Query sequences:

1. ACCGGCCTACTTTGCCACCACGGGTCCAGGAGGAGCTGTACCACCCAACCACCTGCCATTTGCCACACAGCGTCATGCAGCCAGCGAGTACCAGGCTGGACT
2. TCCTGGGACTGCACCAACTAAAAACGCTCAATCTGTACGACAATCAAATCTCATGCGTTATGCCCGGTTCCTTTGAGCATCTCAACTCTCTGACGTCGCTGAACCTCGCATCGAATCCAT
3. TGCGGCAAATGCAGAGTCAGCACCAAAAGGCTAAACCTAAACAAGTTCTGCAAAAGAGACTACGCAATAATGGCCAAAGTAATTGGTCGCGATACGAGCAGCGAGGCGGTCAGCAGGGAAGTTCAACGGCGCG
4. ATGCGTTGTTCACCAAAAACACTGTATTCAACATTGCTTTGCACAACAATCTCATTAAAATCAGCAAAACATTTCAAACGACGCAAATGCACACTAAAG
5. CTTGTGGTAGTCAGACCCTAGCTGCCAATCCCCCAGATGCCGACCAAAAAGGACCCGTCTTCCTCAAGGAACCCACCAACCGCATTGACTTCTCCAACTCCACGGGCGCAGAGATCGAGTGCAAGGCCAGCGGCAATCCCATGCCCGAGATTATTTGGATCAGGAGCGACGGTACCGCCG
6. TCATGCCGCGCTGCCTGGACGTGCTCTTTCGCACCATTTCAGATTACCAGGCCAAGAAGTTCGTCTTCAAGCCGGACAAGCTTAATGGCTTCGAGATCCTGTCCGAAGAGGATGCGCTTTTGGAGCGCCAACACGAGATGAACCAGCGTTTTG

1. FLYBASE / BLAST

BLAST (**basic local alignment search tool**) is an algorithm and program used in bioinformatics for comparing primary biological sequence information, such as the amino-acid sequences of proteins or the nucleotides of DNA and/or RNA sequences. A BLAST search enables a researcher to compare a subject protein or nucleotide sequence (*called a query*) with a library or database of sequences, and identify database sequences that resemble the query sequence above a certain threshold. For example, following the discovery of a previously unknown gene in the mouse, a scientist will typically perform a BLAST search of the human genome to see if humans carry a similar gene; BLAST will identify sequences in the human genome that resemble the mouse gene based on similarity of sequence [*source: https://en.wikipedia.org/wiki/BLAST_(biotechnology)*].

- Open flybase.org → Open "BLAST" (top left) → select: "Database / Annotated genes (NT)" → select: "Program / blastn: NT->NT" (default) → make sure "Drosophila melanogaster" is ticked under "Species (optional)" → paste the sequence into the "Sequence" box → press "BLAST" (*a window will show stating "Job queued" which will renew every 10 seconds until the search is performed*)
- High scores of ≥200 are indicated by red bar and are top in the table below
- Look at the table: genes are given before hyphen, the isoform behind hyphen: gene XXX was found at high score of 264.146 and very low E-value in the three isoforms C, B and A.

BLAST Hit Summary

☑	Description	Species	Score	E value
☑	XXX-RC	Dmel	264.146	3.24763e-70
☑	XXX-RB	Dmel	264.146	3.24763e-70
☑	XXX-RA	Dmel	264.146	3.24763e-70
☑	CG34357-RE	Dmel	38.1576	0.034741
☑	CG34357-RD	Dmel	38.1576	0.034741
☑	CG34357-RC	Dmel	38.1576	0.034741
☑	Inos-RA	Dmel	32.2105	2.14332
☑	CG13954-RD	Dmel	32.2105	2.14332
☑	CG13954-RC	Dmel	32.2105	2.14332
☑	CAP-RAF	Dmel	32.2105	2.14332

Expectation Values and Blast Scores

The **BLAST Score** indicates the quality of the best alignment between the query sequence and the found sequence (hit). The higher the score, the better the alignment. Scores are reduced by mismatches and gaps in the best alignment. Calculation of the score is complex, involving a substitution matrix, which is a table that assigns a score to each pair of residues aligned. The most widely used matrix for protein alignment is known as **BLOSUM62**.

The **expectation value E** of a hit tells whether the hit likely results from chance likeness between hit and query, or from common ancestry of hit and query (*if **E** is smaller than 10^{-100}, it is sometimes given as 0.0*). The expectation value is the number of hits you would expect to occur purely by chance if you searched for your sequence in a random genome the size of the human genome. **E** = 25 means that you could expect to find 25 matches in a genome of this size, purely by chance. So a hit with E = 25 is probably a chance match, and does not imply that the hit sequence shares common ancestry with your search sequence. Expectation values of around 0.1 may or may not be biologically significant (*other tests would be needed to decide*). But very small values of **E** mean that the hit *is* biologically significant, i.e. the correspondence between your search sequence and this hit must arise from common ancestry of the sequences, because the odds are are simply too low that the match could arise by chance. For example, **E** = 10^{-18} for a hit in the human genome means that you would expect only one chance match in one billion billion different genomes the same size of the human genome.

One place to find out more about BLAST searches and statistics is The BLAST Sequence Analysis Tool in the NCBI Handbook.

- type the gene name (before hyphen) into the field "Jump to Gene" (*very top right in the open window*) → the gene page opens:
  - "General Information & Genomic Location" → identifiers, locators
    - "Key Links" → explained below
    - "Genomic Maps" → explained below
  - "Function & Summaries" → crude descriptors and summaries of the gene's nature and function
  - "Gene Model and Products" → graphic display and tabular information of genomic area, different transcripts and protein isoforms
  - "Expression data" → detailed information on the spatiotemporal expression of the gene taken from big data and individual reports
  - "Alleles, Insertions, ..." → available genetic tools to manipulate the gene's function
    - "Classical and Insertion Alleles" → e.g. point mutations or small deletions from classical or transposon-induced screens, transposon excisions etc.
    - "Transgenic Constructs" → transgenic flies carrying constructs to express or knock-down the gene
    - "Aberrations" → chromosomal aberrations (large deletions, inversions, duplications, translocations)
  - "Phenotypes" → crude descriptions of mutant phenotypes linked to this gene
  - "Orthologs" → conserved genes in other species including humans
  - "Paralogs" → homologous (potentially redundant) genes in the *Drosophila melanogaster* genome
  - "Human Disease Associations" → potential links to human diseases (see also Section 5)
  - "Interactions" → predictions of physical interactors and genes that functionally (genetically) interact
  - "Pathways" → further big data predictions
  - "Genomic location..." → all locators and mapping data from different classical and modern approaches
  - "Stocks and Reagents"
    - "Stocks" → genetic tools readily available from stock centres
    - "Antibody Information" → potential published or commercially available gene-specific antibodies (e.g. monoclonal antibodies at Developmental Studies Hybridoma Bank)

- "External Crossreferences and Linkouts" → "Other crossreferences" → "Linkouts": summaries of current understanding of this gene and its functions
- "References" → list of all publications where this gene has been named or used sorted by different categories (*left-hand panel*) → you can export them for easier digest:
  - select "Research paper" in left-hand navigation bar
  - Click symbol top left and choose "Citation Manager (RIS)"
  - dialog box opens: "Save File" to your desktop (saves as "flybase.ris")
  - Open your reference data base (*here described for Endnote*):
    - File/New.. → name new library "XXX" and save on your computer
    - File/Import → "Import File:" choose "flybase.ris" from your desktop → set "Import Option:" to "Reference Manager (RIS)" → click "Import"

2. FLYBASE / JBROWSE

JBrowse was developed and is maintained by the Generic Model Organism Database (GMOD) consortium [*source: https://wiki.flybase.org/wiki/FlyBase:JBrowse*].

- "Genomic Location" → "Genomic Maps" → "JBrowse" button
- To gain a quick info overview, have the following items in left-hand navigation bar ticked:
  - Reference Genome
    - → RNA → are there vast differences in splice versions?
    - → CDS (coding sequence) → do the splice versions generate protein isoforms?
  - Aberrations → Deficiencies → long bars indicate areas which are deleted in those mutant fly stocks
  - Mutations... → Mutations... → FlyBase Mutations → when hovering over them or clicking them, more detailed information is given for each mutation (*far more information is available when visiting the "Alleles, Insertions, ..." section mentioned above*).
  - Further down: Protein Level Features → Protein Domains → Pfam / SMART → functional domains are shown (*see Appendix 5*) → right-click protein domains shown in the schematic view and select "View details" → positional information is provided that allows you map the domain into the protein (*see summary at end of Appendix 5*)
  - Expression
    - → Developmental stages, Tissues / CNS and adult head, Knoblich → L3 CNS neurons
    - → Fly Atlas transcriptomes → Nervous system
    - the wavy landscapes show how strong each exon is expressed in the chosen tissue, indicating whether the gene is expressed, its strength of expression and whether there is alternative splicing.

If you want to see where within the gene your original sequence localises use the "**BLAST**" function (top left) again but select in "Database" the "**Genome Assembly (NT)**" option. Upon BLAST, go to the top scoring sequence (*click red bar*) and click "**JBrowse**" → the yellow vertical bar indicates the query sequence relative to the gene (*horizontal bar at the top*). Click the symbol to get an overview of the whole region. To match against the position in the encoded protein, you might have to translate your NT sequence into aa sequence, e.g. using the **EXPASY TRANSLATE TOOL**: https://web.expasy.org/translate. Consider that there are 3 reading frames in each direction. To find the actual protein sequence take sequences that are uninterrupted and BLAST them in FlyBase using "Database / Annotated proteins (AA)". One of them should give you a perfect match.

3. ENSEMBL.ORG

Ensembl is a genome browser acting as a single point of access to annotated genomes, including information about genes, transcripts and further annotations that can be retrieved at the genome, gene and protein level (*e.g. protein domains, genetic variation, homology, syntenic regions and regulatory elements*). Ensembl imports genome sequences from consortia which keeps us consistent with many other bioinformatics projects. Each species in Ensembl has its own home page, where you can find out who provided the genome sequence and which version of the genome assembly is represented. For an example, see the *Drosophila* home page [*source: https://tinyurl.com/2fvjkxyv*].

- Open ensembl.org → In "Search" select "Drosophila melanogaster" from drop-down menu and enter the name of gene "XXX" in the search box.
- Select the hit that represents the "Gene" data set
- Click "Show transcript table" and select the largest isoform (*compare in schematic view below*) and BLAST either their transcript or protein sequence:
- BLAST transcript sequence:
  - click "**Transcript ID**" → select "**cDNA**" from left-hand navigation bar → click "BLAST this sequence"
  - The "BLAST/BLAT" page opens (*can also be reached from very top left tab next to* e!Ensembl): the sequence data box shows the selected sequence (the chevron at the top is for you to make notes (*e.g. type the gene name and Dm: ">XXX-Dm" – this will show in the query results; see Appendix 1*)
  - tick "DNA"
  - "Search against": click "Change species" → select "Primates" tab → tick "*Human*" → "Apply"
  - BLAST against DNA database / Genomic sequence"; don't remove the "Drosophila" choice.
  - When results show, click "View results" → both Jobs can be selected from the left-hand navigation bar → likely the only hit you have where there is an alignment over a long stretch ("Length") is the actual query gene in *Drosophila* because conservation at the DNA level is low because of the redundant genetic code.
- BLAST protein sequence:
  - Return to the original e!Ensemblene page for gene XXX, this time clicking "**Protein**" in the table and then "**Protein**" in the left-hand navigation bar → the aa sequence shows indicating different exons by alternating black and blue colour (*App. 2*).
  - "BLAST this sequence" against *Drosophila* and Human as above, but now selecting "Protein" and "Protein database" → you likely find high scoring hits in both jobs reflecting potential paralogues in the fly genome and orthologues in the human genome (*App. 3*).

If you want to use genomic or protein sequence for other purposes, you can **copy them out** into Word (or equivalent) either from the "Exon" or "Protein" view (*keeping exon colour code but also all other formatting items especially for NT*) or from the BLAST window (*stripped of all format except single spaces or paragraph marks behind each line*). Spaces/Paragraph marks can be easily removed using the "Replace" function in Word: select the entire sequence except the chevron line, then press "Ctrl H" → to replace space: type a space into the "Find what" box then press "Replace All" → to replace paragraph marks: select "More>> / Special / Paragraph Mark" then press "Replace All". Note, it is advised to use **Courier** font which is designed so that all numbers, letters and symbols have the same length and fall into column.

- Choose **one high-scoring human hit** from the table:
  - Get a feel for the hit: **click "[Alignment]"** behind the hit
    - → note that **start and end of the aligning fragments** are provided for subject and query sequence in both the table and in the info given above the alignment (*all yellow highlighted in Appendix 3*)
    - → the **aligned area** can be seen in black bold mapped into the query sequence when selecting "Ticket / Job / **Query Sequence**" in the left-hand panel (*Appendix 4*).
  - In the **table** generated by the BLAST, click the gene name in "**Overlapping Gene(s)**" → a table opens → click "**Protein**" in the longest isoform → select "**Protein summary**" in the left-hand bar → are there similar functional domains shown in the schematic representation (*Appendix 6*) as you saw in the *Drosophila* gene (*App. 5*)?

4. UNIPROT / ALIGN

- The alignment in Appendix 3 indicated 38.67% identity, meaning that the 319 aa linked by a vertical bar make up about 1/3 of the aligned fragment of 825 aa. However, it does not indicate the similarity, i.e. incorporating conservative exchanges between residues that have comparable properties and are more likely to maintain similar folding and function.
- To find out, use direct alignment, e.g. via **UniProt / Align** (https://www.uniprot.org/align): paste the two protein sequences of the aligned region each starting with a chevron line (*Appendix 7*) into the box → Click "Run alignment" → the alignment shows
- Choose the “TextOutput” tab → identical (*) and similar/highly similar (. / :) residues are indicated. Identical residues are provided under the “Percentage Identity Matrix” tab, similarity values seem no longer to be provided.

5. OTHER TOOLS AND DATABASES

- The functional domains in the first example do not include transmembrane domains, although the presence of FNIII and Ig domains suggests this protein to be a membrane receptor. Other tools can be used, such as TMpred (https://embnet.vital-it.ch/software/TMPRED_form.html)
  - paste the protein sequences from App. 2 and App. 6 singly into the query box.
  - Both show more than one suggested TM domain, with the one around 900 being the likely domain, since it is closest to the FNIII domain, as would be typical of many IgCAM receptors (Appendix 9).
- In FlyBase, human disease models were explained. Another resource to gain insight into the disease relevance is OMIM (Online Mendelian Inheritance in Man):
  - Open OMIM (https://www.ncbi.nlm.nih.gov/omim) and type "robo" into the query box
  - Scroll down until you find Robo1, Robo2 and Robo3 which all featured in the BLAST query (Appendix 3) → browse them one by one
  - In cases where a box is shown, clearly documented human disease genes were found → click "Phenotype MIM number" to learn more about the disease (Appendix 8).
- A rich resource informing about fly genes is "Interactive Fly" (http://www.sdbonline.org/sites/fly/aimain/1aahome.htm) → browse what's there about Robo.

Appendix 1

robo1:
http://www.ensembl.org/Drosophila_melanogaster/Gene/Summary?db=core;g=FBgn0005631;r=2R:22692524-22700970

Name	Transcript ID	bp	Protein	Biotype	UniProt Match	Flags
robo1-RC	FBtr0343013	6698	1395aa	Protein coding	Q9W213	APPRIS P1
robo1-RA	FBtr0071922	6392	1395aa	Protein coding	C4IXX6 O44924 Q058V6 Q9W213	APPRIS P1

```
>robo1-RC Dm FBtr0343013
CGTTCGGGAAATGTGGAAAATGTGCGCGAGGAGAGATGATGGAGAAAACAACCCCGAAAAGCGAGCAATCGTCGAATAAAT
ACCAACAAAAACTTACACTGCGCGGCTGATTAATAAAATCCAAAAAATAAAAAGGAAAATAAAAACCAAATCAGCGAACGA
TACAAAAAAAGCGTTAGAAAGAAAGTGTGAGAACGTGAAATCAAAACTCCGAATCAGAATTTGGATTGTAAATTTTCGGGT
CGGTGCGACTGAGACTGAGCGATAAACACAAAGGCAGATAAATACGGAAATAAAATTAATGGAAAACCAGCATAGACACAT
GTGTGTGGGTGTGTGAGTGTCCAAGTGTGCGTGAGTGTGTGCGTGTGTGTGTGCTTATCGCATGGAAATGCAAGCGACCCA
AAAACAACGCCAAAATGATCGCCGCTGCATTGGGTCATCTGTAGAGTCGTCATTGCAACAGCACCAACACCATAAAATGCA
TCCCATGCATCCCGAAAACCACGCCATCGCCCGGAGCACGAGCACCACTAATAACCCATCTCGCAGTCGGAGCAGCAGGAT
GTGGCTCCTGCCCGCCTGGCTGCTCCTCGTCCTGGTGGCCAGCAATGGCCTGCCAGCAGTCAGAGGCCAGTACCAATCGCC
ACGTATCATCGAGCATCCCACGGATCTGGTCGTTAAGAAGAATGAACCCGCCACGCTCAACTGCAAAGTGGAGGGCAAGCC
GGAACCCACCATTGAGTGGTTTAAGGATGGCGAACCCGTCAGCACCAACGAAAAGAAATCGCACCGCGTCCAGTTCAAGGA
CGGCGCCCTCTTCTTCTACAGGACAATGCAGGGCAAGAAGGAACAGGACGGCGGAGAGTACTGGTGCGTGGCCAAGAACCG
AGTGGGCCAGGCCGTTAGTCGCCATGCCTCGCTCCAGATAGCTGTTTTGCGCGACGATTTTCGCGTGGAGCCCAAAGACAC
GCGAGTGGCCAAAGGCGAGACGGCTCTGCTGGAGTGTGGGCCGCCCAAAGGCATTCCAGAGCCAACGCTGATTTGGATAAA
GGACGGCGTTCCCTTGGACGACCTGAAAGCCATGTCGTTTGGCGCCAGCTCCCGCGTTCGAATTGTGGACGGTGGCAACCT
GCTGATCAGCAATGTGGAGCCCATTGATGAGGGCAACTACAAGTGCATTGCCCAGAATCTGGTAGGCACCCGCGAGAGCAG
CTATGCCAAGCTGATTGTCCAGGTCAAACCATACTTTATGAAGGAGCCCAAGGATCAGGTGATGCTCTACGGCCAGACAGC
CACTTTCCACTGCTCAGTGGGCGGTGATCCGCCGCCGAAAGTGTTGTGGAAAAAGGAGGAGGGCAATATTCCGGTGTCCAG
AGCGCGAATCCTTCACGACGAGAAAAGTTTAGAGATATCCAACATAACGCCCACCGATGAGGGCACCTATGTCTGCGAGGC
ACACAACAATGTCGGTCAGATCAGCGCTAGGGCTTCTCTTATAGTCCATGCTCCGCCGAACTTTACGAAAAGACCCAGTAA
CAAGAAAGTGGGACTAAATGGTGTTGTCCAACTACCTTGCATGGCCTCCGGAAACCCACCGCCATCTGTTTTCTGGACCAA
GGAAGGAGTATCCACTCTCATGTTCCCAAATAGTTCGCACGGGAGGCAGCATGTGGCTGCCGATGGAACTCTGCAGATTAC
GGATGTGCGGCAGGAAGACGAAGGCTACTATGTGTGTTCCGCTTTCAGTGTAGTCGATTCCTCTACAGTACGGGTTTTCCT
GCAAGTCAGCTCGTTAGACGAGCGTCCACCTCCGATTATTCAAATCGGACCTGCCAATCAAACACTGCCCAAGGGATCAGT
TGCTACTTTACCCTGTCGGGCCACCGGAAATCCCAGTCCCCGTATTAAGTGGTTCCACGATGGACATGCCGTACAAGCGGG
CAATCGATACAGCATCATCCAAGGAAGCTCACTGAGAGTCGATGACCTTCAACTTAGTGACTCTGGTACCTACACCTGCAC
TGCATCTGGCGAACGAGGAGAAACTTCCTGGGCTGCCACACTAACGGTGGAAAAACCCGGTTCTACATCTCTTCACCGGGC
AGCTGATCCTAGCACTTATCCTGCTCCTCCAGGAACACCTAAAGTCCTGAATGTCAGTCGCACCAGCATTAGTCTTCGTTG
GGCTAAAAGCCAAGAGAAACCCGGAGCTGTGGGCCCAATCATTGGATACACTGTAGAGTACTTCAGTCCGGATCTGCAAAC
TGGTTGGATTGTGGCTGCCCAACGAGTTGGCGACACTCAAGTCACTATCTCGGGTCTCACTCCTGGCACTTCGTATGTGTT
CCTAGTTAGAGCTGAGAATACTCAGGGTATTTCTGTGCCTTCCGGCTTATCAAATACTATTAAAACCATTGAGGCGGATTT
CGATGCAGCTTCTGCCAATGATTTGTCAGCAGCTCGAACTTTGCTGACGGGAAAGTCGGTGGAGCTAATAGATGCCTCGGC
TATCAATGCTAGTGCCGTAAGACTTGAGTGGATGCTCCACGTGAGCGCTGATGAGAAATATGTAGAGGGCCTGCGCATACA
CTATAAGGATGCCAGTGTACCATCCGCACAGTATCACTCGATCACTGTTATGGATGCCTCTGCAGAATCGTTTGTGGTGGG
AAACCTTAAGAAGTACACCAAGTATGAGTTCTTCCTAACACCCTTTTTTGAGACAATTGAAGGACAGCCCAGTAACTCCAA
GACAGCCCTCACCTATGAAGATGTTCCCTCCGCACCACCGGATAACATTCAGATTGGCATGTACAACCAAACAGCCGGTTG
GGTGCGTTGGACTCCGCCACCCTCCCAGCACCACAATGGCAATTTGTATGGCTACAAGATTGAGGTCAGCGCCGGTAACAC
CATGAAGGTGCTGGCCAATATGACTCTTAATGCTACCACCACATCTGTGCTCCTAAATAACCTAACCACCGGAGCTGTGTA
CAGCGTGAGGTTGAACTCCTTTACCAAGGCAGGAGATGGACCTTACTCCAAACCGATATCACTATTCATGGACCCCACCCA
TCATGTGCATCCGCCACGGGCACATCCAAGCGGCACCCATGATGGGCGACATGAGGGACAGGATCTCACGTATCATAACAA
TGGCAACATACCACCTGGCGACATTAATCCCACCACTCATAAAAAGACCACTGACTACCTATCTGGACCATGGCTAATGGT
GCTGGTCTGCATCGTTCTTCTAGTCCTGGTTATTTCGGCGGCTATTTCGATGGTCTACTTCAAGCGCAAGCATCAAATGAC
CAAGGAATTGGGTCACTTAAGTGTGGTCAGTGACAACGAAATAACCGCATTGAACATCAATAGCAAAGAGAGCCTTTGGAT
AGACCATCATCGTGGATGGCGAACTGCCGATACTGACAAAGACTCAGGATTAAGCGAATCGAAGCTACTATCCCACGTAAA
CAGCAGCCAATCCAACTACAATAACTCCGATGGAGGAACCGATTATGCAGAAGTTGACACCCGTAACCTTACCACCTTCTA
CAATTGTCGCAAGAGCCCCGATAATCCCACGCCGTACGCCACCACTATGATCATTGGTACCTCTTCCAGTGAGACCTGCAC
CAAGACAACATCTATAAGTGCCGATAAGGACTCGGGAACTCATTCGCCCTATTCTGACGCATTTGCCGGTCAGGTGCCAGC
GGTTCCTGTTGTCAAATCCAACTATCTTCAGTATCCGGTTGAACCGATCAACTGGTCAGAGTTTCTACCCCCGCCGCCAGA
```

ACACCCACCTCCGTCTTCTACCTATGGATACGCACAAGGATCTCCTGAATCTTCACGGAAGAGCTCCAAAAGCGCAGGTTC
CGGCATTTCTACAAATCAAAGCATTCTGAACGCCTCCATACACAGCAGCTCCTCGGGCGGCTTTTCAGCTTGGGGAGTTTC
GCCCCAATATGCTGTCGCCTGTCCACCGGAAAACGTTTATAGCAATCCGCTGTCGGCAGTGGCTGGCGGCACCCAGAACCG
CTATCAGATAACGCCAACAAACCAACATCCGCCACAGTTACCGGCCTACTTTGCCACCACGGGTCCAGGAGGAGCTGTACC
ACCCAACCACCTGCCATTTGCCACACAGCGTCATGCAGCCAGCGAGTACCAGGCTGGACTGAATGCAGCGCGATGTGCCCA
AAGCCGCGCCTGCAACAGCTGCGATGCCTTGGCCACACCCTCGCCCATGCAACCCCCACCGCCAGTTCCCGTACCCGAGGG
CTGGTACCAACCGGTGCATCCCAATAGCCACCCGATGCACCCGACCTCCTCCAACCACCAGATCTACCAGTGCTCCTCCGA
GTGCTCGGATCACTCGAGGAGCTCGCAGAGTCACAAGCGGCAGCTGCAGCTCGAGGAGCACGGCAGCAGTGCCAAACAACG
CGGAGGACACCACCGTCGACGAGCCCCGGTGGTGCAGCCGTGCATGGAGAGCGAGAACGAGAACATGCTGGCGGAGTACGA
GCAGCGCCAGTACACCAGCGATTGTTGCAATAGCTCCCGCGAGGGCGACACCTGCTCCTGCAGCGAGGGATCCTGTCTTTA
CGCCGAGGCGGGCGAGCCGGCGCCTCGTCAAATGACTGCTAAGAACACCTAAGGATGAAGTGTTTGAGACTCTCCGAATAG
GGGGATTAGGATTAAAGATAGCATCTAGTCTAACTCTTCATTCATATCGAACTGGAAATCAGTCTAGGAACTGAGTGAACT
CAAAGTTTCTGCATCAGTTATCGAATAGAATAATTGATGTCTGGAGTCCACTACAACATGCGTCTTCGAAAACTGATGCTA
GCAGATAAAGAGGAAACATGTTGCAGTGGAATCTTCGGTATTACACTCAGATACACAAAACCACACTTGTCACAATCATGG
GGGCATTGTGCAGCTTTAGCAATTACGAATTTTTCCATTATCACATCTACCAGCGTAGAGACTCATAGACAGGCACAGTGC
CCCTTCCCCATCCATGCACCACAGAAATGTTTCCCCAATGGGAAACGAGAACCTTTGTACTTAAGAGACGCGAGTGCTGTA
CGATAGATATTGCTTCATGAGCGCGAGGAGATAGTAATCACTACATAGCAGGGATCTGTGTCCCAGCAGAACTTCATTAAA
ATACGTACTCAAAGGAACTTATAGCAGATGATAAACCAGATCAGGCCATAGCAAGTACACACAAAGACGGTGGTTACTCAA
GACGAATGTAGTTAGAACCAAAGTACACCATTTGTACAGCATGTTTTGGGGGCTTTTGAATCGGTCCTAGTCAATAGGCTT
GATAGCGATACAATACGATACTGATTACGATACTTGTATGCAAACTCTTTCATATCTACGATCACAATCGAGCACTTGTCT
CCGACTCGAACTTGTGGGGAGTTGTTGAGTGGAGTTTGAAGGCGACAAGTGAAACGGCAGTTGCAATGTTTAATGGTTAGT
GCTTTTCTATAACAAGGAATAATATACTTCATTGACCCAGTGAATATATATATATATATATACATACTAAACGACACAC
CACACTACTACTTACATACATAGCATTTAGACGAGTTGAGGAGCTGCACACGGCCAGCCGAACCAACCCCAGTTGTCCGTT
CCGTTGTCCTGTCCCTGTGCAGATCCCTCCAAGACACAATGTTTTCTACAATCTTGCATTTTGAAAAGTCTAAAATGTACC
TGTGATATAACCCCATTATACATTATAGTATTTTAATAATTTTTATATTTATGGTTTCCTAATTTTGAGTCGCGTTTTTTG
TTCATTCTGTTCGATCGACGAGGAATGTTTGTTAGAAATCCCCACACAGACGGATTTTTTTTGGGCGAGTATATAGCATAA
TTTAGATCTTAGGCACGTATCACATTAGAAATGCAAAATGGATATATCATCAAATCAAATCGAATGGAATCGAGGAGGAGT
AGAGTAACGCTCAGATAGTTCGGCTCGCCGAAGAGGCCCAGACACTTTTCCACTAAACCCCGGATCGTGAGTTGAATTGGG
AAAAAAGGGGGCTCGCGAGAAGTGGTTGGGCATACGAGCTGTGGATATACGAAGTATCCAGGGGGGGACTAGAGTCAACGA
ATCGTTTAAACAATACTAAACACATATGCATTACACTTACCTATATAGAGATCATATAGCACCTAGAACCGCGCATGTATC
AAATCAAATCAAAATATGCTTTAGCATACACTCACACAAATCCACTTCCTTGCTAGCCGCCCAGATCGGCCGGCATGTACA
TATATATATTATATATATATACATAAAACATACTCCTAATCCCAGTATCGAAACCAACTTGCTTGAGTTGAACTGCCCTTC
CGTAGCCATATACATATAGAGAAAGCATAGTTGGGTCTGGGTTTTAATAATTAGCAACTAGTTTCTAGCGTTGTGTAACGG
GAGTATTTTAAAGTCATAGCGTAAACGAATTTTAAGAACGAAACGAATCCTTTAACGAAGCGGACATTTAATTTTTAGTGC
GTAATGCCATAAACGAACTTTTTCAAAAGCAAAAATAATAACTGACATTTACTTACGATATAAATACTATATCTAGATATA
TACATATAGCATTTAAAATTACATAAACAATAAAGGAACATTTTACATATATGTGG

Appendix 2

>robo1-RC Dm FBtr0343013 1395aa
MHPMHPENHAIARSTSTTNNPSRSRSSRMWLLPAWLLLVLVASNGLPAVRGQYQSPRIIEHPTDLVVKKNEPATLNCKVEG
KPEPTIEWFKDGEPVSTNEKKSHRVQFKDGALFFYRTMQGKKEQDGGEYWCVAKNRVGQAVSRHASLQIAVLRDDFRVEPK
DTRVAKGETALLECGPPKGIPEPTLIWIKDGVPLDDLKAMSFGASSRVRIVDGGNLLISNVEPIDEGNYKCIAQNLVGTRE
SSYAKLIVQVKPYFMKEPKDQVMLYGQTATFHCSVGGDPPPKVLWKKEEGNIPVSRARILHDEKSLEISNITPTDEGTYVC
EAHNNVGQISARASLIVHAPPNFTKRPSNKKVGLNGVVQLPCMASGNPPPSVFWTKEGVSTLMFPNSSHGRQHVAADGTLQ
ITDVRQEDEGYYVCSAFSVVDSSTVRVFLQVSSLDERPPPIIQIGPANQTLPKGSVATLPCRATGNPSPRIKWFHDGHAVQ
AGNRYSIIQGSSLRVDDLQLSDSGTYTCTASGERGETSWAATLTVEKPGSTSLHRAADPSTYPAPPGTPKVLNVSRTSISL
RWAKSQEKPGAVGPIIGYTVEYFSPDLQTGWIVAAQRVGDTQVTISGLTPGTSYVFLVRAENTQGISVPSGLSNTIKTIEA
DFDAASANDLSAARTLLTGKSVELIDASAINASAVRLEWMLHVSADEKYVEGLRIHYKDASVPSAQYHSITVMDASAESFV
VGNLKKYTKYEFFLTPFFETIEGQPSNSKTALTYEDVPSAPPDNIQIGMYNQTAGWVRWTPPPSQHHNGNLYGYKIEVSAG
NTMKVLANMTLNATTTSVLLNNLTTGAVYSVRLNSFTKAGDGPYSKPISLFMDPTHHVHPPRAHPSGTHDGRHEGQDLTYH
NNGNIPPGDINPTTHKKTTDYLSGPWLMVLVCIVLLVLVISAAISMVYFKRKHQMTKELGHLSVVSDNEITALNINSKESL
WIDHHRGWRTADTDKDSGLSESKLLSHVNSSQSNYNNSDGGTDYAEVDTRNLTTFYNCRKSPDNPTPYATTMIIGTSSSET
CTKTTSISADKDSGTHSPYSDAFAGQVPAVPVVKSNYLQYPVEPINWSEFLPPPPEHPPPSSTYGYAQGSPESSRKSSKSA
GSGISTNQSILNASIHSSSSGGFSAWGVSPQYAVACPPENVYSNPLSAVAGGTQNRYQITPTNQHPPQLPAYFATTGPGGA
VPPNHLPFATQRHAASEYQAGLNAARCAQSRACNSCDALATPSPMQPPPPVPVPEGWYQPVHPNSHPMHPTSSNHQIYQCS
SECSDHSRSSQSHKRQLQLEEHGSSAKQRGGHHRRRAPVVQPCMESENENMLAEYEQRQYTSDCCNSSREGDTCSCSEGSC
LYAEAGEPAPRQMTAKNT

Appendix 3

BLAST/BLAT Alignment

Overlapping Gene(s)	Query name	Query start	Query end	Length	Score	E-val	%ID
IGSF10, MED12L	FBpp0309765	75	523	480 [Sequence]	150	7.78e-33	28.542 [Alignment]
IGSF10, MED12L	FBpp0309765	55	528	507 [Sequence]	146	9.92e-32	28.205 [Alignment]
IGSF10	FBpp0309765	108	567	479 [Sequence]	137	5.50e-29	27.975 [Alignment]
IGSF10, MED12L	FBpp0309765	47	436	405 [Sequence]	136	1.55e-28	28.889 [Alignment]
MXRA5	FBpp0309765	108	531	449 [Sequence]	135	2.32e-28	27.617 [Alignment]
MXRA5	FBpp0309765	208	534	350 [Sequence]	126	1.45e-25	28.857 [Alignment]
MXRA5Y	FBpp0309765	49	516	509 [Sequence]	125	4.42e-25	27.701 [Alignment]
MXRA5Y	FBpp0309765	113	531	445 [Sequence]	124	6.50e-25	27.416 [Alignment]
MXRA5	FBpp0309765	75	521	482 [Sequence]	123	1.14e-24	27.178 [Alignment]
ROBO2	FBpp0309765	49	151	106 [Sequence]	121	4.07e-24	52.830 [Alignment]
ROBO1	FBpp0309765	49	151	106 [Sequence]	119	2.30e-23	53.774 [Alignment]

BLAST/BLAT type	BLASTP
Query location	robo1 56 to 863 (+) → *Drosophila* robo1 aa positions
Database location	ENSP00000420321 68 to 872 (+) → human Robo1 aa positions
Genomic location	3 78651928 to 78938898 (-)
Alignment score	544
E-value	1e-165
Alignment length	825
Percentage identity	38.67
Markup	loaded

```
  1 PRIIEHPTDLVVKKNEPATLNCKVEGKPEPTIEWFKDGEPVSTNEK--KSHRVQFKDGAL     60
  1 ||| ||| || | | |||||||| || | ||||| | || | |      |||     | |     60
  1 PRIVEHPSDLIVSKGEPATLNCKAEGRPTPTIEWYKGGERVETDKDDPRSHRMLLPSGSL     60

 61 FFYRTMQGKKEQ-DGGEYWCVAKNRVGQAVSRHASLQIAVLRDDFRVEPKDTRVAKGETA    120
 61 || |   | |   | | | ||| |  | |||  |||  | ||||||  | |  || || |    120
 61 FFLRIVHGRKSRPDEGVYVCVARNYLGEAVSHNASLEVAILRDDFRQNPSDVMVAVGEPA    120

121 LLECGPPKGIPEPTLIWIKDGVPLDDLKAMSFGASSRVRIVDGGNLLISNVEPIDEGNYK    180
121   || || | ||||  | ||| ||||          |  |  || | |      | | |     180
121 VMECQPPRGHPEPTISWKKDGSPLDD-------KDERITI-RGGKLMITYTRKSDAGKYV    180

181 CIAQNLVGTRESSYAKLIVQVKPYFMKEPKDQVMLYGQTATFHCSVGGDPPPKVLWKKEE    240
181 |   | || |||  | | |   | | | |          | | |   ||| | | | |      240
181 CVGTNMVGERESEVAELTVLERPSFVKRPSNLAVTVDDSAEFKCEARGDPVPTVRWRKDD    240

241 GNIPVSRARILHDEKSLEISNITPTDEGTYVCEAHNNVGQISARASLIVHAPPNFTKRPS    300
241 |  | ||  |  |   | |   |  | | | | | | ||   | | | |  || |   |     300
241 GELPKSRYEI-RDDHTLKIRKVTAGDMGSYTCVAENMVGKAEASATLTVQEPPHFVVKPR    300

301 NKKVGLNGVVQLPCMASGNPPPSVFWTKEGVSTLMF---PNSSHGRQHVAADGTLQITDV    360
301    | |   |   | | ||| |  ||  ||   | |   |  |  |  |   | | || |    360
301 DQVVALGRTVTFQCEATGNPQPAIFWRREGSQNLLFSYQPPQSSSRFSVSQTGDLTITNV    360

361 RQEDEGYYVCSAFSVVDSSTVRVFLQVSS-LDERPPPIIQIGPANQTLPKGSVATLPCRA    420
361    | ||| |    |  |      | |      |||| |  || |||        | | |    420
361 QRSDVGYYICQTLNVAGSIITKAYLEVTDVIADRPPPVIRQGPVNQTVAVDGTFVLSCVA    420

421 TGNPSPRIKWFHDGHAVQAGN-RYSIIQGSSLRVDDLQLSDSGTYTCTASGERGETSWAA    480
421 || | | | |  ||  |     |        |      | | | ||| ||   ||  | |    480
421 TGSPVPTILWRKDGVLVSTQDSRIKQLENGVLQIRYAKLGDTGRYTCIASTPSGEATWSA    480

481 TLTVEKPG-STSLHRAADPSTYPAPPGTPKVLNVSRTSISLRWAKSQEKPGAVGPIIGYT    540
481    |   |      |  ||   |  |  | |  |||    | |   |           |     540
```

```
481 YIEVQEFGVPVQPPRPTDPNLIPSAPSKPEVTDVSRNTVTLSW---QPNLNSGATPTSYI    540

541 VEYFSPDLQTGWIVAAQRVGDTQVTISGLTPGTSYVFLVRAENTQGISVPSGLSNTIKTI    600
541  | ||      |   |  |      | || |   | ||||| |  ||| ||  |   ||     600
541 IEAFSHASGSSWQTVAENVKTETSAIKGLKPNAIYLFLVRAANAYGISDPSQISDPVKTQ    600

601 EADFDAASANDLSAARTLLTGKSVELIDASAINASAVRLEWMLHVSADEKYVEGLRIHYK    660
601           |       |      |        |     |   |     |  |  | |     660
601 DV-LPTSQGVDHKQVQRELGNAVLHLHNPTVLSSSSIEVHWT--VDQQSQYIQGYKILYR    660

661 DASVPSAQYHS------ITVMDASAESFVVGNLKKYTKYEFFLTPFFETIEGQPSNSKTA    720
661     ||   |         |      | |   | |   ||    |||    |  |  | |    720
661 ----PSGANHGESDWLVFEVRTPAKNSVVIPDLRKGVNYEIKARPFFNEFQGADSEIKFA    720

721 LTYEDVPSAPPDNIQIGMY--NQTAGWVRWTPPPSQHHNGNLYGYKIEVSAGNTMKVLAN    780
721  | |  |||||          | ||  | | |||    ||    ||     ||      |    780
721 KTLEEAPSAPPQGVTVSKNDGNGTAILVSWQPPPEDTQNGMVQEYKVWC-LGNETRYHIN    780

781 MTLNATTTSVLLNNLTTGAVYSVRLNSFTKAGDGPYSKPISLFMD                   825
781  |    | ||    |  |  |||     | || |  | |     |                   825
781 KTVDGSTFSVVIPFLVPGIRYSVEVAASTGAGSGVKSEPQFIQLD                   825
```

>human ROBO2-202

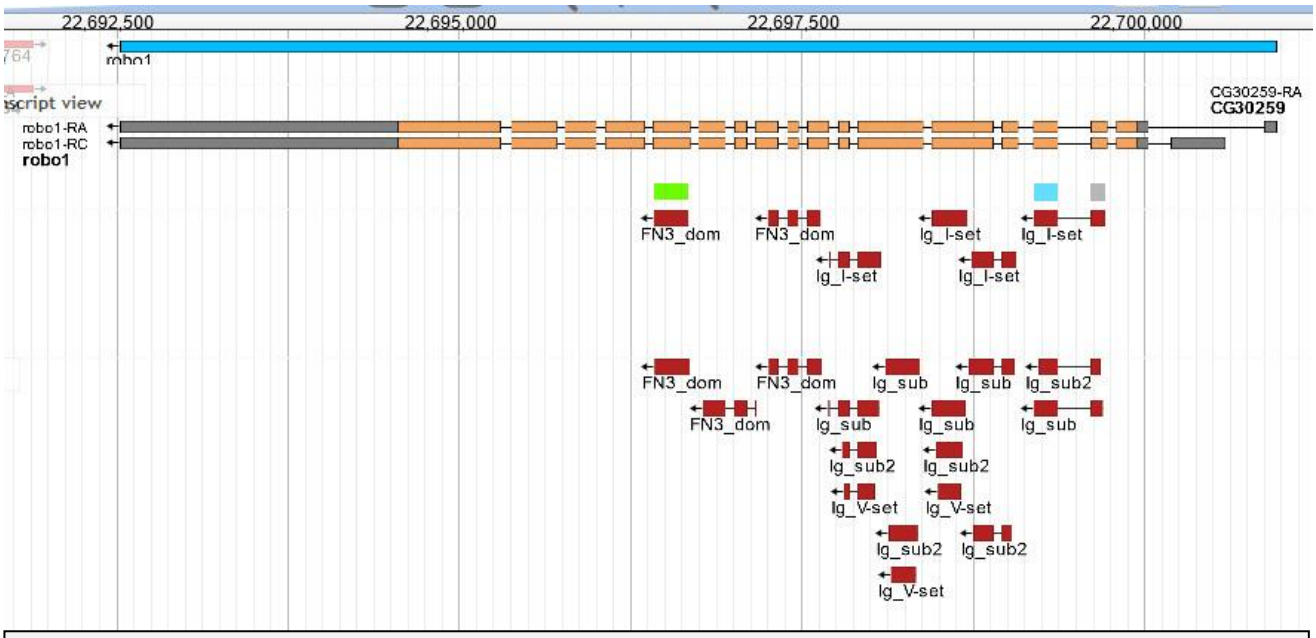
```
MSLLMFTQLLLCGFLYVRVDGSRLRQEDFPPRIVEHPSDVIVSKGEPTTLNCKAEGRPTPTIEWYKDGERVE
TDKDDPRSHRMLLPSGSLFFLRIVHGRRSKPDEGSYVCVARNYLGEAVSRNASLEVALLRDDFRQNPTDVVV
AAGEPAILECQPPRGHPEPTIYWKKDKVRIDDKEERISIRGGKLMISNTRKSDAGMYTCVGTNMVGERDSDP
AELTVFERPTFLRRPINQVVLEEEAVEFRCQVQGDPQPTVRWKKDDADLPRGRYDIKDDYTLRIKKTMSTDE
GTYMCIAENRVGKMEASATLTVRAPPQFVVRPRDQIVAQGRTVTFPCETKGNPQPAVFWQKEGSQNLLFPNQ
PQQPNSRCSVSPTGDLTITNIQRSDAGYYICQALTVAGSILAKAQLEVTDVLTDRPPPIILQGPANQTLAVD
GTALLKCKATGDPLPVISWLKEGFTFPGRDPRATIQEQGTLQIKNLRISDTGTYTCVATSSSGETSWSAVLD
VTESGATISKNYDLSDLPGPPSKPQVTDVTKNSVTLSWQPGTPGTLPASAYIIEAFSQSVSNSWQTVANHVK
TTLYTVRGLRPNTIYLFMVRAINPQGLSDPSPMSDPVRTQDISPPAQGVDHRQVQKELGDVLVRLHNPVVLT
PTTVQVTWTVDRQPQFIQGYRVMYRQTSGLQATSSWQNLDAKVPTERSAVLVNLKKGVTYEIKVRPYFNEFQ
GMDSESKTVRTTEEAPSAPPQSVTVLTVGSYNSTSISVSWDPPPPDHQNGIIQEYKIWCLGNETRFHINKTV
DAAIRSVIIGGLFPGIQYRVEVAASTSAGVGVKSEPQPIIIGRRNEVVITENNNSITEQITDVVKQPAFIAG
IGGACWVILMGFSIWLYWRRKKRKGLSNYAVTFQRGDGGLMSNGSRPGLLNAGDPSYPWLADSWPATSLPVN
NSNSGPNEIGNFGRGDVLPPVPGQGDKTATMLSDGAIYSSIDFTTKTSYNSSSQITQATPYATTQILHSNSI
HELAVDLPDPQWKSSIQQKTDLMGFGYSLPDQNKGNNGGKGGKKKKNKNSSKPQKNNGSTWANVPLPPPPVQ
PLPGTELEHYAVEQQENGYDSDSWCPPLPVQTYLHQGLEDELEEDDDRVPTPPVRGVASSPAISFGQQSTAT
LTPSPREEMQPMLQAHLDELTRAYQFDIAKQTWHIQSNNQPPQPPVPPLGYVSGALISDLETDVADDDADDE
EEALEIPRPLRALDQTPGSSMDNLDSSVTGKAFTSSQRPRPTSPFSTDSNTSAALSQSQRPRPTKKHKGGRM
DQQPALPHRREGMTDEEALVPYSKPSFPSPGGHSSSGTASSKGSTGPRKTEVLRAGHQRNASDLLDIGYMGS
NSQGQFTGEL
```

Appendix 4

```
>robo1:1:1395:1
  1 MHPMHPENHAIARSTSTTNNPSRSRSSRMWLLPAWLLLVLVASNGLPAVRGQYQSPRIIE    60
 61 HPTDLVVKKNEPATLNCKVEGKPEPTIEWFKDGEPVSTNEKKSHRVQFKDGALFFYRTMQ   120
121 GKKEQDGGEYWCVAKNRVGQAVSRHASLQIAVLRDDFRVEPKDTRVAKGETALLECGPPK   180
181 GIPEPTLIWIKDGVPLDDLKAMSFGASSRVRIVDGGNLLISNVEPIDEGNYKCIAQNLVG   240
241 TRESSYAKLIVQVKPYFMKEPKDQVMLYGQTATFHCSVGGDPPPKVLWKKEEGNIPVSRA   300
301 RILHDEKSLEISNITPTDEGTYVCEAHNNVGQISARASLIVHAPPNFTKRPSNKKVGLNG   360
361 VVQLPCMASGNPPPSVFWTKEGVSTLMFPNSSHGRQHVAADGTLQITDVRQEDEGYYVCS   420
421 AFSVVDSSTVRVFLQVSSLDERPPPIIQIGPANQTLPKGSVATLPCRATGNPSPRIKWFH   480
481 DGHAVQAGNRYSIIQGSSLRVDDLQLSDSGTYTCTASGERGETSWAATLTVEKPGSTSLH   540
541 RAADPSTYPAPPGTPKVLNVSRTSISLRWAKSQEKPGAVGPIIGYTVEYFSPDLQTGWIV   600
601 AAQRVGDTQVTISGLTPGTSYVFLVRAENTQGISVPSGLSNTIKTIEADFDAASANDLSA   660
661 ARTLLTGKSVELIDASAINASAVRLEWMLHVSADEKYVEGLRIHYKDASVPSAQYHSITV   720
721 MDASAESFVVGNLKKYTKYEFFLTPFFETIEGQPSNSKTALTYEDVPSAPPDNIQIGMYN   780
781 QTAGWVRWTPPPSQHHNGNLYGYKIEVSAGNTMKVLANMTLNATTTSVLLNNLTTGAVYS   840
```

```
 841 VRLNSFTKAGDGPYSKPISLFMDPTHHVHPPRAHPSGTHDGRHEGQDLTYHNNGNIPPGD    900
 901 INPTTHKKTTDYLSGPWLMVLVCIVLLVLVISAAISMVYFKRKHQMTKELGHLSVVSDNE    960
 961 ITALNINSKESLWIDHHRGWRTADTDKDSGLSESKLLSHVNSSQSNYNNSDGGTDYAEVD   1020
1021 TRNLTTFYNCRKSPDNPTPYATTMIIGTSSSETCTKTTSISADKDSGTHSPYSDAFAGQV   1080
1081 PAVPVVKSNYLQYPVEPINWSEFLPPPPEHPPPSSTYGYAQGSPESSRKSSKSAGSGIST   1140
1141 NQSILNASIHSSSSGGFSAWGVSPQYAVACPPENVYSNPLSAVAGGTQNRYQITPTNQHP   1200
1201 PQLPAYFATTGPGGAVPPNHLPFATQRHAASEYQAGLNAARCAQSRACNSCDALATPSPM   1260
1261 QPPPPVPVPEGWYQPVHPNSHPMHPTSSNHQIYQCSSECSDHSRSSQSHKRQLQLEEHGS   1320
1321 SAKQRGGHHRRRAPVVQPCMESENENMLAEYEQRQYTSDCCNSSREGDTCSCSEGSCLYA   1380
1381 EAGEPAPRQMTAKNT                                                1395
```

Appendix 5



Note the directionality of transcription of the *robo1* gene: it occurs on the minus strand (yellow highlighted in the domain descriptions below and indicated by small left-pointing arrows in the graph above). Once the mRNA is generated, this direction is no longer relevant and you can treat the sequence from start to end.

NameFN3_dom (info obtained upon right-click of green-highlighted domain in image above and selecting "view details")

Position2R: 22696425..22696682 (- strand)
Length 258 bp
Aa_position 770-855
DbxrefPfam: PF00041
INTERPRO: IPR003961
Domain_name_abbreviation FN3_dom
Id fic878592
Seq_id 2R
Source dmel_domain_PFAM

>2R 2R:22696425..22696682 (- strand) class=match_part length=258
CCACCGGATAACATTCAGATTGGCATGTACAACCAAACAGCCGGTTGGGTGCGTTGGACTCCGCCACCCTCCCAGCACC
ACAATGGCAATTTGTATGGCTACAAGATTGAGGTCAGCGCCGGTAACACCATGAAGGTGCTGGCCAATATGACTCTTAA
TGCTACCACCACATCTGTGCTCCTAAATAACCTAACCACCGGAGCTGTGTACAGCGTGAGGTTGAACTCCTTTACCAAGG
CAGGAGATGGACCTTACTCC

To try it out: paste this sequence into the **EXPASY TRANSLATE TOOL** (https://web.expasy.org/translate), identify the correct protein sequence (6 reading frames!) and find it in the full protein sequence.

NameIg_I-set (info obtained upon right-click of grey/blue-highlighted domain in image above and selecting "view details")

Position2R:22699613..22699720 (- strand)
Length 108 bp
Aa_position 56-91
DbxrefPfam: PF07679
INTERPRO: IPR013098
Domain_name_abbreviationIg_I-set
Id fic878768
Seq_id2R
Source dmel_domain_PFAM

>2R 2R:22699613..22699720 (- strand) class=match_part length=108
CCACGTATCATCGAGCATCCCACGGATCTGGTCGTTAAGAAGAATGAACCCGCCACGCTCAACTGCAAAGTGGAGGGC
AAGCCGGAACCCACCATTGAGTGGTTTAAG

Position2R:22699200..22699376 (- strand)
Length 177 bp
Attributes
Aa_position 92-150
DbxrefPfam: PF07679
INTERPRO: IPR013098
Domain_name_abbreviationIg_I-set
Id fic878751
Seq_id2R
Source dmel_domain_PFAM

>2R 2R:22699200..22699376 (- strand) class=match_part length=177
GATGGCGAACCCGTCAGCACCAACGAAAAGAAATCGCACCGCGTCCAGTTCAAGGACGGCGCCCTCTTCTTCTACAGGA
CAATGCAGGGCAAGAAGGAACAGGACGGCGGAGAGTACTGGTGCGTGGCCAAGAACCGAGTGGGCCAGGCCGTTAG
TCGCCATGCCTCGCTCCAGATA

Summary of results:
First (grey/blue) and last (green) functional protein domains relative to sequence and exons (alternating blue and black) and the area identified as being conserved with human Robo1 (bold)

```
1     MHPMHPENHA IARSTSTTNN PSRSRSSRMW LLPAWLLLVL VASNGLPAVR GQYQSPRIIE   60
61    HPTDLVVKKN EPATLNCKVE GKPEPTIEWF KDGEPVSTNE KKSHRVQFKD GALFFYRTMQ  120
121   GKKEQDGGEY WCVAKNRVGQ AVSRHASLQI AVLRDDFRVE PKDTRVAKGE TALLECGPPK  180
181   GIPEPTLIWI KDGVPLDDLK AMSFGASSRV RIVDGGNLLI SNVEPIDEGN YKCIAQNLVG
241   TRESSYAKLI VQVKPYFMKE PKDQVMLYGQ TATFHCSVGG DPPPKVLWKK EEGNIPVSRA
301   RILHDEKSLE ISNITPTDEG TYVCEAHNNV GQISARASLI VHAPPNFTKR PSNKKVGLNG
361   VVQLPCMASG NPPPSVFWTK EGVSTLMFPN SSHGRQHVAA DGTLQITDVR QEDEGYYVCS
421   AFSVVDSSTV RVFLQVSSLD ERPPPIIQIG PANQTLPKGS VATLPCRATG NPSPRIKWFH
481   DGHAVQAGNR YSIIQGSSLR VDDLQLSDSG TYTCTASGER GETSWAATLT VEKPGSTSLH
541   RAADPSTYPA PPGTPKVLNV SRTSISLRWA KSQEKPGAVG PIIGYTVEYF SPDLQTGWIV
601   AAQRVGDTQV TISGLTPGTS YVFLVRAENT QGISVPSGLS NTIKTIEADF DAASANDLSA
661   ARTLLTGKSV ELIDASAINA SAVRLEWMLH VSADEKYVEG LRIHYKDASV PSAQYHSITV
721   MDASAESFVV GNLKKYTKYE FFLTPFFETI EGQPSNSKTA LTYEDVPSAP PDNIQIGMYN  780
781   QTAGWVRWTP PPSQHHNGNL YGYKIEVSAG NTMKVLANMT LNATTTSVLL NNLTTGAVYS  840
841   VRLNSFTKAG DGPYSKPISL FMDPTHHVHP PRAHPSGTHD GRHEGQDLTY HNNGNIPPGD  900
901   INPTTHKKTT DYLSGPWLMV LVCIVLLVLV ISAAISMVYF KRKHQMTKEL GHLSVVSDNE
961   ITALNINSKE SLWIDHHRGW RTADTDKDSG LSESKLLSHV NSSQSNYNNS DGGTDYAEVD
1021  TRNLTTFYNC RKSPDNPTPY ATTMIIGTSS SETCTKTTSI SADKDSGTHS PYSDAFAGQV
1081  PAVPVVKSNY LQYPVEPINW SEFLPPPPEH PPPSSTYGYA QGSPESSRKS SKSAGSGIST
1141  NQSILNASIH SSSSGGFSAW GVSPQYAVAC PPENVYSNPL SAVAGGTQNR YQITPTNQHP
```

```
1201    PQLPAYFATT GPGGAVPPNH LPFATQRHAA SEYQAGLNAA RCAQSRACNS CDALATPSPM
1261    QPPPPVPVPE GWYQPVHPNS HPMHPTSSNH QIYQCSSECS DHSRSSQSHK RQLQLEEHGS
1321    SAKQRGGHHR RRAPVVQPCM ESENENMLAE YEQRQYTSDC CNSSREGDTC SCSEGSCLYA
1381    EAGEPAPRQM TAKNT
```

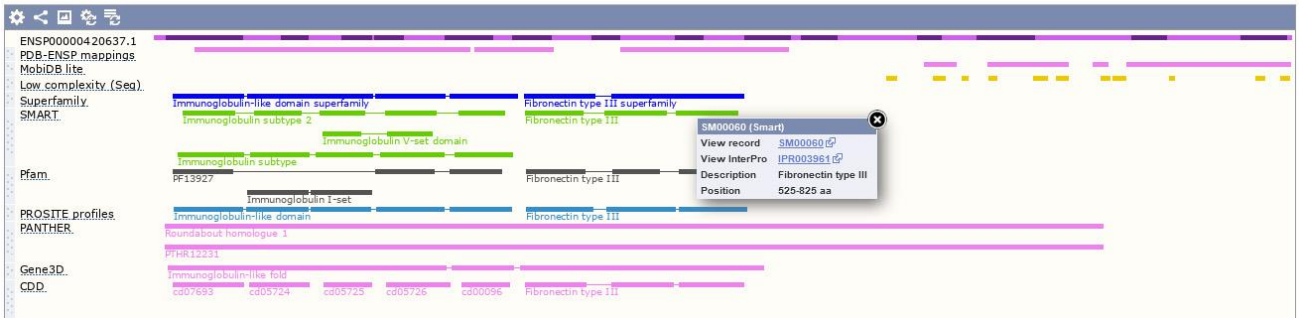
Appendix 6

Show/hide columns (1 hidden) Filter

Name	Transcript ID	bp	Protein	Biotype	CCDS	UniProt Match	RefSeq Match	Flags
ROBO1-202	ENST00000464233.6	6927	1651aa	Protein coding	CCDS54611	Q9Y6N7-1	NM_002941.4	TSL:5 GENCODE basic APPRIS ALT2 MANE Select v0.92
ROBO1-209	ENST00000495273.5	5600	1606aa	Protein coding	CCDS46872	Q9Y6N7-5	-	TSL:1 GENCODE basic APPRIS P4
ROBO1-204	ENST00000467549.5	4841	1551aa	Protein coding	CCDS54610	Q9Y6N7-6	-	TSL:1 GENCODE basic APPRIS ALT2
ROBO1-201	ENST00000436010.6	7501	1531aa	Protein coding	-	A0A0A0MSX4	-	TSL:5 GENCODE basic
ROBO1-212	ENST00000618833.4	7492	1606aa	Protein coding	-	A0A087WTM1	-	TSL:5 GENCODE basic APPRIS ALT2
ROBO1-213	ENST00000618846.4	7483	1525aa	Protein coding	-	A0A087WZ85	-	TSL:5 GENCODE basic
ROBO1-210	ENST00000495961.1	577	192aa	Protein coding	-	H7C4Y7	-	CDS 5' and 3' incomplete TSL:4
ROBO1-205	ENST00000472273.1	569	190aa	Protein coding	-	H7C545	-	CDS 5' and 3' incomplete TSL:4
ROBO1-208	ENST00000492990.1	444	41aa	Nonsense mediated decay	-	F8WEV8	-	TSL:3
ROBO1-203	ENST00000466906.1	550	No protein	Processed transcript	-	-	-	TSL:4
ROBO1-211	ENST00000498428.5	3566	No protein	Retained intron	-	-	-	TSL:2
ROBO1-206	ENST00000484514.1	3236	No protein	Retained intron	-	-	-	TSL:2
ROBO1-207	ENST00000488735.1	593	No protein	Retained intron	-	-	-	TSL:4

Protein summary

Protein domains for ENSP00000420637.1



```
>ENSP00000420321 ROBO1 Hs
MKWKHVPFLVMISLLSLSPNHLFLAQLIPDPEDVERGNDHGTPIPTSDNDDNSLGYTGSRLRQEDFPPRIVEHPSDLIVSK
GEPATLNCKAEGRPTPTIEWYKGGERVETDKDDPRSHRMLLPSGSLFFLRIVHGRKSRPDEGVYVCVARNYLGEAVSHNAS
LEVAILRDDFRQNPSDVMVAVGEPAVMECQPPRGHPEPTISWKKDGSPLDDKDERITIRGGKLMITYTRKSDAGKYVCVGT
NMVGERESEVAELTVLERPSFVKRPSNLAVTVDDSAEFKCEARGDPVPTVRWRKDDGELPKSRYEIRDDHTLKIRKVTAGD
MGSYTCVAENMVGKAEASATLTVQEPPHFVVKPRDQVVALGRTVTFQCEATGNPQPAIFWRREGSQNLLFSYQPPQSSSRF
SVSQTGDLTITNVQRSDVGYYICQTLNVAGSIITKAYLEVTDVIADRPPPVIRQGPVNQTVAVDGTFVLSCVATGSPVPTI
LWRKDGVLVSTQDSRIKQLENGVLQIRYAKLGDTGRYTCIASTPSGEATWSAYIEVQEFGVPVQPPRPTDPNLIPSAPSKP
EVTDVSRNTVTLSWQPNLNSGATPTSYIIEAFSHASGSSWQTVAENVKTETSAIKGLKPNAIYLFLVRAANAYGISDPSQI
SDPVKTQDVLPTSQGVDHKQVQRELGNAVLHLHNPTVLSSSSIEVHWTVDQQSQYIQGYKILYRPSGANHGESDWLVFEVR
TPAKNSVVIPDLRKGVNYEIKARPFFNEFQGADSEIKFAKTLEEAPSAPPQGVTVSKNDGNGTAILVSWQPPPEDTQNGMV
QEYKVWCLGNETRYHINKTVDGSTFSVVIPFLVPGIRYSVEVAASTGAGSGVKSEPQFIQLDAHGNPVSPEDQVSLAQQIS
DVVKQPAFIAGIGAACWIILMVFSIWLYRHRKKRNGLTSTYAGIRKVPSFTFTPTVTYQRGGEAVSSGGRPGLLNISEPAA
QPWLADTWPNTGNNHNDCSISCCTAGNGNSDSNLTTYSRPADCIANYNNQLDNKQTNLMLPESTVYGDVDLSNKINEMKTF
NSPNLKDGRFVNPSGQPTPYATTQLIQSNLSNNMNNGSGDSGEKHWKPLGQQKQEVAPVQYNIVEQNKLNKDYRANDTVPP
TIPYNQSYDQNTGGSYNSSDRGSSTSGSQGHKKGARTPKVPKQGGMNWADLLPPPPAHPPPHSNSEEYNISVDESYDQEMP
CPVPPARMYLQQDELEEEEDERGPTPPVRGAASSPAAVSYSHQSTATLTPSPQEELQPMLQDCPEETGHMQHQPDRRRQPV
SPPPPPRPISPPHTYGYISGPLVSDMDTDAPEEEEDEADMEVAKMQTRRLLLRGLEQTPASSVGDLESSVTGSMINGWGSA
SEEDNISSGRSSVSSSDGSFFTDADFAQAVAAAAEYAGLKVARRQMQDAAGRRHFHASQCPRPTSPVSTDSNMSAAVMQKT
RPAKKLKHQPGHLRRETYTDDLPPPPVPPPAIKSPTAQSKTQLEVRPVVVPKLPSMDARTDRSSDRKGSSYKGREVLDGRQ
VVDMRTNPGDPREAQEQQNDGKGRGNKAAKRDLPPAKTHLIQEDILPYCRPTFPTSNNPRDPSSSSSMSSRGSGSRQREQA
NVGRRNIAEMQVLGGYERGEDNNEELEETES
```

Appendix 7

```
>robo1-RC Dm 56 to 863
PRIIEHPTDLVVKKNEPATLNCKVEGKPEPTIEWFKDGEPVSTNEKKSHRVQFKDGALFFYRTMQGKKEQDGGEYWCVAKN
RVGQAVSRHASLQIAVLRDDFRVEPKDTRVAKGETALLECGPPKGIPEPTLIWIKDGVPLDDLKAMSFGASSRVRIVDGGNLLI
SNVEPIDEGNYKCIAQNLVGTRESSYAKLIVQVKPYFMKEPKDQVMLYGQTATFHCSVGGDPPPKVLWKKEEGNIPVSRARIL
HDEKSLEISNITPTDEGTYVCEAHNNVGQISARASLIVHAPPNFTKRPSNKKVGLNGVVQLPCMASGNPPPSVFWTKEGVSTL
MFPNSSHGRQHVAADGTLQITDVRQEDEGYYVCSAFSVVDSSTVRVFLQVSSLDERPPPIIQIGPANQTLPKGSVATLPCRAT
GNPSPRIKWFHDGHAVQAGNRYSIIQGSSLRVDDLQLSDSGTYTCTASGERGETSWAATLTVEKPGSTSLHRAADPSTYPAP
PGTPKVLNVSRTSISLRWAKSQEKPGAVGPIIGYTVEYFSPDLQTGWIVAAQRVGDTQVTISGLTPGTSYVFLVRAENTQGISV
PSGLSNTIKTIEADFDAASANDLSAARTLLTGKSVELIDASAINASAVRLEWMLHVSADEKYVEGLRIHYKDASVPSAQYHSITV
MDASAESFVVGNLKKYTKYEFFLTPFFETIEGQPSNSKTALTYEDVPSAPPDNIQIGMYNQTAGWVRWTPPPSQHHNGNLY
GYKIEVSAGNTMKVLANMTLNATTTSVLLNNLTTGAVYSVRLNSFTKAGDGPYSKPISLFMD

>ROBO1 Hs 68 to 872
PRIVEHPSDLIVSKGEPATLNCKAEGRPTPTIEWYKGGERVETDKDDPRSHRMLLPSGSLFFLRIVHGRKSRPDEGVYVCVARN
YLGEAVSHNASLEVAILRDDFRQNPSDVMVAVGEPAVMECQPPRGHPEPTISWKKDGSPLDDKDERITIRGGKLMITYTRKS
DAGKYVCVGTNMVGERESEVAELTVLERPSFVKRPSNLAVTVDDSAEFKCEARGDPVPTVRWRKDDGELPKSRYEIRDDHTL
KIRKVTAGDMGSYTCVAENMVGKAEASATLTVQEPPHFVVKPRDQVVALGRTVTFQCEATGNPQPAIFWRREGSQNLLFSY
QPPQSSSRFSVSQTGDLTITNVQRSDVGYYICQTLNVAGSIITKAYLEVTDVIADRPPPVIRQGPVNQTVAVDGTFVLSCVATG
SPVPTILWRKDGVLVSTQDSRIKQLENGVLQIRYAKLGDTGRYTCIASTPSGEATWSAYIEVQEFGVPVQPPRPTDPNLIPSAP
SKPEVTDVSRNTVTLSWQPNLNSGATPTSYIIEAFSHASGSSWQTVAENVKTETSAIKGLKPNAIYLFLVRAANAYGISDPSQIS
DPVKTQDVLPTSQGVDHKQVQRELGNAVLHLHNPTVLSSSSIEVHWTVDQQSQYIQGYKILYRPSGANHGESDWLVFEVRT
PAKNSVVIPDLRKGVNYEIKARPFFNEFQGADSEIKFAKTLEEAPSAPPQGVTVSKNDGNGTAILVSWQPPPEDTQNGMVQE
YKVWCLGNETRYHINKTVDGSTFSVVIPFLVPGIRYSVEVAASTGAGSGVKSEPQFIQLD
```

Job status: COMPLETED

```
robo1-RC      1   PRIIEHPTDLVVKKNEPATLNCKVEGKPEPTIEWFKDGEPVSTN--EKKSHRVQFKDGAL    58
ROBO1         1   PRIVEHPSDLIVSKGEPATLNCKAEGRPTPTIEWYKGGERVEIDKDDPRSHRMLLPSGSL    60
                  ***:***:**:*.*.********.**:* *****:*.** *.*:  : :***: : .*:*

robo1-RC     59   FFYRTMQGKKE-QDGGEYWCVAKNRVGQAVSRHASLQIAVLRDDFRVEPKDIRVAKGETA   117
ROBO1        61   FFLRIVHGRKSRPDEGVYVCVARNYLGEAVSHNASLEVAILRDDFRQNPSDVMVAVGEPA   120
                  ** * ::*:*.  * * * ***:* :*:***::***::*:****** :*.*. ** ** *

robo1-RC    118   LLECGPPKGIPEPTLIWIKDGVPLDDLKAMSFGASSRVRIVDGGNLLISNVEPIDEGNYK   177
ROBO1       121   VMECQPPRGHPEPTISWKKDGSPLDDKDE--------RITIRGGKLMITYTRKSDAGKYV   172
                  ::** **:* ****: * *** **** .            : **:*:*: ..  * *:*

robo1-RC    178   CIAQNLVGIRESSYAKLIVQVKPYFMKEPKDQVMLYGQTATFHCSVGGDPPPKVLWKKEE   237
ROBO1       173   CVGTNMVGERESEVAELTVLERPSFVKRPSNLAVTVDDSAEFKCEARGDPVPTVRWRKDD   232
                  *:. *:** ***. *:* *  :* *:*.*.: .:  .::* *:*.. *** *.* *:*::

robo1-RC    238   GNIPVSRARILHDEKSLEISNITPTDEGTYVCEAHNNVGQISARASLIVHAPPNFTKRPS   297
ROBO1       233   GELPKSRYEI-RDDHTLKIRKVTAGDMGSYTCVAENMVGKAEASATLTVQEPPHFVVKPR   291
                  *::* ** .* :*:::*:* ::*  * *:*.* *.* **: .* *:* *: **:*. :*

robo1-RC    298   NKKVGLNGVVQLPCMASGNPPPSVFWIKEGVSILMFPN---SSHGRQHVAADGTLQITDV   354
ROBO1       292   DQVVALGRTVTFQCEATGNPQPAIFWRREGSQNLLFSYQPPQSSSRFSVSQTGDLTITNV   351
                  :: *.*. .* : * *:*** *::** :** ..*:*     .* .*  *:  * * **:*

robo1-RC    355   RQEDEGYYVCSAFSVVDSSTVRVFLQVS-SLDERPPPIIQIGPANQTLPKGSVATLPCRA   413
ROBO1       352   QRSDVGYYICQTLNVAGSIITKAYLEVIDVIADRPPPVIRQGPVNQTVAVDGTFVLSCVA   411
                  ::.* ***:*.::.*..*  .:.:*:*:  : :****:*: **.***:  ... .* * *

robo1-RC    414   TGNPSPRIKWFHDGHAVQAG-NRYSIIQGSSLRVDDLQLSDSGTYTCTASGERGETSWAA   472
ROBO1       412   TGSPVPTILWRKDGVLVSTQDSRIKQLENGVLQIRYAKLGDTGRYTCIASTPSGEATWSA   471
                  **.* * * * :**  *.:  .* . ::.. *::   :*.*:* *** **   **::*:*

robo1-RC    473   TLTVEKPGST-SLHRAADPSTYPAPPGTPKVLNVSRTSISLRWAKSQEKPGAVGPIIGYT   531
ROBO1       472   YIEVQEFGVPVQPPRPTDPNLIPSAPSKPEVTDVSRNTVTLSWQPNLNSG---ATPTSYI   528
                   : *:: *   .  * :**.  *: *..*:* :***.:::* *  . :.    .   .*

robo1-RC    532   VEYFSPDLQTGWIVAAQRVGDTQVTISGLTPGTSYVFLVRAENTQGISVPSGLSNTIKTI   591
ROBO1       529   IEAFSHASGSSWQTVAENVKTETSAIKGLKPNAIYLFLVRAANAYGISDPSQISDPVKIQ   588
                  :* **    :.* ..*:.*     :*.**.*.: *:***** *: *** ** :*: :**

robo1-RC    592   EADFDAASANDLSAARTLLTGKSVELIDASAINASAVRLEWMLHVSADEKYVEGLRIHYK   651
ROBO1       589   DVLPTSQGVDH-KQVQRELGNAVLHLHNPTVLSSSSIEVHWTV--DQQSQYIQGYKILYR   645
                  :.   : ..:. . .:  * .  :.* : :.:.:*::.:.* :  . :.:*::* :* *:

robo1-RC    652   DASVPS--AQYHSITVMDASAESFVVGNLKKYTKYEFFLTPFFETIEGQPSNSKTALTYE   709
ROBO1       646   PSGANHGESDWLVFEVRTPAKNSVVIPDLRKGVNYEIKARPFFNEFQGADSEIKFAKTLE   705
                   :..    :::  : *   : :*.*: :*:* .:**:   ***: ::*  *: * * * *

robo1-RC    710   DVPSAPPDNIQIGMY--NQTAGWVRWIPPPSQHHNGNLYGYKIEVSAGNTMKVLANMILN   767
ROBO1       706   EAPSAPPQGVTVSKNDGNGTAILVSWQPPPEDTQNGMVQEYKVWCL-GNETRYHINKIVD   764
                  :.*****:.: :.    * **  * * ***.: :** :  **:    **  :   * *::

robo1-RC    768   ATITSVLLNNLTTGAVYSVRLNSFTKAGDGPYSKPISLFMD                      808
ROBO1       765   GSTFSVVIPFLVPGIRYSVEVAASTGAGSGVKSEPQFIQLD                      805
                  .:* **::  *. *  ***.: : * **.*  *:*  : :*
```

310 identical positions (38% identity) + 264 similar positions (71% similarity)

Appendix 8

Alternative titles; symbols

ROUNDABOUT, DROSOPHILA, HOMOLOG OF, 3
RB-INHIBITING GENE 1; RBIG1; RIG1

HGNC Approved Gene Symbol: ROBO3

Cytogenetic location: 11q24.2 ***Genomic coordinates (GRCh38): 11:124,865,431-124,881,470*** (from NCBI)

Gene-Phenotype Relationships

Location	Phenotype	Phenotype MIM number	Inheritance	Phenotype mapping key
11q24.2	Gaze palsy, familial horizontal, with progressive scoliosis, 1	607313	AR	3

Appendix 9

>robo1-RC Dm FBtr0343013 1395aa

Inside to outside helices : 7 found

from		to		score	center
29	(29)	49	(49)	1932	39
355	(355)	378	(378)	552	366
418	(418)	434	(434)	261	426
609	(609)	625	(625)	56	617
815	(819)	841	(838)	671	829
920	(920)	940	(938)	2974	930
1147	(1151)	1169	(1169)	753	1159

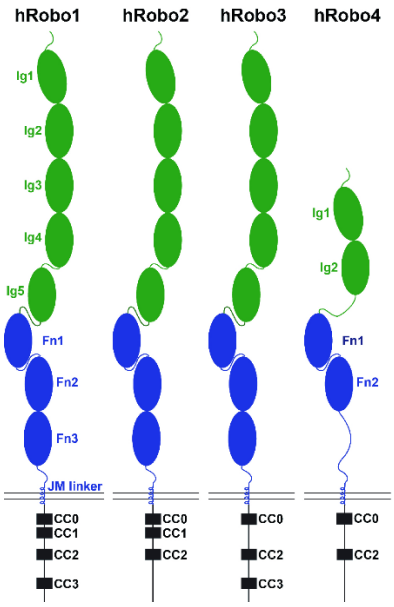
Outside to inside helices : 7 found

from		to		score	center
31	(31)	49	(47)	1991	39
356	(360)	379	(379)	158	368
415	(418)	434	(434)	222	426
610	(610)	627	(627)	239	618
815	(819)	841	(838)	771	829
916	(919)	940	(938)	3126	930
1147	(1149)	1170	(1168)	612	1159

ENSP00000420321 ROBO1 HSI

Inside to outside helices : 6 found

from		to		score	center
6	(6)	24	(24)	1294	16
423	(423)	445	(442)	577	433
465	(468)	488	(484)	1225	476
830	(832)	850	(850)	703	840
900	(900)	917	(917)	2658	909
1395	(1395)	1416	(1413)	193	1405

Outside to inside helices : 5 found

from		to		score	center
5	(5)	24	(24)	1008	15
423	(423)	445	(441)	543	433
470	(470)	488	(488)	998	478
829	(831)	850	(850)	824	840
898	(901)	918	(918)	2775	909