbjct 364 ATGATGGCGTTACATCGACAGACAATATGAGAAATGGATTGGTTAGCAGTGAAGTGCACA 423

Query 438 ATGAAGATGGAAGAAATGGAGATGTCTCTCAGTTCCATATGTGGAATTTACAGGAAGAG 497

|

Sbjct 424 ACGAAGACGGAAGAAATGGAGATGTCTCTCAGTTCCATATGTGGAATTTACTGGAAGAG 483

Query 498 ATAGTGTCACCTGCCCTACTTGTGAGGGAACAGGAAGAATTCCTAGGGGGCAAGAAAACC 557

|||||||

Sbjct 484 ATAGTGTCACCTTGCCCACTTGCCAAGGAACAGGAAGAATTCCTAGGGGACAAGAAAACC 543

Query 558 AACTGGTGGCATTGATTCCATATAGTGATCAGAGATTAAGGCCAAGAAGACAAAGCTGT 617

|||||

Sbjct 544 AACTGGTGGCATTGATTCCATATAGTGATCAGCGTTACGGCCAAGAAGACAAAGCTGT 603

Query 618 ATGTGATGGCTTCTGTGTTTGTCTGTCTACTCCTTCTGGATTGGCTGTGTTTTCTTT 677

|||||

Sbjct 604 ATGTGATGGCGTCTGTGTTTGTCTGCCTGCTCCTGTCTGGATTGGCTGTGTTTTCTTT 663

Query 678 TCCCTCGCTCTATCGACGTGAAATACATTGGTGTAATCAGCCTATGTCAGTTATGATG 737

|||||

Sbjct 664 TCCCTCGATCTATTGAGGTGAAGTACATTGGAGTAAATCAGCCTATGTCAGCTACGACG 723

Query 738 TTCAGAAGCGTACAATTTATTTAAATATCACAAACACACTAAATATAACAAACAATAACT 797

|

Sbjct 724 CTGAAAAGCGAACCATATATTTAAATATCACGAACACACTAAATATAACAAATAATAACT 783

Query 798 ATTACTCTGTGCGAAGTTGAAAACATCACTGCCAAGTTCAATTTTCAAAAACAGTTATTG 857

||||

Sbjct 784 ATTATTCTGTTGAAGTTGAAAACATCACTGCTCAAGTCCAGTTTTCAAAAACCGTGATTG 843

Query 858 GAAAGGCACGCTTAAACAACATAACCATTATTGGTCCACTTGATATGAAACAAATTGATT 917

|||||

Sbjct 844 GAAAGGCTCGTTTAAACAACATAACTAACATTGGCCCACTTGATATGAAGCAGATTGATT 903

Query 918 ACACAGTACCTACCGTTATAGCAGAGGAAATGAGTTATATGTATGATTCTGTACTCTGA 977

|

Sbjct 904 ATACGGTACCCACAGTTATTGCAGAGGAAATGAGTTACATGTATGATTCTGTACTCTGC 963

Query 978 TATCCATCAAAGTGCATAACATAGTACTCATGATGCAAGTTACTGTGACAACAACATACT 1037

|

Sbjct 964 TCTCCATCAAAGTGCACAACATAGTACTCATGATGCAAGTTACTGTAACAACAGCATACT 1023

Query 1038 TTGGCCACTCTGAACAGATATCCAGGAGAGGTATCAGTATGTCGACTGTGGAAGAAACA 1097

||||

Sbjct 1024 TTGGACACTCTGAGCAGATATCTCAGGAAAGGTACCAGTATGTCGACTGTGGAAGGAACA 1083

Query 1098 CAACTTATCAGTTGGGGCAGTCTGAATATTTAAATGTACTTCAGCCACAACAGTAAAAAC 1157

|

Sbjct 1084 CGACTTACCAGTTGGCCAGTCTGAGTATCTAAATGTCCTTCAGCCACAACAATAAACTC 1143



Query 1738 ACTTATAAGCCAAAATAATCTTTGCAAAATTCATACCTAAAAATTTTGAAAGCCCCTAAT 1797

| ||| |||| | | || |||| | | || |||| || |

Sbjct 1723 ATTTACCAACCATACCATTC--GGCCACATTCTGTTC--TACATTTTGTAGAACC---- 1774

Query 1798 GTTTTCACACATCTTCTGTATTAGTTATAGTTTTGTGAAATCTTTGTGTGATCTTCAA 1857

| ||| | |||| | ||| ||||| | ||||| |||||

Sbjct 1775 --TCTCA-----GTATGTGTTTCATGTATAACTTTGTGAGCTTCTGTGTGATCTTCAA 1827

Query 1858 CATTATCATTTAATGTACAATACTGTAAATAAACTGTGCATGGCTTTTATACAGCTTTAG 1917

||| || ||||| ||||| ||||| ||||| ||||| |||||

Sbjct 1828 CATATTCCTTTAATGTACAATATTGTAAATAAA--GTGCATGGCTTTTATACAGCTTTGA 1885

Query 1918 TAAATGTCAAATAAAGTGGTACAGACT--CATTACAACAAGTTTCTCATAAAAATACAAT 1975

|||| ||||| |||| ||||| | ||| |||| ||||| |||||

Sbjct 1886 TAAAGGTCAAATGAAGTAGTACAGATTAAGAATAAAGCAGGTTGTCATAA----ACAAT 1941

Query 1976 AA-ATAGGAAAAATGAAATTCAGAAACCCATAGACTGGGAATAGGTTCCAGTTACAGCTTG 2034

| | ||||| ||||| ||||| |||| | ||||| |||| | |||

Sbjct 1942 AAGGTAGGAAATTGAAATTAAGAAACCTATAAAATATCAGTAGGTTCC-GCTACATGTTG 2000

Query 2035 GATCTGGCATAAAATAAATTTGAAATAAAATATTTTGATGCTCCAttttttATGTTGCT 2094

||||| |||| |||| | ||||| |||| ||||| | ||||| ||

Sbjct 2001 GATCTGACATAACATAGATGCACAGTAAACATTTTCGATGCTCTA-CTTTTATATT--- 2056

Query 2095 TTTCATACTAAAGAATGGTGTAGACATGTTTTCGAACTGTTAGGTACCCAGTTATCAATT 2154

||| |||| | || ||| ||||| | |||||

Sbjct 2057 -----AATTATATAGATGCTTTTCCAGCAATTAGGTA-----TCATCAATT 2098

Query 2155 TTATCAATGTTTGTAGAGGAGGAAATTAAttttttGGTAGAAATTGTTCAAGAAATCCTTA 2214

|||| ||||| |||| |||| ||| ||||| ||||| |||||

Sbjct 2099 TTATGAATGTTTGTAGGAAGAAAACCATTTCTTT-CTAGAAATTAGTCAAGGAAT----- 2152

Query 2215 ATTGAATGTCATTAAATGATGGTGGC-----CAAAATAAAACCTA-----TTTAGAA 2261

||||||| |||| | ||||| ||||| |||||

Sbjct 2153 ----AATTTATTGAACACTGGTGACTACATCAAAATAAAACCTATAGTTTGCTTAAAAA 2208

Query 2262 ATTTAATCACTTTCACATCACTTGGGAATATGATGCCTCTAGTAGTTACTtttttatagt 2321

||||||| ||||| ||| ||||| ||||| ||||| |||

Sbjct 2209 ATTTAATCA--TTACCCATCATCCTGAGT-CGATGCTTCTAGCAGTTAC---AGTAAACA 2262

|||| | |||| | |||| | |||| | |||| | |||| | |||| | |||| | |||| | |||| |

||||| |||| |

Range 2: 4006 to 5767

Query 4259 TAATACTGCTGGACTAAGATTTTGGTAGCATTGTTTT-----CTAAAATATTT-TAAATG 4312

||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| |||||

Query 4313 GAGAATGAACACTTATAAAATGCTTTGGAACATAATCTTTAGCTTAATTTTC---TGTTA 4369

| | | | | | | | | | | | | | | | | | | | | |

Query 4370 AAATTTAGTACCCCTTCATCATTCCAATAAAGATAAGACTGATCCATTGTCTAAGGAAAT 4429

| | | | | | | | | |    |    | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |

Query 4430 TATTTATAAATAATAGAGATTAATTTATTTGAGATTTGAAATAAGAATAGTATGAAAATA 4489

|||||

Query 4490 TTAGATACCACATAAATTGTTTGAAATTACT-GAATAACCATCTTAAGTATGGA-ACATT 4547

||||| | ||||||| ||||||| ||||||| ||||||| | ||| |||||||

Query 4548 TAAATGGCTATATTT-TATTTGTGTA-CAGTTTTTCTGTGCCT--TGTTAGGCCAGT--- 4600

|||||

Query 4601 ---GAAGCAATTATTTCTCTAAGAAA--ATGACAA---TAAAATATAACACACTTCAGA 4652

| | | | | | | | | | | | | | | | | | | | | |



Sbjct 4355 AATGAAGTAATTCTTTTTCTTAAACCTATTTAAAGAGTAAATATACTGACTTCAGA 4414

Query 4653 TTGTCTGATTACAGTTTGAAAGGACACCGCAATGTCAAATAGGTAGGAGAC---CAT 4709

| | | | | | | | | | | | | | | | | | | | | |

Sbjct 4415 TCTTATAATTTGCACTTAGGAAATGGTGGTGAAGTGTAAGGATGTGGGAGGTTGACAT 4474

Query 4710 CAAAAACACAATTAAAGTAACATATTAGGAGACTTGAACTTCAGCCTAATAAATCCTTC 4769

| | | | | | | | | | | | | | | | | | | | | |

Sbjct 4475 CCAAA-CACCGTCTAAGTAACAT--TAGGACAATTGGAACCTGGCCCTAATGGG-CCTTG 4530

Query 4770 ATGTTCTTAGCCTTATTATTGTGATATAATTCTAGATATTTCTTGAGGGCATGTGCC 4829

| | | | | | | | | | | | | | | | | | | | | |

Sbjct 4531 -TGTTTTTAGCCT----ACTATGGAATAATTTGGCTATTATGTTGGCAGG--TGTGCC 4583

Query 4830 CAACTCTCCCGACCCCATTTGTTTGTCTTTAAAGTTCTTAGAATAAACAGTTCTTTA 4889

| | | | | | | | | | | | | | | | | | | | | |

Sbjct 4584 GACTTTTCC-----ATTCTGC----CTTTAAAGTGCTTA-AATCAACTGGTTGTTA 4630

Query 4890 TATAATAATTATATTTTATTTAAGAAAATAGTTTGTAGGTAC-TTTTAAAGATGTAA 4948

| | | | | | | | | | | | | | | | | | | | | |

Sbjct 4631 ---ACACATTGCGTCTTATTTTGAAGATAGTTTCATTCATATGTTTTAAATAGGAGT 4687

Query 4949 AT-TTTTAAATTACAAATACATATGGGTCTT---TGATAAGCAATAGGAATTGAATTAC 5004

| | | | | | | | | | | | | | | | | | | | | |

Sbjct 4688 ACCTTTTACATTTTCAAATACATTTTGTCTCTTTGATAAGTAATAAGAGCTAAACATT 4747

Query 5005 AAGTTACTAGGGTTATAAGCAAAAGGTTGCTTACCATAATGTCATTAGGTCACGATTTTT 5064

| | | | | | | | | | | | | | | | | | | | | |

Sbjct 4748 TAGCCATT-GGTTTATAATCAAATGGTAA---ACCGTGATGCC--TGAGACACTAAAT- 4800

Query 5065 AGCTCACATCTGGAAGCAGCAACTACTTGGCTCAA-GTACATATAAGAGTAATTAGTTTT 5123

| | | | | | | | | | | | | | | | | | | | | |

Sbjct 4801 -GC---CAT-----AGCTGCTTGGCTCAAAGTGTGTCTTAGAGTGGGCAGCCTT 4845

Query 5124 ATTCTCTCTTT-TTTATAAAATCGGGTTTCAGATGAGATGTTTATCTTAGACTATTTAG 5182

| | | | | | | | | | | | | | | | | | | | | |

Sbjct 4846 ATTC-CTCTGACTTTGTGAACTAGATTTTAAATGAAACATTTATCTTTAACT-TTCTAA 4903

Query 5183 GGAAAAATTTTACATGTTTGAGATGGTGGAGTAAAAAGACTGTTAAACATTTCTTTTAA 5242

| | | | | | | | | | | | | | | | | | | | | |



Sbjct 5467 AA-GGTAATATTACTACCATGTAGATGATGACAGTTCAAATTGTCCCACGTACCCAGAA 5525

Query 5831 TTTTAGAAACTAGAAAGTCTGGGAGGTACTATATCAGCTGTAGTTGGGTAATTC-CAAGTG 5889

|||||

Sbjct 5526 GTTTAGAAACTAGAAAGTCTGGGAGACACCACATCAGTTGTA-TTGGGTGATTCTTAAGTG 5584

Query 5890 CTGATAGTACTATTCATCTTTTTTAT-TATTGTGTCAGATGAAACAAATGC----CAAGT 5944

|||

Sbjct 5585 CTGTACACGCTATTTCATCTTGCAATGTCCAGCACAGAAGGAGC-GATGCCAGGCAAGC 5643

Query 5945 TGCAAAATATGCAGATTTTTATTA--TATAATGGTTTTAGGCATAAATTATTAACAAGCC 6002

|||||

Sbjct 5644 TGC-AAAGATGCAGAATTCATGACGCAGGCTGGTGTGACACAT-AAGTATCAGCCAGTC 5701

Query 6003 ATGCCTTATGTGTTTCATCTTATTTTTCTTTAGAACTAACTATAACAGATTTTGAA 6062

|||

Sbjct 5702 ATGTC-CGTGTGTTTTATCTTAGAGTTTCACTTAGAGCTTCACCCCTTA-AGA-GTTGGAA 5758

Query 6063 AATGATTTG 6071

|||||

Sbjct 5759 AATGATTTG 5767

Range 3: 5858 to 6074

Score:207 bits(229), Expect:4e-55,

Identities:183/223(82%), Gaps:8/223(3%), Strand: Plus/Plus

Query 6214 CCTCCATCATGACACACTTACTACATTTATGAATTGAGCAGTTCTGTAATTGTAATTATT 6273

|||||

Sbjct 5858 CCTCCATTATGTTACATCTGTTAAATTCATTATTTGTACAGTTCTG-AATT---ATTATG 5913

Query 6274 ATTGCTGTTTCATGTAACAAAACATGCTTATAATAGCAAACA-AATAGAAATGCCCCAAA 6332

|||||

Sbjct 5914 ATTGCTGTTTCATGTAGGGAAACATGGTTATGATAGGAACAAGAATAGAAATGTCCCAA- 5972

Query 6333 ATGCTAttttttAATTCAGTTATAACTGTTACTCTTGTAGTTGTGTATGACGCAATAAA 6392

|||||

Sbjct 5973 -TGCTATTTTTTAAATTCAGTTGTAAGTGTACTCTTATACTTGTGTATGACAAAATAAA 6031

Query 6393 ATTTGTaaaaaa-aTTTCAGCATGAAAAATAAAATTTGTATCA 6434

|||||

Sbjct 6032 ATTTGTAAAAACACTTTAGCATGAAAAATAAAATTTGTATCA 6074

#### Annex4. Alineament resultant CLUSTALO A-B.

CLUSTAL O(1.2.4) multiple sequence alignment

genomicA	cagaagaattgcttgaaccagggaggtggaggttgacagagatcacgccactgc	60
genomicB	-----gctgggatg--tggggagcagtgctctgaggctgagcag-gac	40
	* ** * ** *	
genomicA	actcctgcttaagtgaacagagtgagactccatctcaaaaaaaaaaaaaattcctatta	120
genomicB	agtgagccttggcctggcct-----ctgaaccatttttccacctaggcctc	90
	* * ** * * * ** *	
genomicA	tgtgcttgagtaataaccaccactctggcaaatcttaaaaaagctcttggccgggtgcag	180
genomicB	tgagcctgtgtcctataactattgagcgtgttagaagc-----aggcagac	138
	** ** ** ** ** ** ** * * * * *	
genomicA	tggtcatgacctgtaatccccagaagaattgcttgaaccagggaggtggaggttgacag	240
genomicB	tacttctggtatgcttggctgcttgaattttttctgcca-----	179
	* * ** ** * ***** *	
genomicA	agcagagatcacgccactgcactcctgcttaagtgaacagagtgagactccatctcaaaaa	300
genomicB	---ga-----tatcctaggtcatcactctATGAGTGTGGATCCAGCTTGT---	221
	** * ** * * ***** *	
genomicA	aaaaaaaaaattcctattatgtgcttgagtaataaccaccactctggcaaatcttaaaa	360
genomicB	-----CCCCAAAGCTTGCTTGCTTGAA	245
	* ** * ** ** *** **	
genomicA	aagctcttggccgggtgcagtggtcatgacctgtaatcccATGGGAAAGTCTCTTTCTCA	420
genomicB	GCATCat--gggagagctgtctctaaagatctctaaagtgactttgagccttttgcctca	303
	* * ** ** ** * * * * *	
genomicA	TTTGCTTTGCATTCAAGCAAGAAGATGCagttccccatttctgtcgccacacctctga	480
genomicB	ttgtcttgatattagccctt-----ggcacccttttagtcacgctaataccacctta	354
	** * * ** * * * * *	
genomicA	gatggtgacctgtgtctgtcattgtttcttgaatcaatctagacctcagttctaaagaacc	540
genomicB	gcaagtgttgctccacagcctgttat-----atcctctctc--aataatgc	401
	* *** ** * ***** *	
genomicA	ctaaaaactctgtccgtgaatcttgggggaaggaggaagtcaatgtaaaatacttccata	600
genomicB	tttttattcttgccacatgg--ctggctacgagtttccaaacttgta---tgcttt	455
	* * *** * * * * * *	
genomicA	ttgtatttctaagatgtctatttcccttt--gtgattattttgactgcaagtgtccgtg	658
genomicB	gtttcccttttaaatgtaagtttcagcttttaagtcatttctttgcatggggagcagatga	515
	* * * * * * * * * * *	
genomicA	aatcttgggggaaggaggaagtcaatgtaaaatacttccatattgtatttctaagatgtc	718
genomicB	atcatatggtgagagaggaagtcacagagagactaggatgtggtaccagactcttaag	575
	* * ** ** * ***** *	
genomicA	tatttcccccttgtgattattttgactgcaagATGAGTGTGGATCCAGCTTGTCCTCCAAA	778
genomicB	caatcaaactctcagtgaaactaactgagcaagaa-----gtgacttatcaccaag	625
	* * * * * ***** *	
genomicA	GCTTGCTTGCTTTGAAGCATCagttccccattt-----ctgtcgccacacctc	827
genomicB	gggtgttaaccattcatgaggatctgcccacatgatccaatcacctcccaccaggaaat	685
	* ** * * * * * * * *	
genomicA	tgagatggtgacctgtgtctgtcattgtttcttgaatcaatctagacctcagttctaaaga	887
genomicB	cacattggtttccaaATGGGAAAGTCTCTTTCTCATTTG-----CCTTTGCATTCA	736
	*** * * * * * * * *	
genomicA	accctaaaaactcagttccccatttctgtcgccacacctctgagatggtgcctgtgtctg	947
genomicB	AGCAAAGAAGATGCTgcccacatgatccaatcacctcccaccaggaaatcacattgggaa	796
	* * * * * * * * * *	
genomicA	tcattgtttcttgaatcaatctagacctcagttctaaagaaccctaaaaactc	1000
genomicB	tcac-----	800
	***	

## Annex5. Megablast A-B.

RID: B63R11NS11N

Job Title:genomicA\_vs\_B

Program: BLASTN

Subject:genomicB ID: lcl|Query\_63835(dna) Length: 800

Query #1: genomicA Query ID: lcl|Query\_63833 Length: 1000

Sequences producing significant alignments:

	Max	Total Query	E	Per.
Description	Score	Score	cover	Value Ident Accession
genomicB	95.3	188	10%	2e-23 100.00 Query_63835

Alignments:

>genomicB

Sequence ID: Query\_63835 Length: 800

Range 1: 201 to 251

Score:95.3 bits(51), Expect:2e-23,

Identities:51/51(100%), Gaps:0/51(0%), Strand: Plus/Plus

Query 751 ATGAGTGTGGATCCAGCTTGTCCTTGAAGCATCA 801

|||||

Sbjct 201 ATGAGTGTGGATCCAGCTTGTCCTTGAAGCATCA 251

Range 2: 701 to 750

Score:93.5 bits(50), Expect:6e-23,

Identities:50/50(100%), Gaps:0/50(0%), Strand: Plus/Plus

Query 401 ATGGGAAAGTCTTTCTCATTGCCTTGCATTCAAGCAAAGAAGATGC 450

|||||

Sbjct 701 ATGGGAAAGTCTTTCTCATTGCCTTGCATTCAAGCAAAGAAGATGC 750

## Annex6. Blastx.

RID: B65Z619G114

Job Title:genomicC

Program: BLASTX

Subject:NP\_001127704 length=274 ID: lcl|Query\_64215(amino acid) Length: 274

Query #1: genomicC Query ID: lcl|Query\_64213 Length: 21894

Sequences producing significant alignments:

Description	Max Score	Total Query Score	E Value	Per. cover
NP_001127704 length=274	135	606	5%	9e-39

85.14 Query\_64215

Alignments:

>NP\_001127704 length=274

Sequence ID: Query\_64215 Length: 274

Range 1: 1 to 73

Score:135 bits(339), Expect:9e-39,

Method:Compositional matrix adjust.,

Identities:63/74(85%), Positives:67/74(90%), Gaps:1/74(1%)

Query 8096 MGKSLSHLPIHTCKEDGYDGGTVSDNMRNGLVHSESHGEDGRCGDVVSQFPYVEFTGRDSV 8275

MGKSLSHLP+H+ KED YDG T S+NMRNGLV+SE H EDGR GDVSQFPYVEFTGRDSV

Sbjct 1 MGKSLSHLPLHSSKEDAYDGV-TSENMRNGLVNSEVHNEDGRNGDVVSQFPYVEFTGRDSV 59

Query 8276 TCPTCQGTGRIPRG 8317

TCPTCQGTGRIPRG

Sbjct 60 TCPTCQGTGRIPRG 73

Range 2: 226 to 274

Score:86.7 bits(213), Expect:4e-22,

Method:Compositional matrix adjust.,

Identities:40/49(82%), Positives:46/49(93%), Gaps:0/49(0%)

Query 16460 VFRLVTVTTSYFGHSEQISREKYQYVDCGGNTTYQLGQSEYLNVLQPPQ 16606

+ ++VTVTT+YFGHSEQIS+E+YQYVDCG NTTYQLGQSEYLNVLQP Q

Sbjct 226 LMMQVTVTTTTYFGHSEQISQERYQYVDCGRNTTYQLGQSEYLNVLQPQQ 274

Range 3: 154 to 198

Score:76.3 bits(186), Expect:1e-20,

Method:Compositional matrix adjust.,

Identities:40/45(89%), Positives:42/45(93%), Gaps:0/45(0%)

NNNYYSVEV NITAQVQFSKTVIGKARLNNIT IGPLDMKQ++ T

Sbjct 154 NNNYYSVEVENITAQVQFSKTVIGKARLNNITIIGPLDMKQIDYT 198

**Annex7. Regions promotores.**

```

>hg38_refGene_NM_018374 range=chr7:12206294-12211293 5'pad=0 3'pad=0 strand=+ repeatMasking=none
tttttttatttgcgtgttccctaggttttagtggttcttttaaatttggtttt
tgttttgtttttgttttttaggattcactgaatttcttgaatatatggattt
acatattcagtcagttttggagatttctaagtttttatcttttcaatatt
tcttctatcccccttctgctccttggattccaagtacatatgtgctaaat
gatttgatattatccataaagatatcagatgtcccttccatttgccttg
attttgcctttttatattcagtttatataatttctactgaccgtctcaagt
ttacaggttttttccctctgttatgtccaatttgggtgataatgctaaaa
agtgaactttttacttctaataatttatttctaataatttcaacttgctttt
tataaatagttcccatcttttgcataaaattccctgtctgttcatgcatg
ttttccattttttatacacagatgctttaacatatctatcaaagtgatttta
aagtctctgccagataataccaaatctgggccatttctgtattttccct
attaatcattttctctcttgacctgggtcacatttcttgccttcat
gtctcataatctttattatagctagatattgtataaaaagaaaacagtaa
ataacatgtacctcaaaaaacatcttgccatttttctatcaagatttatt
gtttggaaggctgaatcaaatctttcaacagtgagtcaaaaacccacatt
ctgggttttggtgaagtttgattcagtttatttactattgctttcaaat
atgcttgagttcttggcactggcaaaatctcaagattaatgtgcacctac
tagctcattccacaaaactattcaataactttcttagctcaaacacctagc
ttccaatatctccacaggttggttctgtttactctccagctctgcccgtca
tttctgtgcctcaggagaatcctgggcagcctgccaccctgcctccagcc
ttcacagagctgtcgcagtacactgggttaaaggcctagaatctgcaaaga
aatttcttcagcatttcatgccctacccccatgcctttaacatatgacattt
gagcacttaatgaagacctttaggagagaaattggagagtgagtgacact
cgttctatcactggaacgacatgagattgtaatctatcatcctgggccac
acacaggggtgtcattcggcaggtctcccttcggccctgtttatgggtgca
ttcctccttctcccaactgtttttaggtgattaaagcaactctaatttcta
ttctctcttaggaagctctccttatcttttggatttaagtgtacttaaat
ttgcgtcctcatttctttgataagttttttaaaaaatgtttaaaaaaaaa
aaaccttctctagcttggtttgttggagaagaacagctttttataacat
aaatcctaattggactctgtgttttccctattttatattcggaaacacctc
ctcaaggatttctggaccaatctcctcttttcagcagcagtgatgatc
acttagcaaaaagagtgaataaccctcttttccccccctaaccagttaac
ttactcacctcatcacttatctttgtgtagtaaaagttgaggacagaa
ttcactcttccaaatttaatacacagatattcatgccacagtggttaatagc
tgtagctttgcagtcaaaaatatccgagttctagttgctagctgtgtggtc
atgtgcagatcataataacctaggatcctctatttccactgtttgaaataa
gatatgggataataacagaacctaatccatagcatttctataaaaaattag
atgatataaactactctttacacaatgcctggcacacaaattcattcattg
gccagtggttattgaaactcctaaaaattattcttactagaaattatgatta
tgacaaaaaacctacaaaagatgacttgaaatttgggaaagaaagtttt
aacacatgggaaagattagctccatgcagggtgtgaagcaagatgaatgag
acagagaaatatagaagatcctgcctaagaaggagtaacagagacataac
caagagtttcagataggttaagttatttgcagaaattaataaagacctctg
cttttaaaataattggaatgacaaagatgtgtcaatagagtcaggagttt
actgcctgtgaatgtaagtcagggtggcacagcgggaaatggtagttttgc
taatggtagtttttgctatgtatctggggcactatctctgaaaatcagcca
gacaaacttcaatcttttcaatccctggtttccaatgtaacctatccagg
tttctaacataatctgaccagcatatcagtggaaggggtgagctgagag
agacttatataattaacatttgacaaattatattttcaaaataaattatgt
tcctttctgcttataaatttttaacagctaatttctagaaaagtaatttt
tgcaaaatttttagttatggctgactgctattaggatttgggtgatttac
cattaaaaagagataaaagagaattgaaatgagagataaggctgttttcaa
gatttagtacaacctagagaaatcccttagcttcagcatcataactaaata
taaccagaaatcattaagaaagaatgttcttctattcaacttatcccag
actgttaaaattactggcctttatctgatagcacctagaccactactg
aactcggaaactcaagaaatcactgttaatcaaataattggaagcttttca
gaaaattttcctttcaaaatgtgcttcaatgaaaaataaggcataggaac
aaatgtaccaaaacattcaatatctgtctttaaagtacttcttgttatat
ggatgataaaattgtaatcataatatctaatatttactgattgctctgca
ggatcagctcccatccataaccctttgagacagacattattactatctt
cattttgcagataataaaaaattaggataaagaaggctaagtattttgcc
cactcttacagagattttgtatatctctgtcaggtgggctcattttcag
gccatatggctcccaaggctgtgttcccaaccctgctactggaactgcttc
ttacagcaacattctcccaggaagacttatttccatgtgaattccaggaa
tttggataaggaactgtgctatatggccagcacgttttaagttaaatgat
gaggttgagaaaaactaaataactcccttaagaattaaagaaaaaagtaa
ctttttcacttaagaataatatattacattatagaaccatttgcctat
ttaacaataataaagaaaaaacatccatgctctcaactacaggcaactg
tttatattttgtatatttctttccatttcaaggcagatatgtaatatatt

```



tatatctatatgttatctatatatgtgaatatagtttacatcctattacaac  
acctcatatcctataactttgcatcctgattttaaattttatatcactcgcg  
gtataacacgctttataagattggttttactggtctttataaaaactcaata  
gatggctattccataaattttaaaactcagtaaatccaaggggaacatca  
tatctccacaggtggactgtcaataaattttgagaaaaattttcccccatg  
attctgatttgtttttaaccaactattaattttgaaatttcaaacagtg  
taatggagacacccctgaccccttccctttggaaaaattgggctaaaaaaa  
agcggttgaaaaataatgattttatacagtctccctgttgacattcagtg  
ttatttaaatggatacacctcatacccaaatcttttatgagttctcatta  
ggtttaggatacttctgaaagggtggaattgctggaccaaacgatgtgaa  
cattttaatagcttctgatacatactgactgtctgccagaaaacttaggag  
ttttaaatctccgacaacatttgtatgagaaagctcatattctcacatagt  
gatctacagtaaatctcttaaattgtatactgtctgactgttatttctg  
acttttacatagcttctatttcccaattcctgttagaattcctgattcta  
acaggtagcacactgactctgttagaagagagaatatatgactctcactt  
cctgtcctctgactgttagaagtgcctaccatttataaacgttaaac  
ctccgcaggaatgagactagagtgccttaatatattatccggcatctca  
ttgatccctcaaaacgactcctgtgctgtgttcttattcttcaagtga  
tgacttagacggttgggggagggcagcccaaccacagcgaaagagcttaa  
gtactgggatccgagacgggattcacacacctgtaactagcaccagagt  
taaggggtggggaggggcagcgtgggcaagcgaaacgaaacccgacaa  
aacaaaaaactacgccccttgcgccttgcgtctcctctagtgtggcaagt  
tctaaattaaaatagaggtagcttctgctgctgcagcggcgtgtgtacgg  
accagtttattccattcctaaggacacagactgcagcgacaactcgatt  
gcgtgagcgaggaggtctcagctgaaagctgaggcagatggtgcacat  
ttataccctttgcaaaaggaaattggtctgatatggaaattgaggaggagc  
actgttcttaggtaattcgtggctacagtggacactttgaagtcgtgggg  
acgcgaaggtaatgaaacctgaagagcgcacaaatctcacacgatcgacgc  
ctctgggctcaaacaggaggaagaaggacgcatgcgtcacgcgcacgcc  
gggtgtctcctccgggagcgtctgcgcaggcgggacgcgcaggttacagc  
agcgtctggcctctgctgtagcgcctgattctacctacccttcccctgc

[illegible]

aggaggggacaagcagccctttttatagagttaggcacacctggcaacag  
gtactgcggggagagcctgacaaaatgccaacactaacagctgctata  
cttacaaccttttaaaacaagctccaacaggggccataccttctaatcgtg  
ccactccttgggagagcatatacacaccatcacattctgcttctcctggc  
cccataggttgtccaaacatatgaatctatggggggccatacctaaacat  
agcataataaaaccatcacagtgatgagcatgtaactgagggactatag  
aaaaataatatagaagacaggtgtaattatggaggtggggaaaagacc  
cttactctggaggctagttggagctgtctctcagttgagagttaaaggatt  
ccagcataaataacacttactgtgatgttgcagacatagaacaatactc  
attttaaaactgtaaaaataacaacaaaaacacccaatagaaaga  
gtgcataagacatcatgtatgttctcagtccttgatattttaagaagcct  
tcacaaacgtgtcttttttttttttttttttttttttttttttttttgg  
aaacttttatcttaactttctgcacccctaggtcgctttacattgtgagt  
tgtaatcttattcaagactccacaggagctctaattccatttaccttct  
taatgaacctttaaagattactgaaaacccaccccaaatgtatgtgtg  
ctactataaagacaatgaagtgtgattctctcttgaggagcccacagcca  
tgttctcagatgagatttatacattctctgagtcctttcatttctattt  
aagtctatgcttaaaactctatgtttaaatgccttttaatacacatcaagt  
gggcttcatattggatcttattgaagggaaggaaggctgttctcacagca  
agtttttaggggcaactagtggtgcaacccaaggaaccacatctcaaaagcaa  
tcattttatcaagtgcactgctggtgtgacttaatacaggtatttagttta  
tgtgtatttagtttcttgcagtggtttgaataggaacagcctctggacc  
ctgagtggtgaaggattggcccaaaggagtagcactattagggaggggtgt  
ggcttaaatggaggaggtgctgactaggagagggccttgggggtctaagg  
agctaattcacttccctgctgtctctgagtggaactctaagctcctcca  
gtatcatgtctgcctggttgcctgcatgttctcctgccaggagcataacgg  
gctaaacctctgaaactgtaagccaaccccaataaatgtattctttata  
agagttgccatggtctacttgcgaagctattgttaactctattgaaatcct  
aagaaactggttacacactgcacacaactttactttagaattctcattat  
aatttagatgaatctcacagagtgaaattgccggacctaaaggaggtgaa  
tgttttaaatagtcttaaaacatcctgcttatgtagatggaaaatacaca  
gctctcacctcaaaaggacaaagagtttattctaaagccaatataggaa  
ccatgacctgaaaaacaaagacttgggctattccaaataccaatgtcccaa  
tgtgataacaatttcatgaaacttttatatacatagaagatggtcataaaa  
tcaaaatactttccaaatacacatacaaggcagtttagaagagcggggacaaa  
attgggagctcctcgccagtgcaagttgtcttttccagtagagaaggaa  
aaaaaaaaaaaaaaaaaagaaaaagaaaaaaaaaaaaaaaaacacaattaag  
agatgcttacttcccagttataactctgttaattttctaggaatttcacc  
agaaaccccatttatagggagctattaaactccttttagcaaccgcatct  
gttcttgcctttcaattttatcttattataatcccatccaggaatcttatg  
agccacttaatacatcttgattatcaggaagacagaaaagaggtcataaac  
aacctttgactataaacactggcgagccactaaaatgagccactgtagcct  
caactcactgtaaatatgactggacaatgaggtgcatggattgggacac  
tcctttaacaactcacttggctcatgagctttggggtacaactcttattgc  
atgcaactgaataatgtgaatttaagtacaaaatcaaaacaagattaaacc  
aaggcgggggcggggggtgcgtcttcccgtgacatttaattctaaatgaa  
aactcaataacaaaaattctggcctgcaactctgccagaaatgtgcaccaa  
ccggctcattcttaacccaggaaggaagttttgtgtctgcaaatctagac  
ttgagaacacgatcgctcaaaattgcgacgggtgcacattcctccccttta  
caaaaggaatcagcctggtggaggggaaaagggaacacgattattcgt  
agccaggcgcaattagtagtattttgaaactttccagctctagagaagaga  
caggtgacaaaaccagaaagctttcacctggcaacttccttcaaacgg  
tgacagtgcgcatgcgtagtgtccattccaggggcctagcccctccttct

>rn6\_refGene\_NM\_001004267 range=chr4:39512679-39517678 5'pad=0 3'pad=0 strand=+ repeatMasking=none  
gccaggaacaccagaaggcagagttgagcaggcaggggcagagagctga  
ggcaaaagacaccagaatgaaagcaagaaggaactgggtgaaagtatgt  
ctgcaccagctcccaggcagggcagggccactggttttgttcccttccctc  
ccctgcctgctgcctccctgctgcgccagatggccagatacagcagtggt  
gacagcagcagcagcagaagttaaaattataactagaaaaatggcagcca  
ggccagatcaccaggtggtcacatggtgaacttggggagttagctctggg  
ccgcgactgtaggtaaacagttcacctatgcctaggatttggcaactaca  
agaaggaaactacctcagaggagagaggaatactcgcccaagagaaaaatg  
tcctagaatgagggtcagtgaagagaccaagtcataaagaccgatatctc  
atcctcaagggcagcacaggggttgcctctgacaagtaggaggaggggagg  
ttctctgaagtttgggctaaagagcaaaaggcctctctatgtggtggtggg  
gggtgggggggaggtgggatgtgaacctgaggagtttctaattccaagtat  
cggctcaaatgtgtttataaaaaatatagagagaaggaatatgtttgct  
ggaagttaaggtaagggtgtggaggaatggggtgcctttgcaggccattg  
ctggcgcatcgctttccacctgaggaccaacaccacaggacagttgcta  
tggttcccttagtagctcgatatatcaggcatgagcgacgtgggttggtg  
ggaggtgaggtggtggagtgttgaggggtgggttcgagggggtattttagt  
ctagggttttggtgtcagggtttaggactatttctgaggagagagtagggc  
gagtggtgtgatctctgtgagagagatgaggtagcttcttcttctgtagg  
tgaaggagtgggtggtgactgtttaccttggggggtcatcagtttgtga  
catggtgttgcctgctgattttctgtttaattttaactatctgtgctgct  
gtctgtgacctgctgtatctgaggaagtgcatggagggggagagggagg  
ggcgaggaggggtgggaggacatggaagatccggagccatagatgataga  
tgggtggtgacacagatgaactatgagggggtacaaggattgatattact  
gatgggtttacatttgggtgtttttgtagtgttgcgagagagggaaagggt  
aagaggacagcgctagagtaagaagcaagaggtgggtggggcaagcagc

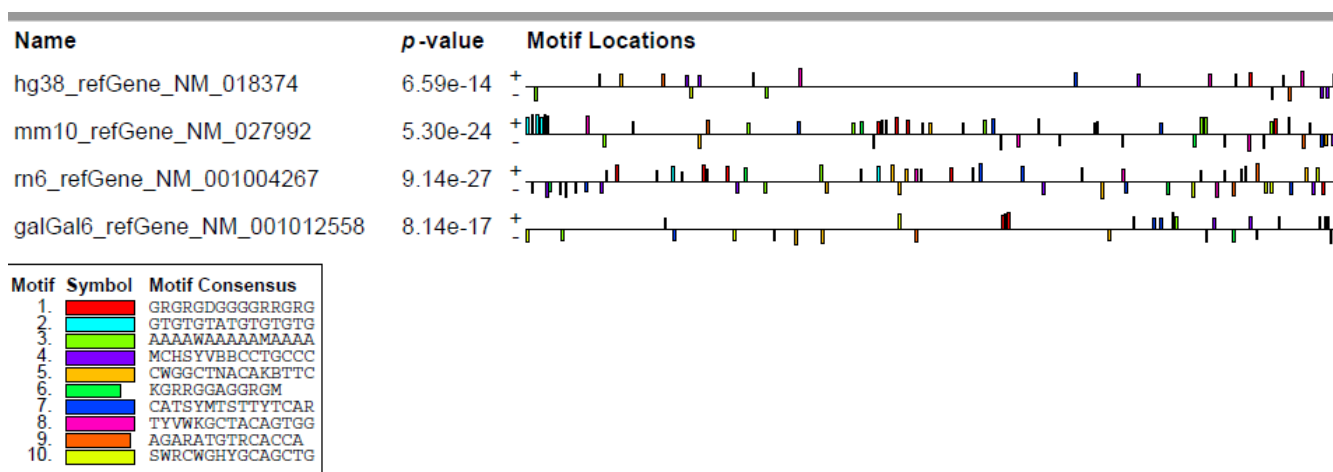
ctgggttatagtgagtcaggcacacctgactgttgccaggttaactggggg  
gtggagcttagacaaaatgccagcactaataatgctagacctacgacct  
gtaaaacaagcattatttgctatctcaattctaaaagcaatagacatatgt  
ataaagaatgaggtttttctttatttccaggctgtttacctactttgtag  
cttagttcatcctcaggcaagcagccttctaaggcaataaccactaaattt  
acaggacttaccttctagagattcattattgaattctagtccagtggtata  
aagactgccccaaatgacaatattgatagagacaattctgtgaaaaacac  
ccttagatattccttttaaggaaggacttattttgcattcaagaataatc  
aggtcagcctatagaaataggcacaaacttcaacattgtatacaataaaat  
atttatttttaaaagtctctcaaaaaggcaactgttggtgggatttctgtccct  
tgtccaaaaaataaaaaaaatagtctaattaaagagattggaattgt  
tacctgaccccatccagcttccctcttagctcggggtcattataacctt  
catgaaaacagtttagatgacaccgagggtcaagtttattgtgaagaat  
gtgaggcattccacctactgagcatggacaaagggtcaagagaccactgca  
gaggaaagctggctaagggttaggggtgatgatgtccagagacctgtgc  
taactgacctacatgttcaggaaatctttgtctgaaacaacttcttgaaa  
tggtatttccctatccaggtgtgtctcaggtccattttgagtaactctta  
aggatgtgtgtgtgcacgtgggtgtcttaggggttactgtgtgaacaga  
catcatgaccaggacaactcttacaaggacaacgtttaattgggactggc  
ttacaggttcagaggttcaggtccattatcacccaggtagaacatggcagc  
ctcaaggcaggcctggcacaggaggaactgagagccttacatcttcatct  
tcagggtgctagcagaatactggcttcagcttaggatgagggtcttatagc  
cacagtgcacacactactccaacagggtccacacttttctaatcatgccatt  
ccctgggcaaaagcatatacaaaaccattaatcctacttctgtgtcccat  
ggctgtgtccaaacatgtgagtaaacacagcataataaagccatcacagt  
gtgtgagggtgttattgaggtcactgtacaaaaataacagaaaggcaggt  
gaagttaagggtcgggggtgggtgtctccttattctggaatctagctga  
aatctttcaccagcttagaggttaagggttcagcacaaatataattcact  
gtgatgttgcaaatagaaacattatccatttaaaagctgtaaaaataaca  
acaaaaacaaaataactcaaaagaaggagtgcataagacttcatgcatgtt  
ctcagtccttgataattttacaaagtctttaaagaacaatacttttctttta  
ggtaaagatatcttaattctctgcacccctaggttgctttatatgtcag  
ttgtaatcctactcaagactccacaggagctacccgccagccaatcccat  
ttaaattccttaatgaggtcttaaaagactactgaaaaccaccccaaaag  
gtatgtgtcgctacagtaaaagacaatgactatttctcagaagtcacag  
ccatgttctcagatgagatttatacattctttgagcccccttctctttca  
taaaagtctgtgttgaaatataatgtttaaatgccctttaatacacttca  
agtaggcttcataattggatcttattgaagggtcaggaagactgttctcgca  
gcaagttttagggtcagtttagtgcaacacaaaggaccacatctcaaaagct  
attaattcaccagtgactactgtgtgactaatcaggtattatagttta  
tgtgactgggtcacagtggtttgaataggaatagcctcctcaacctgtttg  
aaggtttggctcatagggagtggtgcacagttaggaggcgtggcttagaat  
gggtgtggctttgttgggggaggtgtgtgactaggaagtaagctttgaggt  
gtaaaggagcaaatccacttctgtgtgtatgtgaactctcaagcagaacca  
gcaccatctgctgcaggaagataatggactaaacctttgaaactgtaag  
ccaggcccaaaaggttgccgtgtgtgtgtccacttcgcagcaatagaaa  
tcctaagatatgtgtcacccactgcacgcaacttactttggaattctca  
ttatcattcaggtgaattctacagtggaacctgaagagatgaatgtttta  
gtcttgaaacttctctgttagatggaaaaacacagctctcaccttaaa  
agggtctaagagtttattttggagccaaatgtgagggtacatgacctgaaa  
acaaaaactcgggtctattctgaataacctgtcccaactgttaacaatttc  
atgaaacttatacacataaaaagaagatgggtcataaattaaagatactttcc  
aaataaaaaacaggcagtttagaagagcggggcaaaatgggtcgctcctc  
gccagtcgggtgtcttttggcagcagagaagaaagaacaacacaatca  
atggatgtctaaacctttccacttagttatagtcatttaattttctaggt  
aatttcaccagtaacccatttatagggaaattttaaaacttttagcagct  
gcatcagttcttgggtttcaattttatcagattataattccatccagtcct  
cttttgagccacttaatatatcttgattatcaggaagacaggaaaaaaa  
aagtaaccttttaagataacactcgagattcactaaaaatgagccacagta  
gcctcaactcacagtgagttatgactggacaatgaggtgtgtgggtggag  
agggtcctgaggatactcctctaagaactcacttgggtcatgagtatgggg  
tacaactcttattgcacacacagaataatgtgaacttaagtacaaaatca  
aaacaagattaaaccaatgccccgggtgggggtgcgttttccagtgacag  
aaattttctaaatgaaaacaataacaagatgtgtgtaactctgtcagagat  
gtgcaccaactggttcacaccttaccacagggtgaaggtctgtgcggtgca  
aatgcagcattgaaaacaaggccgtcaaaactgcgacagtacacatttc  
tccccctcacaaaagaaatcagccttatggaagggtgaaaggaaaccgat  
ttttcgtaaggcagacgcaataaagtactatttctaaacttgcagcttag  
agaaacagacaggaggtgaaaccaggagccttccccagcaacttcccttc  
agggtgcaaaagctgggaaagggtgcgtactacctattccagggtc  
ccagctcctccttccatgggtgactacgcagggtgcagcgcaggtcctcc  
gggttaccgggtcagctgatgcagttgtgtcctaccctccccgcgcg  
ctctccgggtcactcgggtccggccttagccccgcctgggacggga  
ctccgtgacccgtcctgcgggtgcgcagcaggtatcccggtgacg

>galGal6\_refGene\_NM\_001012558 range=chr2:26565971-26570970 5'pad=0 3'pad=0 strand=+ repeatMasking=none  
cagctgcaattgctgctatgtttcttaggtaaagacaagtgttttgaat  
gtgattttgaataaccgtgttctgaaagaagctattcagaagcattaaac  
tataatcactgggtcattcaccacactgatcagaagatttattattcttt  
gtctagatctgtgggtgtgtggcattttgtatctccccatcaatatcca  
gtttcttgaatttccagttgttttttttctcctccagaagaaagggtg

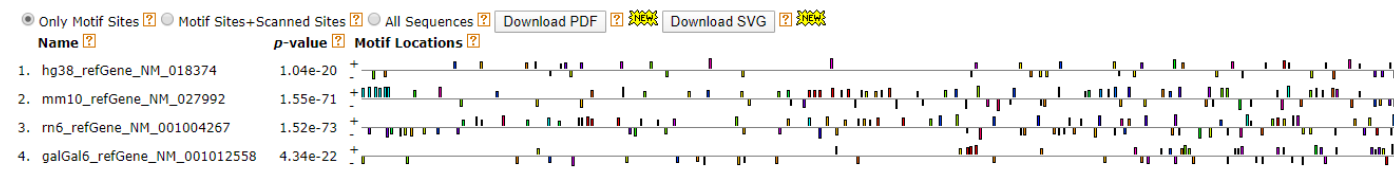
agaatatacaaacagaattaagatgggctatTTTTtcttcaactcctggga  
gcacacaatgcagcaagcaagccatggataatacatagcatcagaactgt  
ggaggactgagcacaatttctaattcaattttcatctatctacagaacat  
aatcatctcattcgtcaactaatgcaaaatttaaggtttgcacacgtaat  
gaacaaggagaaactataaagtatactgcagaagaccatagagtgttg  
caagaatgacagttgtcggtagctgtgcaatatacagaaatacatcagt  
tataaaaggctgttttatgttttattcattcaaaataactcagagcacagg  
ccgtatcatcagaaaaatacatagcaatgaaagtacctgtgacacaccatg  
cttaagcattaaatgcatctaaaaatgtttaccacagaatatcattaaaa  
tcagaaattctgaaatacagagaattcttcacataggtggaaaaggatgt  
ttcaaaaaacagaactcttaaaaaatccgaacagaccaataaagacagtgg  
aatTTTTgatattcataatttttaattcttgtaattctaatttaattagta  
caggcacagctcaaaggaacactgggtgcagtaacaggtatgttgctga  
tcagataacaggcagggtcatgagctatgttgctcctgtgcctggaataga  
ttgctcagattacatacacctacatgagcgtcgcaactattgtataga  
ttggtttgatgtgttctgggattaggatgatttctgaataaaaaaagact  
gttgcccttgtaatagatcaagggcaaaatcaataaacaattatagttgt  
tcttaatttttatcaggatagcttataacctctcttctcctgagtttctcaa  
aaagaaaaatcagagtaaaagggaatcagcattttcacttggttgctccagt  
gagcaggtgtagagaagtctaaaaaaattgttctcctgtgactgagtgatg  
aaggaagaaagtgtccacataggcagctgtaacagatgtgcaaagccta  
ctgtgccttatataaaccagtggttaataaccagctctaaagcagttacaag  
ctttcttaaaaggccttctcccctatcagcaactggtcagtgcccacattc  
tctcacatttgctgttttatattctagattagatagaaatcattgtctttta  
tgttgctttattttattttttgtgggtttggtagtttaggttctttc  
tgatttgtttgatTTTTgtcagggagaagccttgatataatgttagagat  
aaaaactgagctaaaatgctgtatctacctcagcattcagatagcatgcc  
accattttcaggctagcagactgtatgaaaacaagactcagtaactttg  
aaggtgttagccagtttcagctgagtcacagagataaagaaagtttataca  
aatttcataaagatagataaaccaggcagaagagagacaagggattgaatg  
tttgaagccattgtgtaataactttctatgcagtaagcacattttctaac  
aaattctattttgtgaacatggcagcctgcatgttactgtcagaataactct  
gcaggaagacatgagttgtatgcagttcttttgttacaatttggttaa  
acagtcattctgagaggtagatctataggaatgaggctgcattagtttca  
gtgctcaccaaaccagtagctttttatttctgaagggtgaaataaatgcag  
tcagtttgaagtttcttcggataaagcaatgcaatgaggttgaagaagaaa  
taaaattaatttagtctctcaagcatctgaggtcatgcagatatgcactgg  
gagtcctgcacattggtgcagttgattgctcttgagattgctaattccagc  
tcttgcaagaacctgactcacttgctcatatccgagccacaaaactcaact  
tgaacctgtattcatcacttggttcttcttactgatataactactggaagt  
ataaaatgcacctgggtactgatctatgagaactaagagcacttgagct  
gtcccataactcagttcaaaggatgatctcaaatccctgttctctatc  
atttgatgtttccatttcatacccctcacagtgccatttgtgctgtaca  
tcaattgcttatcatgcatttttattattttttgtatctgtgtttaagt  
atgcaatcccagtcagtggtggttagcacagtcgaatcagtttcaggaa  
ctggcaatatttgatgctgtgcagtgtaaaatgtgtaaaataactccttct  
ataaccaaaggtattttgtttcaatcagtcgaagctcagtgggccacttc  
agtgcagctttcaggcatctcatcccatatttaagcacttctaagtact  
tgtaatgtgtacatcttgttatcggtaataaccacataatttctcatgtttt  
tcatgcttgcatagtgagcgtaaagactttacaggacctcatattattgc  
ttatttctcaattcttatcacataacttttagcctataacaaataaatgata  
tttaagacaactattaaggagcttaaggcacataattggtgagcaagatt  
tttctagtgtttcaacttaaaattttctttaacatgcactttctaaatgaa  
aaacaaagcataaaacaaagaggagggaggggaaaggggaggtggggggg  
gagagtaggagagaggagaggaggagggaaggaggaaggaaggcaaat  
gaagggaagaaagcctttcctaattgtatcaaatctacagaaggaaagcct  
ggcagtttggaatgtaatttaatacaaaaacagtttattgtttaaagtaac  
agactcaaaacctcatgcagatttcaactgttggtgatctatcatggaaag  
agaacaatgtctcaggcaactatgtaacatcctgatcggcatttggcct  
ctagataagtttcaatgttggtgaaactgctgcaatcaaaagtgtaa  
aactgtaatttgaccaagactgcactgtatttcaaaactgctagggagaa  
taattttatattgatactctgtcagagtaaaataactctgagtagccag  
aattcagtgtaattcccaaagccatttcaaaaaacctttctgctagactt  
actcctcacgtctcctgtagtataatgattcagatcaatctgaatgactg  
acattattatctgatactattcaaaaggcggagactgtgttctcattttc  
tgtttggatagctggaaactctgatactgtttgatgtggatcaatataaa  
cttgagactgaaccaaaagtcaatagccagtgaaagccagtagccaggctt  
accaaaattcacctcatctgacacttggtgctgaacatgtcatgttcatac  
agggtccaaaacagccagggttctgttcaaaggtttttatgtataatttt  
gtatgtaggaaccacagagtacatggatccatgctgcactggcagc  
tttgatggggacctcaggaaacttgaaacctgcaaaagtgtttggtgag  
gagggacaaagaagaccatgacatggcagctttgccactactttcagaca  
cttgaccatgctgactgcagaggtggcatcaccagcaagcagcacacag  
gcctctcctttacatgaactggctaaccaggagtgaggatgattgctaaa  
caagagcttttaactgccagaaaaaagaaaaaagaaaaaagaaaga  
aaaaaaacccaggaagtccaaggataaagaaaaataatgggtgacat  
tggtcaggtcatgatttggcagaccattttttctgagtgctgtatttt  
tgcaggtgagttattttctttgcttattctttcctggacacagagccaa  
catgagcaataactgccagcagggtcgcacctgcagcagctctggactgg  
tgttagaggtgctgctcagagcagcctggagccatgcctgtcctc  
atgctcaccccatgctgtcataatgcctatttcttatcctctcccttga  
agtctgccttatgtttgccttggaattccgtacatggcctccttcagg

```
ccctccctccacagccccaggtctgatccattttctgctgttgccctgag
gtccccaagctgggtgagttcctccctagccctccaaactccttcaccc
ccagtcgtgccccactgcatggcctcatgtcccttggtgccccacactc
ctcgagctctctccctgggcagccatttcacagaccagctgcaggcacac
agggagagcaacaacctcaagttcattccattccagtggtgtttattttg
ttccggtaaaatgcctggaatcatctcactgctcatgcccctgagtacac
gtgggtgtgcatgatccttgtaaaacgcgatgtgagttacagaaccaa
tgccactcccaacagaggcactgtgctgtttcctactgtctttatgcc
ggtctgtcacagctctccttgccctcagttctcacaggagcgcagtgcg
taatttacctccactatcaccaatcaaacagagaagagaaaatcatgaca
gtgctactatgacaacaacaagcccacctggccttcctgtcctgcac
agcgagggcaccccgctcctgcccgtgcaacagccaccgcgaagcgctc
cccccccgcagcggcacccgcccccgccctgcgcacggcccgatct
```

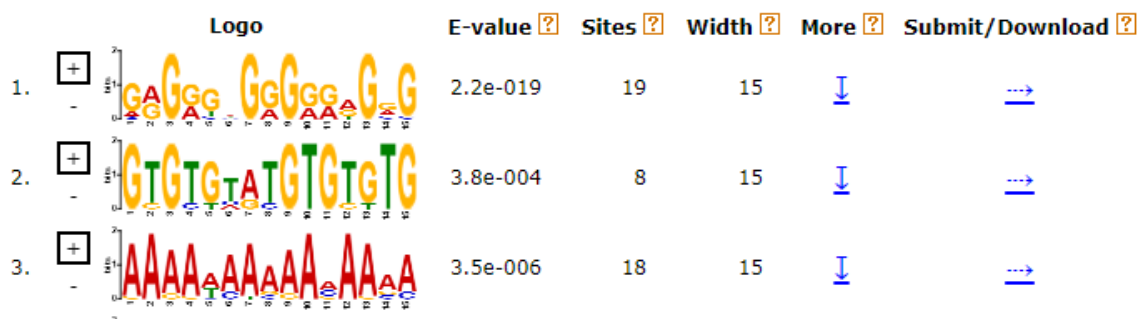
## Annex8.Resultats MEME.



## MOTIF LOCATIONS



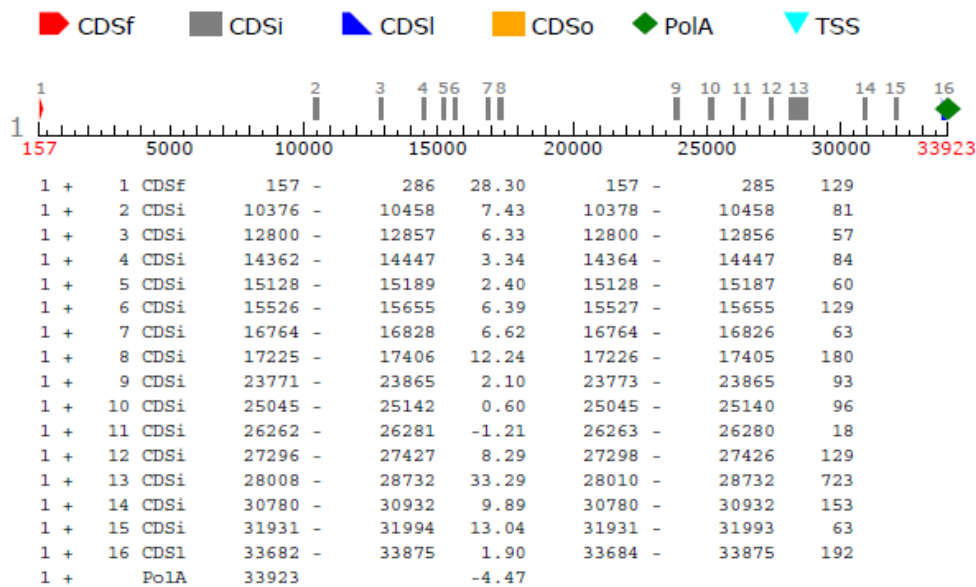
## Annex8b. 10 millors motius de 30.





## Annex9.

FGENESH 2.6 Prediction of potential genes in Homo\_sapiens genomic DNA  
 Seq name: human  
 Length of sequence: 37571  
 Number of predicted genes 1: in +chain 1, in -chain 0.  
 Number of predicted exons 16: in +chain 16, in -chain 0.  
 Positions of predicted genes and exons: Variant 1 from 1, Score:115.993872



Predicted protein(s):

>FGENESH:[mRNA] 1 16 exon (s) 157 - 33875 2277 bp, chain +  
 ATGGCCCCGCCATGCAGCCGGCCGAGATCCAATTTGCCAGCGGCTGGCGTCCAGCGAG  
 AAGGGCATCCGGGACCGAGCGGTGAAGAAGCTGCGCCAGTACATCAGCGTGAAGACGCAG  
 AGGGAGACAGGAGGTTTTCAGTCAGGAAGAACTTCTGAAATCTGGAAGGGGCTCTTCTAC  
 TGCATGTGGGTGCAGGATGAACCCCTTCTACAGGAAGAGCTCGCCAACACCATTGCACAG  
 CTAGTCCATGCTGTTAACTCACTCAGCGGCTCAACACCTGTTTCATTGACACCTTTTGGCAA  
 ACCATGAATCGAGAATGGAAGGAATAGACAGGCTACGCCTGGACAAATACTATATGCTG  
 ATTCGCTGGTCTGAGGCGATCCTTTGAAGTCTTGAAGCGAAATGGCTGGGAAGAAAGC  
 CGAATCAAGGTTTCTTGGATGCTGATGAAGGAGGCTGTGTCCTGAGAGTCAGTCT  
 CCTAATGGAGTGAGATTCACCTTTCATTGATATTTACCTGGATGAACCTCCAAAGTCGGG  
 GGAAGGAGCTTTTAGCAGATCAGAATCTCAAGTTTATCGATCCATTCTGCAAAATTGCT

CGGAAGACGAAGGACCACACCCTGGTACAGACCATAGCTCGGGGTGTCTTCGAAGCTATC  
GTAGATCAGTCTCCTTTTGTGCCTGAAGAGACGATGGAGGAACAGAAGACAAAAGTGGGT  
GATGGTGACCTCTCTGCTGAGGAGATACCTGAAAATGAGGTATCCTTGAGAAAGAGCTGTC  
AGTAAAAAGAAGACAGCACTGGGCAAAAACCATCCAGAAAAGATGGACTCAGTGATGAA  
AGAGGAAGAGATGACTGTGGAACCTTTGAGGACACAGGGCCCCCTCTCCAGTTTGACTAT  
AAGGCTGTTGCTGATCGACTCCTGGAATGACCAGCAGGAAGAACACGCCCCACTTCAAC  
AGGAAGCGCCTCTCCAAACTCATCAAGAAATTCGAAGACCTTTCTGAAGGAAGCAGTATA

TCTCAACTCAGTTTTGCGGAGGACATTTCTGCTGATGAAGATGACCAAACTCCTCAGTCAA  
GGAAAGCATAAGAAGAAAGGAAATAAACTTTTAGAGAAAATAACTTGGAAGAGGAGAAA  
GGAAGCAGAGTCTTTTGTGTAGAGGAAGAGGACAGTGAAAGCAGTCTTCAAAGAGAAGA  
AGGAAGAAGAAGAAGACACCCTGCAGCCTGAAAAATCCAGGCCAGGGGGTGCAGCC  
CCATCCCTGGAACAGAACCGGGGAGGGAGCCCCGAGGCCCTCTGGGCTGAAAGCCCTGAAG  
GCACGTGTGGCCGAGCCAGGTGCAGAGGCCACGTCCAGCACTGGGGAGGAGAGTGGCTCC  
GAGCATCCTCCAGCGTGCCCATGCACAATAAAAGGAAACGGCCACGGAAGAAGAGCCCG  
AGGGCCCCACAGGAAATGTTGGAATCAGCAGTGTGCCCCAGAGGACATGTCTCAGAGT  
GGCCCCAGTGGCAGTCATCCTCAGGGACCTAGAGGGTCCCCGACAGGTGGAGCCCAACTC  
CTAAAAAGGAAGCGGAACTTGGAGTTGTGCGCGTCAATGGCAGTGGCCTGTCCACGCCG  
GCCTGGCCTCCATTGCAGCAGGAAGGCCCTCCACAGGCCCCGAGAGGGGGCGAACAGC  
CACACCAGCTGCCCCAGCGCAGGAGGCTGCAGAAAAAGAAGGCAGGGCCCGGCAGCCTG  
GAGCTCTGTGGCTGCCCCAGCCAGAAAACAGCAAGTTTGAAAAAGAGGAAGAAAATGAGA  
GTGATGTCAAACCTGGTGAGCACAACGGGGTGTGAGTCCGAAGCTGGGCAACCCAG  
GCTCTGGGAAGCAGTGGGACTTGCAGTTCCTGAAGAAGCAGAAGCTGAGGCGAGAGAGC  
GACTTTGTGAAGTTTGACACCCCTTCTTACCAAAGCCCTGTTCTTCAGAAGAGCCAAG  
AGCAGCACTGCCACCCACCTCCAGGCCCTGCCGTCCAGCTAAACAAGACACCATCCAGC  
TCCAAGAAAGTCACCTTTGGGCTGAACAGAAACATGACTGCCGAATTCAGAAGACAGAC  
AAGAGTATCTTGGTCAGTCCACGGGCCCTTCTCGAGTGGCCTTCGACCCTGAACAGAAG  
CCCTCCACGGGGTGTGAAGACCCCAACGCTCACCTGCCAGCTCACCCCTGGTGGCC  
AAGAAGCCCTGACCACCACCAAGGAGAAGGCCAGGGCTATGGATTCTTCTGA  
>FGENESH: 1 16 exon (s) 157 - 33875 758 aa, chain +  
MAPAMQPAEIQFAQRLASSEKGIRDRAVKKLRQYISVKTQRETGGFSQEELLKIWKGLFY  
CMWVQDEPLLQEELANTIAQLVHAVNNSAAQHFIQTTFWQTMNREWKGIDRLRLDKYYML  
IRLVLRQSFVFLKRNGWEESRIKVFLDVLMKEVLCPESSQSPNGVRFHFIDYLDLSKVG  
GKELLADQNLKFIDPFCKIAAKTKDHTLVQTIARGVFVAIVDQSPFVPEETMEEQKTKVG  
DGDLSAEIIPENEVSLRAVSKKKTALGKNHSRKDGLSDEGRDDCGTFEDTGPLLQFDY  
KAVADRILLEMTSRKNTPHFNRKRLSKLIKKFQDLSEGSSISQLSFAEDISADEDDQILSQ  
GKHKKKGKMKLEKTNLEKEKGSRVFCVEEEDSESSLQRRRKKKKHHLQPENPGPGGAA  
PSLEQNRGREPEASGLKALKARVAEPGAEATSSSTGEESGSEHPPAVPMHNKRKRPRKKS  
RAHREMLES AVLPPEDMSQSGSPGSHPGPRGSPTGGAQLKRRKRLGVVPVNGSGLSTP  
AWPPLQQEGPPTGPAEGANSHTTLPQRRRLQKKKAGPGSLELCGLPSQKTASLKKRKKMR  
VMSNLVEHNGVLESEAGQPQALGSSGTCSSLKKQKLRAESDFVKFDTPLPKPLFFRRAK  
SSTATHPPGPAVQLNKTPSSSKKVTFLNLRNMTAEFKKTDKSIILVSPTGPSRVAFDPEQK  
PLHGVLTPTSSPASSPLVAKKPLTTTPRRRPRAMDF

## Annex10.

>FGENESH: 1 16 exon (s) 157 - 33875 758 aa, chain +

MAPAMQPAEIQFAQRLASSEKGIRDRAVKKLRQYISVKTQRETGGFSQEELLKIWKGLFY  
CMWVQDEPLLQEELANTIAQLVHAVNNSAAQHFIQTTFWQTMNREWKGIDRLRLDKYYML  
IRLVLRQSFVFLKRNGWEESRIKVFLDVLMKEVLCPESSQSPNGVRFHFIDYLDLSKVG  
GKELLADQNLKFIDPFCKIAAKTKDHTLVQTIARGVFVAIVDQSPFVPEETMEEQKTKVG  
DGDLSAEIIPENEVSLRAVSKKKTALGKNHSRKDGLSDEGRDDCGTFEDTGPLLQFDY  
KAVADRILLEMTSRKNTPHFNRKRLSKLIKKFQDLSEGSSISQLSFAEDISADEDDQILSQ  
GKHKKKGKMKLEKTNLEKEKGSRVFCVEEEDSESSLQRRRKKKKHHLQPENPGPGGAA  
PSLEQNRGREPEASGLKALKARVAEPGAEATSSSTGEESGSEHPPAVPMHNKRKRPRKKS  
RAHREMLES AVLPPEDMSQSGSPGSHPGPRGSPTGGAQLKRRKRLGVVPVNGSGLSTP  
AWPPLQQEGPPTGPAEGANSHTTLPQRRRLQKKKAGPGSLELCGLPSQKTASLKKRKKMR  
VMSNLVEHNGVLESEAGQPQALGSSGTCSSLKKQKLRAESDFVKFDTPLPKPLFFRRAK  
SSTATHPPGPAVQLNKTPSSSKKVTFLNLRNMTAEFKKTDKSIILVSPTGPSRVAFDPEQK  
PLHGVLTPTSSPASSPLVAKKPLTTTPRRRPRAMDF

>/tmp/05\_08\_20-11:44:37.fasta|GENSCAN\_predicted\_peptide\_1|897\_aa

MAPAMQPAEIQFAQRLASSEKGIRDRAVKKLRQYISVKTQRETGGFSQEELLKIWKGLFY

CMWVQDEPLLQEELANTIAQLVHAVNNSAAQHFIQTFWQTMNREWKIDRLRLDKYYML  
 IRLVLRQSFEVLKRNGWEESRIKVFLDVLMEVLCPEQSPNGVRFHFIDIYDELKSVG  
 GKELLADQNLKFIDPFCKIAAKTKDHTLVQTIARGVFEAIVDQSPFVPEETMEEQKTKVG  
 DGDLSAEEIPENEVSLRAVSKKKTALGKNHSRKDGLSDERGRDDCGTFEDTGPLLQFDY  
 KAVADRLEMTSRKNTPHFNKRKLSKLIKQDLSEGSSISQLSFAEDISADEDDQILSQ  
 GKHKKKGNKLEKTNLEKEKGKQELQGALGGGCLMTTRDLWFLPLSPKISNGTISVPYV  
 FINGQKEGFQSQLGMEEVGPDDKGSRVFCVEEEDSESSLQKRRRKKKKHHLQPENPGPG  
 GAAPSLEQNRGREPEASGLKALKARVAEPGAEATSSTGEESGSEHPPAVPMHNKRKRPRK  
 KSPRAHREMLESVLPPEMDSQSGPSGSHPGPRGSPTGGAQLLKRKRKLGVVPVNGSGL  
 STPAWPPLQKEGPPTGPAEGANSHTTLPQRRRLQKKKAGPGSLELCGLPSQKTASLKKRK  
 KMRVMSNLVEHNGVLESEAGQPQALAAHLNLPPEPVCQRHWAHTSESQVRDPVSLWVA  
 VSCCTRNECPGPASVVLCKVPELCRMEGLSASAVRKTAGRRGSSGTCSSLKKQKLRAESD  
 FVKFDTPLPKPLFFRRAKSSTATHPGPAVQLNKTSSSKKVTFGLNRNMTAEFKKTDK  
 SILVSPTGPSRVAFDPEQKPLHGVLTPTSSPASSPLVAKKPLTTTPRRRPRAMDFF

>human\_1|geneid\_v1.2\_predicted\_protein\_1|622\_AA

MAPAMQPAEIQFAQLRASSEKGIKRAVKKLRQYISVKTQRETGGFSQEELLKIWKGLFY  
 CMWVQDEPLLQEELANTIAQLVHAVNNSAAQACVWFFSRIKVFLDVLMEVLCPEQSPN  
 GVRHFHIDIYDELKSVGGKELLADQNLKFIDPFCKIAAKTKDHTLVQTIARGVFEAIVD  
 QSPFVPEETMEEQKTKVGDDLSAEEIPENEVSLRAVSKKKTALGKNHSRKDGLSDERG  
 RDDCGTFEDTGPLLQFDYKAVADRLEMTSRKNTPHFNKRKLSKLIKQDLSEGSSISQ  
 LSFAEDISADEDDQILSQGKHKKKGNKLEKTNLEKEKGSRVFCVEEEDSESSLQKRRRK  
 KKKKHHLQPENPGPGGAAPSLEQNRGREPEASGLKALKARVAEPGAEATSSTGEESGSEH  
 PPAVPMHNKRKRPRKSPRAHREMLESVLPPEMDSQSGPSGSHPGPRGSPTGGAQLLK  
 RKRKLGVVPVNGSGLSTPAWPPLQKEGPPTGPAEGANSHTTLPQRRRLQKKKAGPGSLEL  
 CGLPSQKTASLKKRKKMRVMSNLVEHNGVLESEAGQPQALVRWEHPQASSPQRHSLASMG  
 LHCLLRGRVGAGGQASGLSSS\*

## Annex11.

>FGENESH: 1 16 exon (s) 157 - 33875 758 aa, chain +

MAPAMQPAEIQFAQLRASSEKGIKRAVKKLRQYISVKTQRETGGFSQEELLKIWKGLFY  
 CMWVQDEPLLQEELANTIAQLVHAVNNSAAQHFIQTFWQTMNREWKIDRLRLDKYYML  
 IRLVLRQSFEVLKRNGWEESRIKVFLDVLMEVLCPEQSPNGVRFHFIDIYDELKSVG  
 GKELLADQNLKFIDPFCKIAAKTKDHTLVQTIARGVFEAIVDQSPFVPEETMEEQKTKVG  
 DGDLSAEEIPENEVSLRAVSKKKTALGKNHSRKDGLSDERGRDDCGTFEDTGPLLQFDY  
 KAVADRLEMTSRKNTPHFNKRKLSKLIKQDLSEGSSISQLSFAEDISADEDDQILSQ  
 GKHKKKGNKLEKTNLEKEKGSRVFCVEEEDSESSLQKRRRKKKKHHLQPENPGPGGAA  
 PSLEQNRGREPEASGLKALKARVAEPGAEATSSTGEESGSEHPPAVPMHNKRKRPRKSP  
 RAHREMLESVLPPEMDSQSGPSGSHPGPRGSPTGGAQLLKRKRKLGVVPVNGSGLSTP  
 AWPPLQKEGPPTGPAEGANSHTTLPQRRRLQKKKAGPGSLELCGLPSQKTASLKKRKKMR  
 VMSNLVEHNGVLESEAGQPQALGSSGTCSSLKKQKLRAESDFVKFDTPLPKPLFFRRAK



SSTATHTPPGPAVQLNKTSSSKKVTFGLNRMNTAEFKKTDKILVSPTGPSRVAFDPEQK

PLHGVLTPTSSPASSPLVAKKPLTTTPRRRPRAMDF

>/tmp/05\_08\_20-11:44:37.fasta|GENSCAN\_predicted\_peptide\_1|897\_aa

MAPAMQPAEIQAQRLASSEKGIRDRAVKLRQYISVKTQRETGGFSQEELKIWKGLFY

CMWVQDEPLLQEELANTIAQLVHAVNNSAAQHFIQTFWQTMNREWKIDRLRLDKYYML

IRLVLRQSFEVLKRNGWEESRIKVFLDVLMKEVLCPEQSQSPNGVRFHFIDIYDELDSKVG

GKELLADQNLKFDIDPFCKIAAKTKDHTLVQTIARGVFEAIVDQSPFVPEETMEEQKTKVG

DGDLSAEIPENEVSLRRRAVSKKKTALGKNHSRKDGLSDERGRDDCGTFEDTGPLLQFDY

KAVADRLEMTSRKNTPHFNKRRLSKLIKKFQDLSEGSSISQLSFAEDISADEDDQILSQ

GKHKKKGKNGKLEKTNLEKEKGKQELQGALGGGCLMTTRDLWFLPLSPKISNGTISVPYV

FINGQKEGFQSQLGMEEVGPDDKGSRVFCVEEDESSESLQKRRRKKKKHHLQPENPGPG

GAAPSLQNRGREPEASGLKALKARVAEPGAEATSSTGEESGSEHPPAVPMHNKRKRPRK

KSPRAHREMLESVLPEDMSQSGPSGSHPGQPRGSPTGGAQLLKRKRKLGVVPVNGSGL

STPAWPPLQQEGPPTGPAEGANSHTTLQRRRLQKKKAGPGSLELCGLPSQKTASLKKRK

KMRVMSNLVEHNGVLESEAGQPQALAAHLNLPEPPVCRQRHWAHTSESQVRDPVSLWVA

VSCCTRNECPGPASVVLCKVPELCRMEGLSASAVRKTAGRRGSSGTCSSLKKQKLRAESD

FVKFDTPLPKPLFFRRAKSSTATHTPPGPAVQLNKTSSSKKVTFGLNRMNTAEFKKTDK

SILVSPTGPSRVAFDPEQKPLHGVLTPTSSPASSPLVAKKPLTTTPRRRPRAMDF

>human\_1|geneid\_v1.2\_predicted\_protein\_1|622\_AA

MAPAMQPAEIQAQRLASSEKGIRDRAVKLRQYISVKTQRETGGFSQEELKIWKGLFY

CMWVQDEPLLQEELANTIAQLVHAVNNSAAQACVWFFSRIKVFLDVLMKEVLCPEQSQSPN

GVRHFHFIDIYDELDSKVGKELLADQNLKFDIDPFCKIAAKTKDHTLVQTIARGVFEAIVD

QSPFVPEETMEEQKTKVGDDLSAEIPENEVSLRRRAVSKKKTALGKNHSRKDGLSDERG

RDDCGTFEDTGPLLQFDYKAVADRLEMTSRKNTPHFNKRRLSKLIKKFQDLSEGSSISQ

LSFAEDISADEDDQILSQGKHKKKGKNGKLEKTNLEKEGSRVFCVEEDESSESLQKRRRK

KKKKHHLQPENPGPGGAAPSLQNRGREPEASGLKALKARVAEPGAEATSSTGEESGSEH

PPAVPMHNKRKRPRKKSRAHREMLESVLPEDMSQSGPSGSHPGQPRGSPTGGAQLL

RKRKLGVVPVNGSGLSTPAWPPLQQEGPPTGPAEGANSHTTLQRRRLQKKKAGPGSLEL

CGLPSQKTASLKKRKKMRVMSNLVEHNGVLESEAGQPQALVRWEHPQASSPQRHSLASMG

LHCLLRGRVGAGGQASGLSSS\*

### Annex13.SNPs RRP1B.

chr21	43659829	43659830	rs225448	0	-	
chr21	43660136	43660137	rs73362750	0		+
chr21	43660254	43660255	rs35000523	0		+
chr21	43660278	43660279	rs9978671	0	+	
chr21	43660467	43660468	rs78030818	0		+
chr21	43660590	43660591	rs178743	0	-	
chr21	43660687	43660688	rs78427898	0		+
chr21	43660720	43660720	rs34256469	0		+
chr21	43661117	43661118	rs73905589	0		+
chr21	43661485	43661486	rs170445	0	-	
chr21	43661714	43661715	rs11330761	0		+
chr21	43661715	43661716	rs79698070	0		+
chr21	43661725	43661726	rs35519182	0		+
chr21	43661729	43661730	rs77321313	0		+
chr21	43661730	43661731	rs11330762	0		+
chr21	43661730	43661731	rs78878629	0		+

chr21	43661739	43661740	rs57248208	0	+
chr21	43661919	43661920	rs13052462	0	+
chr21	43661930	43661931	rs67287505	0	+
chr21	43661942	43661943	rs76650127	0	+
chr21	43661943	43661944	rs35522463	0	+
chr21	43661971	43661972	rs79432021	0	+
chr21	43662086	43662087	rs76643017	0	+
chr21	43662088	43662088	rs68105436	0	+
chr21	43662088	43662089	rs74886708	0	+
chr21	43662096	43662096	rs72009683	0	+
chr21	43662106	43662106	rs35823612	0	+
chr21	43662107	43662107	rs72212204	0	+
chr21	43662144	43662145	rs75365894	0	+
chr21	43662340	43662341	rs9979482 0	+	
chr21	43662625	43662626	rs225449 0	-	
chr21	43662660	43662661	rs78356453	0	+
chr21	43662795	43662796	rs178744 0	-	
chr21	43662820	43662820	rs35948230	0	+
chr21	43663244	43663245	rs76690582	0	+
chr21	43663317	43663318	rs7283699 0	+	
chr21	43663551	43663555	rs10590039	0	+
chr21	43663552	43663553	rs3787976 0	+	
chr21	43663552	43663556	rs57250024	0	+
chr21	43663554	43663555	rs3787977 0	+	
chr21	43663556	43663557	rs691748 0	-	
chr21	43663556	43663557	rs67420624	0	+
chr21	43663556	43663558	rs1620215 0	+	
chr21	43663556	43663562	rs57323059	0	+
chr21	43663556	43663564	rs61243535	0	+
chr21	43663557	43663558	rs36033318	0	+
chr21	43663562	43663563	rs8133786 0	+	
chr21	43663564	43663565	rs60279758	0	+
chr21	43663568	43663569	rs60745406	0	+
chr21	43663586	43663590	rs72216559	0	+
chr21	43663587	43663591	rs72172023	0	+
chr21	43663590	43663592	rs72232474	0	+
chr21	43663597	43663601	rs10590040	0	+
chr21	43663598	43663599	rs28377560	0	+
chr21	43663598	43663602	rs71878824	0	+
chr21	43663600	43663601	rs3787978 0	+	
chr21	43663826	43663827	rs76031955	0	+
chr21	43663937	43663937	rs11402831	0	+
chr21	43663937	43663938	rs66687826	0	+
chr21	43663945	43663946	rs61440193	0	+
chr21	43663946	43663946	rs34349454	0	+
chr21	43663947	43663947	rs67848320	0	+

## Annex14.

### CNS retrieval options:

☐ Remove gaps    Add  extra bases upstream (5') and downstream (3') of CNS

```
>Human Mar. 2006 chr21:43659925-43660040 (+)
AACACAAACGTATAGTACAAAATGT-AACAAATAAGTAGTGTCTACCTGTCTCCCATGTA
GATACATATACAGGATTTGCTTTAAAAAACAACAAAACCTTAAAAACACTGACCTA
>Mouse Jul. 2007 chr17:31982234-31982339 (+)
AACACAAAATGTA-TACAAAATAGAAAAAATAAGTAGTGT----CTGTTTCCTATTTA
TATACACATACAGTATTTGCTTTAAAAAGA-----AACTTTACAAACACTGACCTA
= length = 117bp, identity = 76.1%, type = UTR
>Human Mar. 2006 chr21:43660184-43660269 (+)
CAGCCTC-CAATATTGCACAACCTCACCCAAGTCAGGTGTTCTTGTGGACCCC-ATCAT
GAAATGCATAAA-----CGTCAGCAGCAGCGGA
>Mouse Jul. 2007 chr17:31982457-31982548 (+)
CAGCCCCCTGGTATTGCACAATTCCC-CCAAGTCAGGTGTTCTAGCGGGACCCTGATAAT
GAAATGCATAAAATCTGATAATTAGCAGCAGA
= length = 93bp, identity = 72.0%, type = UTR
>Human Mar. 2006 chr21:43660818-43661021 (+)
CCCAGACACG-----CCCCTCAATTATC---TTTCTTCTCAAAGTTTCCAGTCCCTACT
CAAAGTTGAGGTCACTCAGTGCAAGAAAAAATAAAACAAA-----AATAAACT
---AAAAAGCCAAACACCTAACGTATTGCTTATTTTCGCTTCTAGAAAGTTCTTGGC-TT
TGAGGCCAGA---TGTCTTCACCTCCTCTGAAGTTATTGCT-GTAAGAGTCA
>Mouse Jul. 2007 chr17:31982939-31983161 (+)
CCCA----CGTGACATACCTCAATTATTATTTTCTTCCAGAAAGTT-----TCCTACT
TATCATTTGAGGTCACTCAGTGCAAGGAGAAAAACAAAACAAAAACAACAACAAAAAT
CCTAAAAAGCCCATACACCAAAACGTATTGCTTGTTCACGTCTAGAAAGTTCTTGGCTTT
TGAGTCCGGAAACCATCTTCACATACTCCGAAATTATTGCTTGGGAAGAGTCA
= length = 232bp, identity = 71.1%, type = UTR
>Human Mar. 2006 chr21:43661050-43661293 (+)
TCACTGCACCAGGACAAACGTGCCTAGGAGCAGGGCATCAGGTCCCTCCATCTCACAGTC
```

CCCCTGCAGCAGCCCCAGGGGCTCACAACCTGGGGCCAGC----CTGGCCAGGCGTGGTG  
 GGGGCACAGCGGGGAGGGCGGTGGGGCCGGTGCCAATGTGCAGGTGTGTCTCCAGGAGCT  
 GCGCCGCTGAGGCCACCGGGGACGCGCCGGTCTGCAGGAGTGGGGGCGGCAGCAGCGGA  
 GCCCCGAC  
 >Mouse Jul. 2007 chr17:31983228-31983474 (+)  
 TCACTGTACCAGGACGAATGTCCCCCGTGGCCAGAGGTCAAGTCCCTCCATCTCACAGTC  
 CCCTGTGGCAGGCCAGCTGGATCACAGCCTGGGGACAGCCTGG-TGAGGCACTGTGGCA  
 GGGGCCCAGTGGGAGGGGCCACTGGGCCAGCACTGATGTGCAAGTGGGTATCCAGGAGAT  
 GCGCAGCAGACGCCACAGGGGACATGCCAGCCTGGAGAAGGGGTGTTGGCAGCAGTGAA  
 TTCCAGAC  
 = length = 248bp, identity = 71.4%, type = exon  
 >Human Mar. 2006 chr21:43661328-43661425 (+)  
 GGCCATCACAGGGGCGATCACAACCGGGCAGGGGCCGGCTGGGGGGCCCTGGGAGCAGC  
 CGGGTGCAGCGGCC-----GGGTGGTGTCTGTAACGGAGCAGC  
 >Mouse Jul. 2007 chr17:31983482-31983588 (+)  
 GGCTGTACAGGGAGCAAGCACATAGGGGACCGGGGACAGCTGGGGGCCCTGCTGGCAGC  
 CAGATGAGGCGGCGCGGTAGAAGAGTGGTGTCTGTAACGGAGCAGC  
 = length = 107bp, identity = 72.0%, type = exon  
 >Human Mar. 2006 chr21:43661851-43662082 (+)  
 CTCTGCTGCTCTAGCACCTCCTCCAGCAGGCTCCAGCCCTCCCGGCTGCCGGCTGCGCCC  
 CCGTGCAGGCCTGGGCTCTGTGCAGGGGCGTGGAAAGGGCTCAGGCCGCCCTGCTGGCC  
 CGGCTGGCGGGGGCTGGCACACCTGGCGAGCCAGCCCTTGATTTTGTTCAGTCCCAGA  
 AACCCTTTGGTCCGCGTGGTCTTCTCAGCTGCTGCCGAAAGGCCTTCAGCC  
 >Mouse Jul. 2007 chr17:31983978-31984206 (+)  
 CTCTGCTGGTGCAGGACCTCTTCAAGTAGGCTCCTGCCCTCCCGATTGCTGGTCTGTCAG  
 CCTTGCAGGCCGGAGCTTGGGGCTGGGGGTGGGAAGGTACTCATCCCTCCCCGGGGAGTT  
 CGGACGGAGGA---CTGGCACACCTGGCGAGCCAACCCCTTGTATCTTGTTCAGTCCCAGG  
 AACCCCTTGGTCTCGCATTTTCTCCTCAGCTGCTGCCGGAAGGCCTTCAGCC  
 = length = 232bp, identity = 77.6%, type = exon  
 >Human Mar. 2006 chr21:43662570-43662847 (+)  
 TGAGTCAGTGAGGTGTCCGACGCCCCCGCTCCCTCCTGGAAGCTGACAGGGAGCAGAACA  
 GCTCCTCCCAAGCCCCCTGAGCCTGCAGCACTGGGGTGGCGGACTGCGACCCCAAGGAAG  
 GGCAGGGCCAGCCTGACCGGGGAGCAGGCGCCCAGCAGCCCTGAGTGGCCGGGGTGCCA  
 CTGAGCCCCGCGGGGCTTTGCTCGCAGAGAAGGTACAGCAACTGTCAGAGCTGGTTCCC  
 TCTGCAGGACTTGCCGTGGTGGAGGGGAGACGACTAT  
 >Mouse Jul. 2007 chr17:31984842-31985119 (+)  
 TGAGTGAGAGAGGTATCAGACGCTCTCCGTCCCTCCTGGAAGCTGACGGGAGGTAAGACG  
 GCTGTGCCCAGACCTGCCTGAGTCTGGAGCACAGGGGTGGCTGACTGTGATCCCAGGAAG  
 GGCAGGGCCAATCTGACTGGAGAGCTGGTGGCCAGCAGCCCTGGGGTGGCCAGGCCACTG  
 CCAAGCCCTGCAGGACCTTCACTTGCAGAGAAGGGGAGGCAGCTGTCGGAGCTGGTCCCT  
 TCCGAGGGACTCGCCGTGGCAGAGGAGGAGACAATTAT  
 = length = 278bp, identity = 76.3%, type = exon  
 >Human Mar. 2006 chr21:43663331-43663545 (+)  
 GGCGCGGTGAGTGGGGAGAGGCGGGTGGAGACCTCGGCCAGGGTGTGCCTCCGGCCCGTG  
 CTGCTGGGCAGGGACTCCTGCGTGTCTGCTCCTCCTTAGGCCCGGCCCTGCCTGGCC  
 TCCTCACTGATGGCTGTGTCCAGCAGGTGCTTGGGGACACGGGCCGGGGCCGGAACACT  
 CCGCTGCAGCTGGCATCCACCGGAAGAACAAGGG  
 >Mouse Jul. 2007 chr17:31985579-31985793 (+)  
 GGAGGGTTGAGCGGGGAGAAATGGGTGGAGACTTCAGCCAATGTGTGCCTCCGGCCTGTG  
 CTTCGGGCAGGGGTTCCTGGACCTCCTGTTCTCCTCTAGGCTGGGACCTGCCTGGCC  
 TCCTCGCTGATAGCTGTGTCCAGCAGACTGCTGGGGGAGATGGATCGGTGCCGGAACACT  
 CCCTGCACTTGGTATCCAGGGGGAATAATAAGGG  
 = length = 215bp, identity = 81.4%, type = exon  
 >Human Mar. 2006 chr21:43663662-43663786 (+)  
 CAC-TGCAGCGAGCTCTGGAGCTCACAGTCCATCTCGGCCTGGAGGACGGAAGTGCACCAA  
 GGTCTGCGGCTGCGGGCAGCAAGGCAGGTGCGAAAGGGTGGTGGAAAGACCTTCCTG  
 AGGCAC  
 >Mouse Jul. 2007 chr17:31985901-31986026 (+)  
 CACCTGAAGTGAGCTGTGGAGATCACAGTCTATCTCAGCCTGCAGGACAGACTGAGCCAA  
 GGCTTGGGGCTGTGGGCACAGCAGAGAGGGCCGAAAGGGTCACACGGGAGAATTCTTTG  
 AGGAAC  
 = length = 126bp, identity = 77.0%, type = exon  
 >Human Mar. 2006 chr21:43664167-43664313 (+)  
 CTCCAAACCACTGAGGTCCGAGCTCCGAGGCGCGGCTGCCTGGCAGGCCCGGGG-----  
 -CGGGCGCAC---TGGGCATTCCGATACTCCTTGAGCCGCTCAAGGAGGAGGTAATAAAT  
 GGCAGCAAAGTGGTTATAGCTGCTGTTTTGCAGTGA  
 >Mouse Jul. 2007 chr17:31986464-31986619 (+)  
 CTCAGACTGCTGAGGTCTGAGCTTCGGAGCTGGGGCTGTCTGGTGGGTGCAGGGGTGGG  
 CCGGGATGAGGGCTGGGCGCTTCGATGCTCCTTGAGGCGCTCGAGTAGGAGGTAGTAAAT  
 GCGGGCAAAGTGGTTGTAGCTGCTGTTCTGCAGAGA  
 = length = 156bp, identity = 74.4%, type = exon  
 >Human Mar. 2006 chr21:43664542-43664765 (+)  
 CTCACCGCTCCTCTGCCGGTCCACGCCCAGGGTCTGCATGATACCCAGCGCCTGCTCATC  
 GTAGTCGCCCCAGGTGGAGGTGTAGCTGTGTGCGGAGAAGGCGGGGCAGGCGGGTCCCGG  
 CAAGCAGGGCTCAGCCCCGATCCACCGGTGCTGCCGGATCTGGGCGATGGTGATGCGCCT  
 GGCGGGGTCCACCACAGCATGCGGCGGATCAGGCTCTCACAGT  
 >Mouse Jul. 2007 chr17:31986873-31987096 (+)  
 CTCTATAGTCCGCTGCCGGTCTGATGCCGAGGGCTGCATGATCCCTAGCACCTGTTCGTT  
 GTAGTCGCCCCAGGTGGAGGTGTAGCCTTGCATGTCGAAGGCAGGGTTCCTGCTGCAG  
 GAGAGTGGGGTGGGCTGCATCCACCGGTGCTGGCGGATCTGGGCTATGGTGTATGCGCTT  
 AGCGGGGTCCACGACCAGCATGCGTCTGATCAGCGTCTCACAGT  
 = length = 224bp, identity = 78.1%, type = exon

>Human Mar. 2006 chr21:43665318-43665441 (+)  
CTTGAGACATGAAGAAGGGGATGCGGAAGCGGCCCTCCAGCACCCGCTGTCTCAGCGTCCG  
GCAGGTTAGGCCATCGAAGGGGAGAGAACCGCAGACCAGGACGTACAGCACCACGCCCA  
GGCT

>Mouse Jul. 2007 chr17:31987671-31987794 (+)  
CTTGAGACATGAAGAAGGGGATGCGGAAGCGGCCCTCCAGTACCCGCTGTCTCAGCGTAG  
GCAGGTTGGGCCGTCGAAGGGGAGGACCCACAGACCAGGACGTACAGCACCACACCGA  
GGCT

= length = 124bp, identity = 92.7%, type = exon

>Human Mar. 2006 chr21:43665551-43665675 (+)  
CCAGATGTCCAGCTGGGGGCTTCTACTCTCCCTCAAAGACTTCCGGGGCGGCATA  
CGGGGGGCTCCCAACCACGTGGACAGAGGCTCTCCTGACTTGTAGAAATTTCCAAATCC  
AAAAAT

>Mouse Jul. 2007 chr17:31987886-31988010 (+)  
CCAGACGTCCAGCTGGGGACCCTCGTACTCTCCCTCGAAGACTTCCAGGGGCTGCATA  
CGGGGGGCTCCCAACCACGTGGACAGAGGCTCGCCTGGCTTGTAGAAATTTCCAAATCC  
AAAAAT

= length = 125bp, identity = 92.8%, type = exon

>Human Mar. 2006 chr21:43665946-43666107 (+)  
CTGCCAGCTTGATGTCCATGTTGCCATCCAGCAGGAGGTTCTCGGTCTTGAGGTCCCGGT  
GGACGATGTGATGGTCGTGACAGTACTCCACGGCCGACAGGATTTGCCAGAACTTCTTCC  
GCGCTCGTTCTCACTCAGGTGCCCCGTTGGAAGTCAAAATAAT

>Mouse Jul. 2007 chr17:31988180-31988341 (+)  
CTGCTAGCTTGATATCCATGTTGTGTCCAGGAGCAGGTTCTCCGTCTTGAGGTCCCGGT  
GGACAATGTGGTGGTGTGGCAGTACTCCACGGCTGACAGAATCTGCCAGAACTTCTGCC  
TAGCCTCGTTTTCTCACTCAAGTGCCCCGTTGGAAGTCAGATAAT

= length = 162bp, identity = 87.7%, type = exon

>Human Mar. 2006 chr21:43666282-43666345 (+)  
CAAACATTTCTCCATTTTTTAGCAAATTCAGTGACGATGTAAAGCATGTCTTTGTTTCCA  
TAAC

>Mouse Jul. 2007 chr17:31988478-31988541 (+)  
CAAACATTTCTCCATTTTTTGCAAATTCGTGACAATGTAGAGCATATCCTTTGTCTCCA  
TAAC

= length = 64bp, identity = 90.6%, type = exon

>Human Mar. 2006 chr21:43669255-43669355 (+)  
CTACTGTCTTGCAGAAATAAATAATGAAAACC-AGTAAGAGCAGCTGGCTTT--AAAG  
TTTAATAATTATCTGCAAAGGCAGAAAGTTTCCAGTTTCACTCAA

>Mouse Jul. 2007 chr17:31990715-31990816 (+)  
CTACTGCACCC-C-AAGACACTTTAAGGAGACCTGGGAAAGAACAGCTGCTTTTTAAACG  
TTTAGCAATCCTCTGCAAAGGCGGAAGTTCCAGTTTTCGCTCAA

= length = 104bp, identity = 71.2%, type = intron

>Human Mar. 2006 chr21:43669715-43669831 (+)  
CTGGTAAAGCTTTATGATGTGTGGATGGTTTCAAGCTTCATCAGCTGAACCTCACGATA  
GATTTTCTCCAAATTGCTTGAATCTAATCGTGTTTTATCAATTTATTTATTGCAAC

>Mouse Jul. 2007 chr17:31991153-31991269 (+)  
CTGATAAAGCTTGATGATATTTGGGTGGTTCAAAGTTTCATGAGCTGGACCTCCCGGTA  
GATCTTCTCCAGATTGCTAGAACTAACCCTGTCTTGTCAATTATTTTATTGCAAC

= length = 117bp, identity = 85.5%, type = exon

>Human Mar. 2006 chr21:43670331-43670486 (+)  
CTGCGTTTTGGTGACTCGATGCCGCGCCAGCTTACCACCGCGAAGTTGCTTTGCCCAG  
GGTCCGCTCGATGTCGTAAAAACCCACCGGAGGGGCTTCTGTGGCCCTGACCTTGCC  
CGCGGGTCCGCGCTGAACCTCCGACATGATAACCAT

>Mouse Jul. 2007 chr17:31991839-31991994 (+)  
CTGCGTTTTGGTGACTCGGTGCCGCGCCAGCTTAACCACTGCAAAATTGCCTTTGCCCAG  
GGTCCGTTCCACGTGCTGTAAGCCACCGGAGGGGCTTCTGTGGCCCTGGCCAGTGCC  
TGAGGGGACCGCACTGAACCTCCGACATGATACCAT

= length = 156bp, identity = 88.5%, type = exon

>Human Mar. 2006 chr21:43670487-43670500 (+)  
GGCTCCGCGCG-CAC

>Mouse Jul. 2007 chr17:31991995-31992009 (+)  
GGCTTCGGGCGCGAC

= length = 15bp, identity = 73.3%, type = UTR

>Human Mar. 2006 chr21:43671318-43671422 (+)  
CTCCGCCGCCACCGGAGCGCCAGGCCAGGAAGCCGCGCCGCTCGGCGCTGCTC  
GGGTGCCTACTGCTCCGGCTGCCGCGCGCTGCTGCTGCCTCCG

>Mouse Jul. 2007 chr17:31992617-31992721 (+)  
CTCCGCCGCCCTCCGTGAGCCCGACCCGGGATGCCACCCGCTCGGCTGCTAGCTCTGCTC  
CGGTGCCCGCGCTTCGCGCCGCGCCGCAACGCTGCCGCCACCG

= length = 105bp, identity = 74.3%, type = UTR

>Human Mar. 2006 chr21:43671490-43671709 (+)  
CCCGCCCCGGCCCCCCCCCGTAGTCCCCCCCCCTGCCCCGCCCGCCCCGGGCGG  
GGCGCCGCATTTAGCTGCTGCTTTGACGCCAGAGCAGCGCGCGCGAGCGCGGGCGGAG  
GGAGAGCGGGAGATCGGGCGGTTGCCAAGAGACTGGCGGCTCTGACGCGCGCGGG-ATGA  
GGCCGTTGCCCTGGCGACCGCGGCGGTGACGTACGGGGC

>Mouse Jul. 2007 chr17:31992777-31992993 (+)  
CCGCCCCCGCCCCGGCCCCCGCACAAAGCCCGCCTCG-CCCCGCCCGCGCGCAG  
GCACAGCCCATTTGACGTCGTTTTGACGCCAGAGCAGCGCGCGCGCGCTGCTGCGG--A  
GGGGGAGCGGAGATCGGGCGGTTGCCAAGAGACTGGAGGCTCTGACGCGCGCGCTGGTGA  
GGCCGTCGCCATGGCGACCGCGGG-CGGTGACGTACGGGGC

= length = 221bp, identity = 78.3%, type = intergenic

>Human Mar. 2006 chr21:43685080-43685176 (+)  
CTTATTTATTCTGAGAACTCTTTTATATTAAGGCATGAATCCTTTACCTCTTTCTTCC

```

ATGTTTTTTTCTTAGT----TATTTGTCTCTGTGGTTTTT
>Mouse Jul. 2007 chr17:32008698-32008798 (+)
CTTATTTATCCTGAGAGCTCTTTCTATATTAAGGACATTGGTCCTTTATATCTTTCTTAC
CTATCTTGTTTTGCTTTTTTGTTCCTTGGTGGGTATTT
= length = 101bp, identity = 70.3%, type = intergenic
>Human Mar. 2006 chr21:43685318-43685418 (+)
GCTAGTACTTCAGTCAGTGAATTGGCATTGCAAGCGTGTGTCATTAGTACAGACGTCA
GCGGTGAATGAGAAAGCAAAGAACCCGTGGCTGCACCTAAT
>Mouse Jul. 2007 chr17:32008943-32009042 (+)
GCTAGTGCTTCAGCGCGTGG-TTGGCATAAAGAAGTTTATCATCACTTCTACAGACGTCA
CAGGCGCATGTTAGAGCCAAGAAGCCATAGCTGCCCCGTGAT
= length = 101bp, identity = 70.3%, type = intergenic
>Human Mar. 2006 chr21:43685485-43685649 (+)
CTGGTCCAGGTGCAGGTTTTGTAGTCCTCACTCAA-----AAATGAGATAAGGAATGTGA
CTTTGAACCTATTAGCCTTTTGATCAGGCTATGAACCTAAAGGCAATAGAACAGGTGGCC
TGACCCCTGCCAGAGACTGGCACTG-CCTGCGAGGTAGGAGCTCCCAGGTGG
>Mouse Jul. 2007 chr17:32009119-32009278 (+)
CTCCACCCAGTACA-----CAAGTCCTCACTCAATCTTAAAAA-GTGATAAGAAATGTAA
CTTTGATCTGGCTAGCCTTTTGATCAAGCTGTGAATTAAAGGC-AATGGGCGGTGT--GC
CAACCTCTCCCGAGGCTGGCACTGCCCTTCGTGGGAGGAGCCA---GGTGG
= length = 172bp, identity = 68.0%, type = intergenic
>Human Mar. 2006 chr21:43662672-43662847 (+)
GACTGCGACCCAGGAAGGGCGAGGCCAGCCTGACCGGGGAGCAGGCGCCAGCAGCCCC
TGAGTGGCCGGGGTGCCACTGAGCCCCGCGGGGCTTTTGCTCGCAGAGAAGGTGAGACAA
CTGTGACAGCTGGTTCCTCTGCAGGACTTGCCGTGGTGGAGGGGAGACGACTAT
>Mouse Jul. 2007 chr17:31984944-31985119 (+)
GACTGTGATCCCAGGAAGGGCGAGGCCAATCTGACTGGAGAGCTGGTGGCCAGCAGCCCT
GGGGTGCCAGGCCACTGCCAAGCCCTGCAGGACCTTCACTTGACAGAGAAGGGGAGGCAG
CTGTGCGAGCTGGTCCCTTCCGAGGGACTCGCCGTGGCAGAGGAGGAGACAATTAT
= length = 176bp, identity = 76.1%, type = exon
>Human Mar. 2006 chr21:43663331-43663545 (+)
GGCGCGGTGAGTGGGGAGAGCGGGTGAGACCTCGGCCAGGGTGTGCCTCCGGCCCCGTG
CTGCTGGGCAGGGACTCCTGCGTGTCTGCTCCTCTAGGCCCGGCCCTGCCTGGCC
TCCTCACTGATGGCTGTGTCCAGCAGGCTGCTTGGGGACACGGGCGGGGCCGGAACACT
CCGCTGCAGCTGGCATCCACCGGAAGAACAAGGG
>Mouse Jul. 2007 chr17:31985579-31985793 (+)
GGAGGGTTGAGCGGGGAGAAATGGGTGGAGACTTCAGCCAATGTGTGCCTCCGGCCTGTG
CTTCCGGGCAGGGGTTCTGAGACCTCCTGTTCTCCTCTAGGCTGGGACCTGCCTGGCC
TCCTCGCTGATAGCTGTGTCCAGCAGACTGCTGGGGGAGATGGATCGGTGCCGGAACACT
CCACTGCAGTTGGTATCCAGGGGAATAATAAGGG
= length = 215bp, identity = 81.4%, type = exon
>Human Mar. 2006 chr21:43695585-43695695 (+)
CTTCCGCGTGTGGCCATCTGTCTTTCAGAAACCTCTGGCAGAGGGACCCAGCAGATGAC
ACGGGGGCCGTTTACACTCTGAAACCTGCCTG-GAGTCAAAACACTGCTGGC
>Mouse Jul. 2007 chr17:32014877-32014980 (+)
CTTCC--GTTTGGCCATCTGTCTTCCAGAAACCTCTGACAGC-ACATCCGGCAGATGCC
TTGGGG--TGTTCTCAT---AGACCTGTGTGGAAGTCAAAACATGGCTGGC
= length = 112bp, identity = 72.3%, type = UTR

```

#### Conserved intervals sorted by the Human Mar. 2006 coordinates

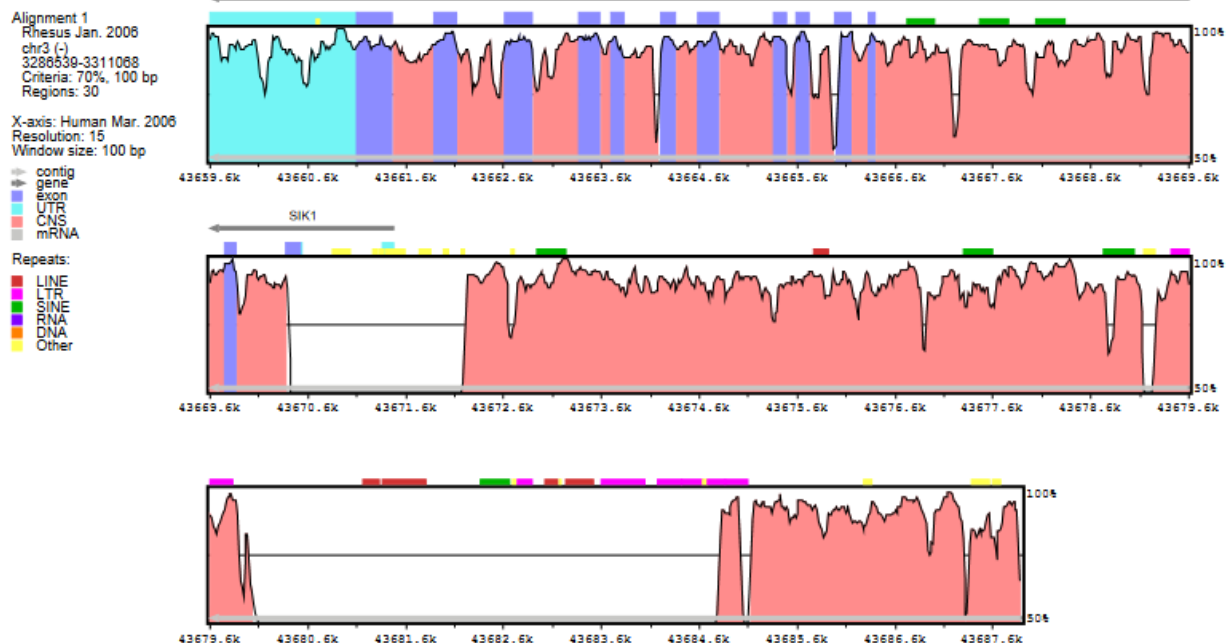
```

***** Conserved Regions - Human Mar. 2006 chr21 (Mouse Jul. 2007 chr17) *****
43659925 (31982234) to 43660040 (31982339) = 117bp at 76.1% UTR
43660184 (31982457) to 43660269 (31982548) = 93bp at 72.0% UTR
43660818 (31982939) to 43661021 (31983161) = 232bp at 71.1% UTR
43661050 (31983228) to 43661293 (31983474) = 248bp at 71.4% exon
43661328 (31983482) to 43661425 (31983588) = 107bp at 72.0% exon
43661851 (31983978) to 43662082 (31984206) = 232bp at 77.6% exon
43662570 (31984842) to 43662847 (31985119) = 278bp at 76.3% exon
43662672 (31984944) to 43662847 (31985119) = 176bp at 76.1% exon
43663331 (31985579) to 43663545 (31985793) = 215bp at 81.4% exon
43663331 (31985579) to 43663545 (31985793) = 215bp at 81.4% exon
43663662 (31985901) to 43663786 (31986026) = 126bp at 77.0% exon
43664167 (31986464) to 43664313 (31986619) = 156bp at 74.4% exon
43664542 (31986873) to 43664765 (31987096) = 224bp at 78.1% exon
43665318 (31987671) to 43665441 (31987794) = 124bp at 92.7% exon
43665551 (31987886) to 43665675 (31988010) = 125bp at 92.8% exon
43665946 (31988180) to 43666107 (31988341) = 162bp at 87.7% exon
43666282 (31988478) to 43666345 (31988541) = 64bp at 90.6% exon
43669255 (31990715) to 43669355 (31990816) = 104bp at 71.2% intron
43669715 (31991153) to 43669831 (31991269) = 117bp at 85.5% exon
43670331 (31991839) to 43670486 (31991994) = 156bp at 88.5% exon
43670487 (31991995) to 43670500 (31992009) = 15bp at 73.3% UTR
43671318 (31992617) to 43671422 (31992721) = 105bp at 74.3% UTR
43671490 (31992777) to 43671709 (31992993) = 221bp at 78.3% intergenic
43685080 (32008698) to 43685176 (32008798) = 101bp at 70.3% intergenic
43685318 (32008943) to 43685418 (32009042) = 101bp at 70.3% intergenic
43685485 (32009119) to 43685649 (32009278) = 172bp at 68.0% intergenic
43695585 (32014877) to 43695695 (32014980) = 112bp at 72.3% UTR

```

## Annex15.

Human Mar. 2006 chr21:43659560-43687848



## Annex16.

Human Mar. 2006 chr21:43659560-43696079

