Basic rsCDS algorithm

1. BLASTX
   - BLAST result
     - Hit 1 e-value 1 → 5' truncated → <1..32
     - Hit 2 e-value 2 → complete CDS → 6..32
     - Hit 3 e-value 3 → coding potential → No CDS

2. Local alignment by FASTY

3. Status determination according to alignment

4. CDS prediction

- CDS’s are predicted if either start or stop codon matches with protein termini.
- Priority: Complete CDS, truncated, E-value
## Status list

<table>
<thead>
<tr>
<th>status</th>
<th>Description</th>
<th>CDS prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete CDS</td>
<td>Both start and stop codon matches with protein termini.</td>
<td>Predicted</td>
</tr>
<tr>
<td>Truncated</td>
<td>Either start or stop codon matches with protein termini. Match to the</td>
<td>Predicted</td>
</tr>
<tr>
<td></td>
<td>termini of cDNA sequence</td>
<td></td>
</tr>
<tr>
<td>Immature</td>
<td>Either start or stop codon matches with protein termini. There is ag,</td>
<td>Not predicted</td>
</tr>
<tr>
<td></td>
<td>gt nearby.</td>
<td></td>
</tr>
<tr>
<td>Alternative terminal</td>
<td>Either start or stop codon matches with protein termini. Start, stop</td>
<td>Predicted but needs</td>
</tr>
<tr>
<td></td>
<td>codon predictable.</td>
<td>special care</td>
</tr>
<tr>
<td>Coding potential</td>
<td>Either start or stop codon matches with protein termini.</td>
<td>Not predicted</td>
</tr>
<tr>
<td>Reverse</td>
<td>Hits in reverse orientation.</td>
<td>Not predicted</td>
</tr>
<tr>
<td>???</td>
<td>No match with neither start nor stop codon.</td>
<td>Not predicted</td>
</tr>
</tbody>
</table>

Additional status: Frame shift(Insertion, Deletion), Unexpected stop codon
Subject protein completely matches cDNA sequence
- There is “ATG” triplet just beginning of matching region
- There is “TAA”, “TAG” or “TGA” at the end of matching region

Homologs of ID10 by BLASTX NCBI-nr
<table>
<thead>
<tr>
<th>query</th>
<th>length</th>
<th>identity</th>
<th>alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mus musculus</td>
<td>92%</td>
<td>402</td>
<td>2136</td>
</tr>
<tr>
<td>Mus musculus</td>
<td>92%</td>
<td>402</td>
<td>2136</td>
</tr>
</tbody>
</table>

Supplementary Information Figure
5' truncated

- Match region starts at first, second or third base in cDNA sequence
- Match region does not start from first amino acid in subject protein
- Match region ends at the end of subject protein, and there is “TAA”, “TAG” or “TGA” at the corresponding region of cDNA sequence
3‘ truncated

- Match region starts at the beginning of subject protein, and there is “ATG” at the corresponding region of cDNA sequence
- Match region ends at the end or 1 or 2 bases upstream from the end of the sequence
- Match region does not end at the end of subject protein
Supplementary Information Figure

5′ immature

- Match region ends at the end of subject protein, and there is “TAA”, “TAG” or “TGA” at the corresponding region of cDNA sequence
- There is “ag” located 2-4 bases upstream of matching start position in cDNA sequence
3' immature

- Match region starts at the beginning of subject protein, and there is “ATG” at the corresponding region of cDNA sequence
- There is “GT” located 0-2 bases downstream of matching end position in cDNA sequence
Alternative N-terminal

- Match region ends at the end of subject protein, and there is “TAA”, “TAG” or “TGA” at the corresponding region of cDNA sequence.
- There is “ATG” located upstream of matching start position in the same reading frame in the sequence.
Alternative C-terminal

- Match region starts at the beginning of subject protein, and there is “ATG” at the corresponding region of cDNA sequence
- There is “TAA”, “TAG” or “TGA” downstream of matching end position in the same reading frame in cDNA sequence

Subject protein

CDS

ATG

stop codon

TAA

homology of ID:373 (by BLASTx NCBI ref)
query 373
1 200 400 600 800
Rattus norvegi
47 100% 250
1 256
602 100% 250
Rattus norvegi
47 92% 250
1 256
602 92% 250
SODIUM- AND CHLORIDE-DEPENDENT GABA TRANSPORTER 3 (GA T3)
SODIUM- AND CHLORIDE-DEPENDENT GABA TRANSPORTER 2
Coding potential with predicted stop codon

- Match region ends at the end of subject protein, and there is “TAA”, “TAG” or “TGA” at the corresponding region of the sequence.
Coding potential with predicted start codon

• Match region starts at the beginning of subject protein, and there is “ATG” at the corresponding region of cDNA sequence
• No stop codon

Homology of ID13849 (by BLASTX NCBI-nr)

<table>
<thead>
<tr>
<th>Query</th>
<th>Match Length</th>
<th>Identity</th>
<th>Expect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homo sapiens</td>
<td>254</td>
<td>74%</td>
<td>241</td>
</tr>
<tr>
<td>Mus musculus</td>
<td>254</td>
<td>79%</td>
<td>240</td>
</tr>
</tbody>
</table>

348 Align Kruppel-like factor 3 (basic)
348 Align hBKLF for basic kruppel like factor
Reverse

- Sequence matching in opposite direction