Contents

1 Inleiding

Dit onderwerp van onderzoek is heel belangrijk, omdat er in de experimentele wetenschap momenteel gezocht wordt naar alternatieven voor het gebruik van dieren in verschillende laboratoriumexperimenten. Aangezien verschillende wormsoorten een belangrijke rol spelen als modellen in medisch en biologisch onderzoek, is het cruciaal om hun genetisch materiaal te

onderzoeken. Dit project richt zich specifiek op de analyse van het genoom van de soort Lumbricida.De soort Lumbricus Terresteris hoort tot de fylogenetische familie Annelida, Clitellata, Oligochaeta, Crassiclitellata, Lumbricina, Lumbricidae (Erxleben and Grüning 12:19:56 +0000). De soort ringwormen (Annelida) zijn de oudste evolutionaire groep. De musculatuur lijkt hier sterk op de dwarsgestreptte musculatuur van dieren.(Pilato n.d.). Daarom is deze soort een van de mensrelevante modellen voor laboratoriumonderzoek.

Genoomannotatie is het proces dat gericht is op het identificeren van functionele componenten binnen een DNA-sequentie. Dit proces van annotatie biedt inzicht in het genoom door de plaats en functie van genen te specificeren, waaronder genen die eiwitten coderen of andere functies vervullen, evenals de bijbehorende regulerende elementen. De assemblage is altijd gebaseerd op de reads die zijn gegenereerd tijdens het sequentieproces. Het proces van genoomassemblage houdt in dat het originele genoom wordt gereconstrueerd uit kleine stukjes DNA, verkregen door middel van sequenting("De Novo Assembly Tutorial" n.d.). Deze reads zorgen ervoor dat het oorspronkelijke genoom meestal meerdere keren wordt gedekt. Bij het analyseren van genomische en metagenomische gegevens, is de gebruikelijke oplossing een verzameling contigs. Een contig is een aaneengeschakelde nucleotidesequentie. Deze contigs kunnen worden samengevoegd tot scaffolds, waarbij scaffolds bestaan uit een reeks contigs met een schatting van de afstanden tussen deze sequenties("The NCBI Eukaryotic Genome Annotation Pipeline" n.d.) processen van genomassemblage en annotatie zijn geïntegreerd in een groter geheel dat zich richt op de identificatie van het genoom. Het annoteren van genoom is nog steeds een proces dat veel tijd kost en verschillende soorten sequentieanalyses samenbrengt. Gezien de grootte en complexiteit van genomen, is de eerste stap naar volledige genoomassemblage meestal het verkrijgen van sequencinggegevens om ruwe assemblage en voorspelling van genmodellen te verkrijgen.

Het hele annotatieproces bestaat over het algemeen uit de volgende stappen : 1) Het maskeren van sterk repetitieve elementen in de genoomsequentie 2) het gebruik van transcripten en eiwitten van dezelfde of verwante soorten om ab initio te voorspellen . De bekende transcripten en eiwitten zijn opgeslagen in genetische databases zoals NCBI en BLAST. 3) gebruik van genzoekalgoritmen om mogelijke genstructuren te identificeren; 4) het combineren van deze gegevens om een eerste reeks genmodellen te creëren; 5) filteren van de resultaten op kwaliteit om de meest waarschijnlijke genmodellen te identificeren die volledige eiwitcoderende regio's.(ncbi.nlm.nih.gov n.d.) . In eerste instantie wordt de kwaliteitsselectie uitgevoerd met behulp van een set positieve controles voor het programma. Na het voltooien van deze controles wordt het percentage fout-positieven resuultaten duidelijk zichtbaar.

Een overzicht van publiek beschikbare genomen en annotaties in de soort Lumbricus terrestris.

Hoewel er in de bestaande literatuur slechts een beperkt aantal studies is dat de assemblage en annotatie van het genoom van

Lumbricus terrestris behandelt, zijn er wel talrijke beschrijvingen van de genoomannotatie van andere organismen. Voor de soorten C.Elegans en Lubricus Rubellis is er bijvoorbeeld een volledige annotatie. Voor het eerst wordt een gedetailleerde genoomassemblage van de genen van een soort Lumbricus terrestris gepubliceerd op 30 oktober (Blaxter, Spurgeon, and Kille 2023a). Door de innovatieve long-read sequencing methoden van Pacific Biosciences hebben de wetenschappers het genoom gesequenced en gepubliceerd. Hoewel het gepubliceerde genoom compleet is op sequentieniveau, is de analyse van de annotaties erg fragmentarisch. Deze genoomassemblage is de eerste die voor het publiek beschikbaar is. Het project legt de focus op het verder onderzoeken van de metadata van genoomassemblages. Tot nu toe zijn er in de literatuur geen uitgebreide en systematische studies gedaan over de annotatie van het Lumbricus terrestris genoom.

2 Samenvatting

2.1 1.Training van het AUGUSTUS-programma voor het ontdekken van nieuwe genmodellen en hun patronen.

De training van AUGUSTUS vond plaats in verschillende stappen. In het begin werd de predictor uitgevoerd met de standaardinstellingen voor caenorhabditis, wat leidde tot 11.000 voorlopige genmodellen voor één chromosoom, maar met een vrij lage nauwkeurigheid in de voorspellingen. Voor het opstellen van een eerste trainingsset van genen werden RNA-sequencingdata gebruikt. De transcriptomereads werden met TopHat op het genoom gemapt (zie documentatie protocol1, data_processing). Dit resulteerde in 9.953 voorspelde genmodellen per chromosoom op basis van het transcriptoom. Gemiddeld waren de genen ongeveer 5.146 baseparen lang, en elk gen had meestal rond de 3,2 exons (zie protocol1, data_processing, genemarkES, genemark.average_gene_length.out). De exons waren gemiddeld 1.719 baseparen, terwijl de introns gemiddeld 4.760 baseparen lang waren.

Daarna werden de genensets gefilterd met het Augustus-programma filterGenemark.pl. Na de filtratie bleven er 1.975 genen over op één chromosoom. Etrain werd uitgevoerd met genen die uit het transcriptoom kwamen. De uiteindelijke parameters werden gebruikt om de gff-annotatie te genereren. Een de novo-model met hoge specificiteit en sensitiviteitsscores van 8-9 voor de Lumbricus Terresstris werd verkregen via de mRna-pijplijn.

2.2 2. De training van het AUGUSTUS-programma met proteïne van langere evolutionaire afstand

Het AUGUSTUS-programma is getraind met proteïne die een langere evolutionaire afstand hebben. Hiervoor is een database uit Ortho DB, Arthropoda ("Bioinformatics Web Server - University of Greifswald" n.d.) gebruikt. Deze database is voorbereid met "ProtHints", wat een onderdeel is van de Braker-pipline("Gaius-Augustus/BRAKER" [2018] 2024). Voor de BLAST-analyse werd de versie ncbi-blast-2.16.0+ toegepast (zie protocol 2 documentatie). De OrthoDB-database diende als referentie. Om redundantie te verminderen, zijn alle trainingsgen aminozuursequenties met elkaar vergeleken en zijn alleen die eiwitsequenties behouden die minder dan 80% redundant zijn met andere sequenties in de set (zie protocol 2, scripts en documenten). Hieruit is een model (species) afgeleid dat de annotatie heeft opgeleverd.

Na het verwijderen van redundante genstructuren in de proteïne-pijplijn, zijn de specificiteit en gevoeligheid gestegen van 0,01 naar 0,4-0,5 punten voor het de novo-model van de eiwitpijplijn.

2.3 De training van het AUGUSTUS-programma met proteïne van kortere evolutionaire afstand

Proteinegegevens werden ingezet voor extern bewijs en als Hints van Extrinsic Evidence. Lumbricus proteoom("(Taxonomy_id:6397) in UniProtKB Search (698) | UniProt" n.d.), Eisenia fetida proteoom("(Taxonomy_id:6393) in UniProtKB Search (633) | UniProt" n.d.) en Genomethreader("Genometools/Genomethreader" [2019a] 2024) zijn toegepast. Door de alignment met GenomeThreader zijn er 20 genen geïdentificeerd voor chromosoom 1. Op chromosoom 1 zijn deze eiwitten en hun coördinaten aangetroffen: Q8MWU7 DNA-binding transcription factor activity, RNA polymerase II-specific, A0A088BZ25_EISFE 26S proteasome non-ATPase regulatory subunit 4, G3LY18_EISFE Superoxide dismutase Eisenia fetida, B9TY06_LUMTE Superoxide dismutase Lumbricus terrestris, B9TY04_LUMRU Superoxide dismutase Lumbricus rubellus, Q9GRJ1 CALM_LUMRU Calmodulin Lumbricus rubellus, Q2I6A7_EISFE Calmodulin Eisenia fetida, V9VGQ0_LUMRU glutathione gamma-glutamylcysteinyltransferase Lumbricus rubellus, P92182 ACT1_LUMTE Actin-1 Lumbricus terrestris, P91754 ACT_LUMRU Actin Lumbricus rubellus, E9KJS6_9ANNE Beta-actin Lumbricus friendi,Q2I743_LUMTE Extracellular hemoglobin linker L2 subunit Lumbricus terrestris, Q0G8J7_LUMTE High-affinity serotonin transporter protein Lumbricus terrestris, V9GWR0_LUMTE Peroxidasin Lumbricus terrestris, Q8MWS8_9ANNE Hox20 Eisenia andrei, Q2I6A6_EISFE HSp60 Eisenia fetida, Q2I741_LUMTE Extracellular hemoglobin linker L4 subunit Lumbricus terrestris, P08924|GLB1_LUMTE Extracellular globin-1 Lumbricus terrestris, P92182ACT1_LUMTE Actin-1 Lumbricus terrestris, Q2I6A1_EISFE Ubiquitin Eisenia fetida, P84589 UBIQ_LUMTE Ubiquitin Lumbricus terrestris, A0A143Y4B3_9ANNE

Myeloid Differentation primary response protein MyD88, Q2I699_EISFE Protein kinase C2 Eisenia fetida.

3 Materialen en methoden

3.1 Alignment. Protocol1.Ruwe gegevens inspecteren

Voor de Alignment zijn Bowtie en Tophat gebruikt. Het transcriptome ID49 is afkomstig van Project: PRJEB59399("ENA Browser" n.d.), dat een verzameling genomische en transcriptomische data bevat voor Lumbricus terrestris, ook wel de gewone regenworm genoemd. Dit project is opgezet om de assemblage en annotatie van het genoom te ondersteunen. Je kunt de ruwe gegevens hier bekijken: https://www.ebi.ac.uk/ena/browser/view/PRJEB59399. Reference genoom: https://ftp.ensembl.org/pub/rapid-release/species/Lumbricus_terrestris/GCA_949752735.1/ensembl/genome/

Eerst wordt de index opgebouwd met bowtie2. Daarna vindt de Alignmenet plaats met Tophat. Cufflinks voegt alle reads samen tot transcripties met: cufflinks accepted_hits.bam. Script om de genoomindex te genereren:

```
bowtie2-build -f  Lumbricus_terrestris-GCA_949752735.1-softmasked.fa  lumter  --large-index
```

Nu we een referentie-index hebben gecreëerd, kunnen we verder met het uitlijnen van de reads. We doen dit om te ontdekken waar in het genoom de reads oorspronkelijk vandaan komen. TopHat is een RNA-Seq lezer die helpt bij het uitlijnen van gesplitste transcripties naar een referentie. Het ondersteunt splice-junctions, wat betekent dat je niet alleen alle exons krijgt, maar ook de verbindingspunten tussen twee exons, die we splice junctions noemen.

```
tophat lumter     sample_1.fastq  sample_2.fastq \
 --output-dir TopHAT \
```

```
 cufflinks  accepted_hits.bam
```

Cufflink zal transcripts.gtf genereren, terwijl TopHat accepted_hits.bam aanmaakt met de resultaten van de alignment en een lijst van verbindingspunten tussen de exons in jucntions.bed. Elke junction bestaat uit twee verbonden BED-blokken, waarbij elk blok zo lang is als de maximale overhang van een lees die de junction overspant. De score is het aantal uitlijningen dat de

junction overspant.

Het bestand introns.gff geeft details over de coördinaten van introns en de strengen (+-) die gebruikt kunnen worden voor ET-training.

OX457036.1 TopHat2 intron 253060 254504 12 + . .

Ten eerste moeten we de ruwe gegevens bekijken die we hebben van de genoom-Alignment . Elke junction bestaat uit twee verbonden BED-blokken,welke bevat informatie over de knooppunten die exons met elkaar verbinden. Junctions.bed (TopHat, protocol1, script2):

```
head(junctions)
```

```
##           V1     V2     V3           V4 V5 V6     V7     V8      V9 V10    V11
## 1 OX457036.1 135689 136300 JUNC00000001 16  - 135689 136300 255,0,0   2 143,22
## 2 OX457036.1 136278 139661 JUNC00000002 13  - 136278 139661 255,0,0   2  22,37
## 3 OX457036.1 139624 150988 JUNC00000003  9  - 139624 150988 255,0,0   2  37,70
## 4 OX457036.1 150918 153142 JUNC00000004  1  - 150918 153142 255,0,0   2  70,16
## 5 OX457036.1 150929 156647 JUNC00000005  1  - 150929 156647 255,0,0   2  59,92
## 6 OX457036.1 155453 155919 JUNC00000006  2  - 155453 155919 255,0,0   2  92,59
##       V12
## 1   0,589
## 2  0,3346
## 3 0,11294
## 4  0,2208
## 5  0,5626
## 6   0,407
```

Cufflink verwerkt de uitgelijnde RNA-Seq-reads die van Tophat komen en bouwt ze op in de transcripten en exonen.

```
transcripts <- read.table("lumbricus/protocol1/data_processing/TOPHAT/transcripts.gtf", sep="\t")

colnames(transcripts) <-  c("chr", "versie", "feature", "start", "end", "score", "strain", "v8")

transcripts  %>%  select(1:5)  %>% head()
```

```
##          chr    versie    feature  start    end
## 1 OX457036.1 Cufflinks transcript 109191 109546
## 2 OX457036.1 Cufflinks       exon 109191 109546
## 3 OX457036.1 Cufflinks transcript 124949 125423
## 4 OX457036.1 Cufflinks       exon 124949 125423
## 5 OX457036.1 Cufflinks transcript 135006 155436
## 6 OX457036.1 Cufflinks       exon 135006 135832
```

1. We gaan de outputbestanden van Tophat+Cufflink, namelijk accepted_hits.bam en junctions.bed, in IGV zetten, samen met het transcriptbestand van Cufflinks. Eerst hebben we een bed-bestand nodig.

```
awk '{if($3=="exon" ) {print $1,$4,$5, $7, $3 }}'  transcripts.gtf  > exon_ids.bed

awk '{if($3=="transcript" ) {print $1,$4,$5, $7, $3 }}'  transcripts.gtf  > tranasctips_ids.b
```

1. Bekijk de bed-bestanden voor de genoombrowser:

```
exons_ids <- read.table("lumbricus/protocol1/data_processing/TOPHAT/igv/exon_ids.bed", sep="\t")

transcript_ids <- read.table("lumbricus/protocol1/data_processing/TOPHAT/igv/transctipts_ids.bed", sep="\t")

head(exons_ids)
```

```
##                               V1
## 1 OX457036.1 109191 109546 . exon
```

```
## 2 OX457036.1 124949 125423 . exon
## 3 OX457036.1 135006 135832 - exon
## 4 OX457036.1 136279 136300 - exon
## 5 OX457036.1 139625 139661 - exon
## 6 OX457036.1 150919 150988 - exon
```

```
head(transcript_ids)
```

```
##                                   V1
## 1 OX457036.1 109191 109546 . transcript
## 2 OX457036.1 124949 125423 . transcript
## 3 OX457036.1 135006 155436 - transcript
## 4 OX457036.1 135006 156649 - transcript
## 5 OX457036.1 135006 156649 - transcript
## 6 OX457036.1 135006 156649 - transcript
```

Bekijk de exonen (diepblauw) en transcripties (lichtblauw) in IGV:

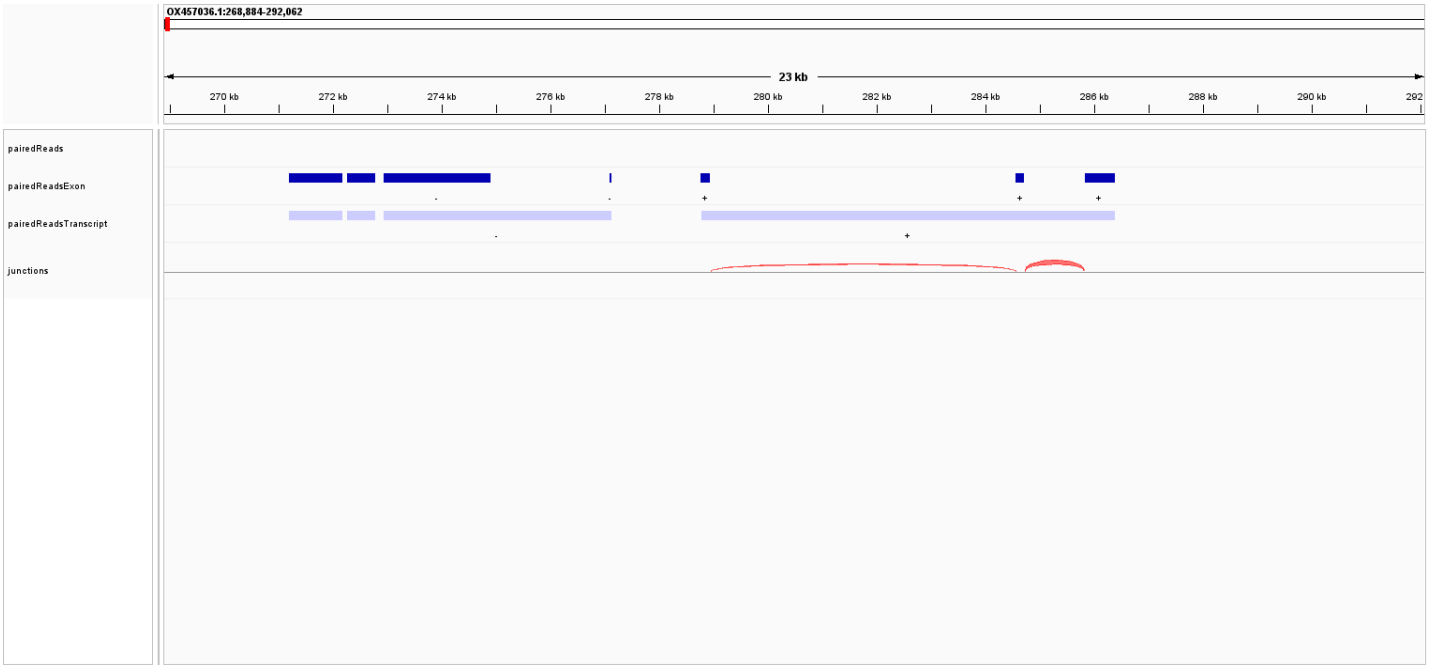


Figure 1: exon-transcripts structure chr1:23kb

Vervolgens plaatsen we junctions.bed (rood) en geaccepteerde hits of reads (grijs) op dezelfde track om de exon-intronstructuur te visualiseren (snapshots 1,2,4,5).

IGV Snapshots:

https://raw.githubusercontent.com/alenagrrr3/OX457036.1.html/refs/heads/main/igv_snapshot1.png

https://raw.githubusercontent.com/alenagrrr3/OX457036.1.html/refs/heads/main/igv_snapshot2.png

https://raw.githubusercontent.com/alenagrrr3/OX457036.1.html/refs/heads/main/igv_snapshot4.png

https://raw.githubusercontent.com/alenagrrr3/OX457036.1.html/refs/heads/main/igv_snapshot5.png

Laten we nu eens kijken naar de introns (protocol1-TopHat->introns.gff). We merken op dat er gaten in de structuur zitten en dat er ook lange introns uit de structuur komen. Het is daarom handig om verschillende RNA-Seq read mappers te gebruiken en te vergelijken. In dit onderzoek zijn Star, TopHat en Minimap toegepast, maar TopHat werd gekozen voor verder onderzoek omdat het goed samenwerkt met genemark. IGV Snapshots met introns (introns worden in geel weergegeven):

https://raw.githubusercontent.com/alenagrrr3/OX457036.1.html/refs/heads/main/igv_snapshot10.png

https://raw.githubusercontent.com/alenagrrr3/OX457036.1.html/refs/heads/main/igv_snapshot10.png

https://raw.githubusercontent.com/alenagrrr3/OX457036.1.html/refs/heads/main/igv_snapshot9.png

https://raw.githubusercontent.com/alenagrrr3/OX457036.1.html/refs/heads/main/igv_snapshot11.png

https://raw.githubusercontent.com/alenagrrr3/OX457036.1.html/refs/heads/main/igv_snapshot13.png

https://raw.githubusercontent.com/alenagrrr3/OX457036.1.html/refs/heads/main/igv_snapshot12.png

In de laatste (snapshot12) is er een enorm transcript te zien waarin niet alle introns en exons worden herkend, en er zijn veel gaten.

Voordat we GeneMarkET uitvoeren, verzamelen we enkele statistieken uit de primaire analyse. Eerst bekijken we de gemiddelde introns, exonen en lengtes.

```
introns  <- read.table("lumbricus/protocol1/data_processing/TOPHAT/introns.gff", sep="\t")


colnames(introns) <- c("chr","source","structure", "start", "end", "score", "strand", "v8", "v9")


head(introns)
```

```
##          chr  source structure  start    end score strand v8 v9
```

```
## 1 OX457036.1 TopHat2    intron 135833 136278    16      -  .  .
## 2 OX457036.1 TopHat2    intron 136301 139624    13      -  .  .
## 3 OX457036.1 TopHat2    intron 139662 150918     9      -  .  .
## 4 OX457036.1 TopHat2    intron 150989 153126     1      -  .  .
## 5 OX457036.1 TopHat2    intron 150989 156555     1      -  .  .
## 6 OX457036.1 TopHat2    intron 155546 155860     2      -  .  .
```

```
introns_length <- introns %>%   mutate(ilength=end-start)


max_intron <-  max(introns_length$ilength)  %>%  round(digits = 1)


 avr_intron <-  mean(introns_length$ilength)  %>%  round(digits = 1)
```

```
exons  <- read.table("lumbricus/protocol1/data_processing/TOPHAT/transcripts.gtf", sep="\t")


exons  <- exons %>%  select(1:5)
colnames(exons) <- c("chr","source","structure", "start", "end")
exons_length <- exons %>%   mutate(elength=end-start)


max_exon <-  max(exons_length$elength)  %>%  round(digits = 1)
max_exon
```

```
## [1] 255549
```

```
avr_exon <-  mean(exons_length$elength) %>%  round(digits = 1)
```

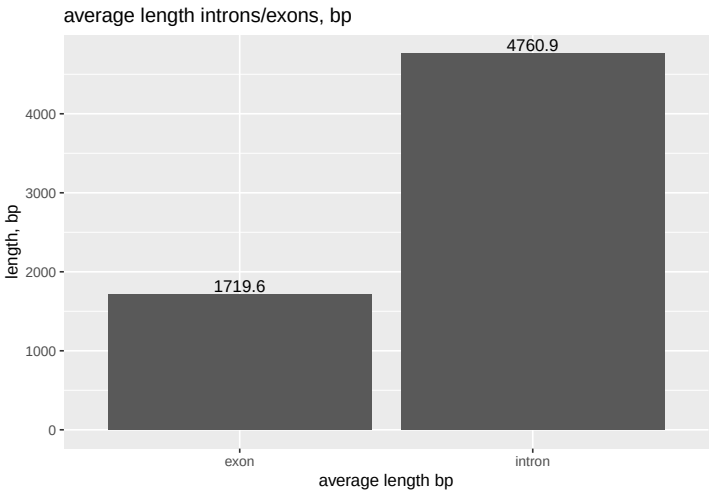
maximale lengte van intron : 2.91919×10^5

gemiddelde intronlengte 4760.9

maximale lengte van exon : 2.55549×10^5

gemiddelde lengte exon 1719.6

```
plot
```



4 GenemarkET. Model opbouwen (protocol 1). mRna pijplijnne

4.1 Deel 1. Model opbouwen

Het script bed_to_gff.pl van GeneMarkES maakt introns.gff aan vanuit de TopHat junctions.bed. Dit bestand bevat informatie over de strengen en kan direct gebruikt worden met GeneMarkET (protocol 1, script 2).

```
bed_to_gff.pl  --bed junctions.bed   --gff introns-gmes.gff  --label TopHat2
```

```
introns  <- read.table("lumbricus/protocol1/data_processing/TOPHAT/introns.gff", sep="\t")

colnames(introns) <- c("chr","aligner","structure", "start", "end", "score", "strand", "v8", "v9")

head(introns)
```

```
##          chr aligner structure  start    end score strand v8 v9
## 1 OX457036.1 TopHat2    intron 135833 136278    16      -  .  .
## 2 OX457036.1 TopHat2    intron 136301 139624    13      -  .  .
## 3 OX457036.1 TopHat2    intron 139662 150918     9      -  .  .
## 4 OX457036.1 TopHat2    intron 150989 153126     1      -  .  .
```

```
## 5 OX457036.1 TopHat2    intron 150989 156555     1      -  .  .
## 6 OX457036.1 TopHat2    intron 155546 155860     2      -  .  .
```

Om genemark met introns.gff uit te voeren:

```
../../gmes_petap.pl     --verbose --sequence   genome.fa   --ET  introns.gff
```

1. GeneMarkET gaat een ghmm-model en genemark.gtf produceren. Dit bestand(gtf) bevat informatie over de start- en eindcoördinaten van genen, die in de daaropvolgende stap gebruikt zal worden.
2. Genemark maakt gebruik van filterGenemark.pl voor kwaliteitscontrole. Dit zorgt ervoor dat alleen de genmodellen die niet geregistreerd zijn in de intronstructuur behouden blijven. (protocol1, script3) Na het filteren van de primaire resultaten wordt er een set van 1.975 genmodellen voor één chromosoom opgeslagen in genemark.f.good.gtf.

```
filterGenemark.pl  --genemark genemark.gtf  --introns introns.gff
```

```
Average gene length: 5146
Average number of introns: 3.26484477042098
Good gene rate: 0.198432633376871
Number of genes: 9972
Number of complete genes: 9953
Number of good genes: 1975
Number of one-exon-genes: 1982
Number of bad genes: 7997
Good intron rate: 0
One exon gene rate (of good genes): 1.00354430379747
```

```
 cut -f 2,3,4,5  lumbricus/protocol1/data_processing/GeneMarkES/genemark.f.good.gtf   | head
```

```
## bash: cannot set terminal process group (2732574): Inappropriate ioctl for device
## bash: no job control in this shell
## GeneMark.hmm3    stop_codon  51009   51011
## GeneMark.hmm3    CDS 51009   54860
```

```
## GeneMark.hmm3    exon    51009   54860
## GeneMark.hmm3    start_codon 54858   54860
## GeneMark.hmm3    stop_codon  82883   82885
## GeneMark.hmm3    CDS 82883   86734
## GeneMark.hmm3    exon    82883   86734
## GeneMark.hmm3    start_codon 86732   86734
## GeneMark.hmm3    stop_codon  116110  116112
## GeneMark.hmm3    CDS 116110  117048
## bash: [2732617: 2 (255)] tcsetattr: Inappropriate ioctl for device
```

1. Genemark.f.good.gtf is nu klaar om een trainingsset te maken van (protocol1, stap 4 en 5). Eerst wordt gtf omgezet naar gb. Zie protocol1, data_processing, Bonafide.

```
computeFlankingRegion.pl bonafide.gtf
```

```
Total length gene length (including introns): 1780572. Number of genes: 1975. Average Length: 901.555443037975
```

```
The flanking_DNA value is: 450 (the Minimum of 10 000 and 450)
```

```
gff2gbSmallDNA.pl bonafide.gtf genome.fa 450 tmp.gb
filterGenesIn_mRNAname.pl bonafide.gtf tmp.gb > bonafide.gb
```

```
cat lumbricus/protocol1/data_processing/bonafide/bonafide.gb | head
```

```
LOCUS       OX457036.1 Lumbricus terrestris genome assembly, chromosome: 1_50559-55310   4752 bp  DNA
FEATURES             Location/Qualifiers
     source          1..4752
     mRNA            complement(451..4302)
                     /gene="1_t"
     CDS             complement(451..4302)
                     /gene="1_t"
```

```
BASE COUNT     1219 a   994 c  853 g   1655 t   31 n
ORIGIN
        1 catccgtctt tttggaatcg atttttatcg tattctgaaa tgttcttatc aatcttacac
       61 cggctgcaaa ttttcttatc cttagtttcc ttatttttcct ggctcgcgta cttatgcgct
      121 agctccttta ctttagcatt tttacagagt ttacagctcg gataacttcc gttcttttgt
      181 cttttatctt tatgaaattc atctaagctt ttttcaatct tacagcagca gcaaactttt
```

Bonafide.gb is klaar om etrain te starten

4.2 Etrain (protocol7)

Op basis van de genen die we hebben verkregen via mRNA-alignment, gaan we een trainingsset opstellen om een nieuw model te trainen. In de vorige sectie hebben we bonafide.gb aangemaakt, waarin 1.975 geverifieerde genen voor een specifiek chromosoom zijn opgenomen. We zijn nu klaar om de ontwikkeling van een nieuwe species te starten.

```
conda activate c
new_species.pl --species=lumter
```

```
etraining --species=lumter bonafide.gb &> bonafide.out
```

Check for Stop Codonds:

```
grep -c "Variable stopCodonExcludedFromCDS set right" bonafide.out
```

0

We hoeven geen bad lijst op te stellen, omdat er geen stopcodons in de CDS aanwezig zijn.

```
grep -c LOCUS bonafide.gb
```

1975

Het randomSplit.pl-script splitst de data op in twee segmenten: een kleinere sectie genaamd test.gb voor trainingsdoeleinden, en een grotere sectie die train.gb wordt genoemd voor de evaluatie van het trainingsproces.

```
randomSplit.pl bonafide.gb 200
mv bonafide.gb.test test.gb
mv bonafide.gb.train train.gb
```

```
etraining --species=lumter train.gb &> etrain.out
```

Deze configuratie kan worden aangepast in het configuratiebestand (map config, species, lumter_parameters.cfg).

tag: 511 (0.259) taa: 700 (0.354) tga: 764 (0.387)

Evaluatie van de voorspelling:

```
augustus --species=lumter test.gb > test.out
```

******* Evaluation of gene prediction *******

```
*******      Evaluation of gene prediction     *******


---------------------------------------------\
                | sensitivity | specificity |
---------------------------------------------|
nucleotide level |        0.95 |       0.977 |
---------------------------------------------/


-----------------------------------------------------------------------------------------------------------------\
           |  #pred |  #anno |      |    FP = false pos. |    FN = false neg. |             |             |
           | total/ | total/ |   TP |--------------------|--------------------| sensitivity | specificity |
           | unique | unique |      | part | ovlp | wrng | part | ovlp | wrng |             |             |
```

```
-------------------------------------------------------------------------------------------------------------|
           |        |        |      |                32 |                40 |             |             |
exon level |    192 |    200 |  160 | ------------------ | ------------------ |         0.8 |       0.833 |
           |    192 |    200 |      |   24 |    0 |    8 |   24 |    0 |   16 |             |             |
-------------------------------------------------------------------------------------------------------------/


-------------------------------------------------------------------------\
transcript | #pred | #anno |   TP |   FP |   FN | sensitivity | specificity |
-------------------------------------------------------------------------|
gene level |   192 |   200 |  160 |   32 |   40 |         0.8 |       0.833 |
-------------------------------------------------------------------------/
```

See also: lumbricus/protocol1/test/test.out

Hier eindigt onze mRNA-pijplijn, waarbij we een hoge specificiteitsscore hebben bereikt voor het model dat we hebben gemaakt voor Lumbricus Terrestris. Dit model zal dienen voor visualisatie.

5 ProtHints en de eiwitpijplijn

5.1 Pipiline met eiwitten van grotere evolutionaire afstand

Er zijn veel genen in verschillende genoom die door hun evolutionaire oorsprong met elkaar verbonden zijn. De gelijkenis tussen eiwitsequenties is goed zichtbaar. OrthoDB is een belangrijke bron voor eiwitten en dient als een database die eiwitten met een uitgebreider evolutionair verleden omvat. Zie protocol 2.

```
../bin/prothint.py  ../OX457036.1.fasta  ../Arthropoda.fa
```

```
grep ">" seed_proteins.faa  | wc -l
```

14733

Prothint heeft een database met eiwitten voorbereid voor startAlign.pl. Het resultaat was 14.733 eiwitten in het bestand seed_proteins.faa. Dit seed-bestand kan worden gebruikt met startAlign.pl om een gth.concat.alg-object te verkrijgen, dat vervolgens wordt gebruikt om bonafide.gb te genereren.

```
head lumbricus/protocol2/data_processing/ProtHints/seed_proteins.faa
```

```
## bash: cannot set terminal process group (2732574): Inappropriate ioctl for device
## bash: no job control in this shell
## >6249_g
## MPSVSGLIEMMMMMTATITVMMTVTTVRIVERLGWGSYDTDGDDGDDDDDDDDDDDDDDDDDDDSNNNSSNPPQVTAELCRRELRRCRHRFRSTSSEMTAPPAASA
##
## >10626_g
## MLGRGDCERKKQGNGILETAIHEHAWLQYLEGTDERNGKKSKAGNLKAKREKLQKMRKGDIEEIGLLRGFAERKEKQGETEGLTGQVEEMEIDGPTTEKARHCLVAK
##
## >2633_g
## MSSAHAHVNASRRQQRQTINVRQRKDGEGRRLKRGVLVGNSEDLTVNWVWKATRCRPVPLRYQGVSNETLRMNCNSTSGEGRFGTAIAIGVRRQKKGAKRQQDEKLP
##
## >1749_g
```

Naast het seed_proteins.faa genereert protHints een prothint_augustus.gff hintsbestand dat je direct kunt gebruiken met augustus.

```
head lumbricus/protocol2/data_processing/\
ProtHints/prothint_augustus.gff
```

```
## bash: cannot set terminal process group (2732574): Inappropriate ioctl for device
## bash: no job control in this shell
## OX457036.1   ProtHint    stop    51009   51011   2   -   0   src=P;mult=9;pri=4;al_score=0.163636;
## OX457036.1   ProtHint    start   52806   52808   2   -   0   src=P;mult=2;pri=4;al_score=0.2;
## OX457036.1   ProtHint    intron  53104   53221   2   -   .   src=P;mult=1;pri=4;al_score=0.361685;
## OX457036.1   ProtHint    intron  53515   53655   0   -   .   src=P;mult=1;pri=4;al_score=0.108387;
## OX457036.1   ProtHint    start   55225   55227   0   -   0   src=P;mult=1;pri=4;al_score=0.104132;
## OX457036.1   ProtHint    stop    82883   82885   2   -   0   src=P;mult=11;pri=4;al_score=0.163636;
## OX457036.1   ProtHint    intron  84978   85095   2   -   .   src=P;mult=1;pri=4;al_score=0.361685;
## OX457036.1   ProtHint    intron  85389   85529   0   -   .   src=P;mult=1;pri=4;al_score=0.108387;
## OX457036.1   ProtHint    start   87099   87101   0   -   0   src=P;mult=1;pri=4;al_score=0.104132;
```

```
## OX457036.1   ProtHint    intron  144544  144597  0   +   .   src=P;mult=1;pri=4;al_score=0.13595;
```

2. We kunnen augustus meteen draaien met de prothint_augustus.gff die door de eiwitten zijn gemaakt, voordat we de trainingsset aanpakken.

```
augustus --species=lumter\
--predictionStart=2000000 --predictionEnd=3000000\
OX457036.1.fasta\
--extrinsicCfgFile=extrinsic.cfg\
--hintsfile=prothint_augustus.gff \
> augustus.hints.prots.orthodb.arthropoda.2-3mb.gff
```

Hierdoor ontstaat een annotatie voor 2mb-3mb van het chromosoom, gebaseerd op de eiwitindicaties van eiwitten die een lange evolutionaire afstand hebben.

```
cat   lumbricus/protocol2/data_processing\
/ProtHints/augustus.hints.prots.orthodb.arthropoda.2-3mb.gff | \
tail -n 50
```

```
## bash: cannot set terminal process group (2732574): Inappropriate ioctl for device
## bash: no job control in this shell
## # 3'UTR exons and introns: 0/0
## # hint groups fully obeyed: 0
## # incompatible hint groups: 2
## #      P:   2 (407821_0:000ad4_584_g)
## # end gene g81
## ###
## # start gene g82
## OX457036.1   AUGUSTUS    gene    2981898 2982863 1   -   .   g82
## OX457036.1   AUGUSTUS    transcript  2981898 2982863 1   -   .   g82.t1
## OX457036.1   AUGUSTUS    stop_codon  2981898 2981900 .   -   0   transcript_id "g82.t1"; gene_id "g82";
## OX457036.1   AUGUSTUS    CDS 2981898 2982863 1   -   0   transcript_id "g82.t1"; gene_id "g82";
```

```
## OX457036.1   AUGUSTUS    start_codon 2982861 2982863 .   -   0   transcript_id "g82.t1"; gene_id "g82";
## # protein sequence = [MDDEETVPYSLPRTTSTPATKGAAEASAFGQSRAEAYRTFEDPEYQFLDLPKKDRKKVLISETTVSDSKRWEDASHLM
## # GPRKIQMKPGKFDGTSSLESFLTQFEVCARHNRWDDSDKVDFLRCALDKAATQLLWDFGARADVTYDQLVGRLRQRYGVEGQAETYRAQLYYRRQRAD
## # ESLSDLLHDIRRLVVLAYPVPSNETTEIVARDSFLEAIRDRELSLKVREREPKSIDEAYRVALRLSAYQQMTDVDDRRRPPNRVRQTQEADAGNQLQT
## # QLDGFLAAQRKWQRDFEDRISLQLNELRNQSQTHPDVAPATRNPASP]
## # Evidence for and against this transcript:
## # % of transcript supported by hints (any source): 100
## # CDS exons: 1/1
## #      P:   1
## # CDS introns: 0/0
## # 5'UTR exons and introns: 0/0
## # 3'UTR exons and introns: 0/0
## # hint groups fully obeyed: 0
## # incompatible hint groups: 1
## #      P:   1 (407821_0:000ad4_584_g)
## # end gene g82
## ###
## # start gene g83
## OX457036.1   AUGUSTUS    gene    2983320 2984174 0.91    -   .   g83
## OX457036.1   AUGUSTUS    transcript  2983320 2984174 0.91    -   .   g83.t1
## OX457036.1   AUGUSTUS    stop_codon  2983320 2983322 .   -   0   transcript_id "g83.t1"; gene_id "g83";
## OX457036.1   AUGUSTUS    CDS 2983320 2984174 0.91    -   0   transcript_id "g83.t1"; gene_id "g83";
## OX457036.1   AUGUSTUS    start_codon 2984172 2984174 .   -   0   transcript_id "g83.t1"; gene_id "g83";
## # protein sequence = [MEKAGLYFNLKKTKLMTTENWTSFEVDGEEMKVVTCFCFFGAMIENDGGCERYCGSLAGGINFFAVCVFFACSERTCL
## # SEPLVASSSCPLEPAPSSRLFARSNLPLRAARCLFELPDRTCPSEPPVRCSSRTCPSEPLAASSSCPLEPAPPSRLSACSGRLRPLRAASSLFELLAR
## # TSLFRTFAAESNLFVRAACLLLRTGLETYKRRKKKKPSFAVGIEVGESQSLRVNPSGVSQRNEKGSSSSVVRSPSPRKVISSIRQSEVSSSFKLRLKL
## # RLNSGQFVVE]
## # Evidence for and against this transcript:
## # % of transcript supported by hints (any source): 0
## # CDS exons: 0/1
```

```
## # CDS introns: 0/0
## # 5'UTR exons and introns: 0/0
## # 3'UTR exons and introns: 0/0
## # hint groups fully obeyed: 0
## # incompatible hint groups: 0
## # end gene g83
## ###
## # command line:
## # augustus --species=lumter --predictionStart=2000000 --predictionEnd=3000000 OX457036.1.fasta --extrinsicC
```

5.2 Protocol 2. Het creëren van genstructuren voor training op basis van eiwitten.

5.3 GenomeThreader

We hebben 14.733 eiwitten verzameld uit de eerdere secties. Nu gaan we een trainingsset opzetten met deze eiwitten. Uit de oorspronkelijke 14.733 eiwitten hebben we een klein deel gekozen om de trainingsset te vormen.

```
startAlign.pl  --genome OX457036.1.fasta  \
--prot seed_proteins.faa  \
--pos OX457036.1:1-10000000 \
--prg gth
```

2. Hierdoor ontstaat het object gth.concat.aln, dat vervolgens kan worden geconverteerd naar het gtf-formaat (protocol2 ,data_processing, protHints).

```
gth2gtf.pl gth.concat.aln bonafide.gtf
```

Converting GenomeThreader file align_gth/gth.concat.aln to gtf format

Controleer het gtf-bestand :

```
head lumbricus/protocol2/data_processing/Bonafid/bonafide.gtf
```

```
## bash: cannot set terminal process group (2732574): Inappropriate ioctl for device
## bash: no job control in this shell
## OX457036.1   gth CDS 51009   54860   .   -   0   gene_id "OX457036.1_g_gene1_mRNA1"; transcript_id "OX45703
## OX457036.1   gth exon    51009   54860   .   -   0   gene_id "OX457036.1_g_gene1_mRNA1"; transcript_id "OX4
## OX457036.1   gth CDS 82883   86734   .   -   0   gene_id "OX457036.1_g_gene2_mRNA2"; transcript_id "OX45703
## OX457036.1   gth exon    82883   86734   .   -   0   gene_id "OX457036.1_g_gene2_mRNA2"; transcript_id "OX4
## OX457036.1   gth CDS 104626  104645  .   -   2   gene_id "OX457036.1_g_gene3_mRNA3"; transcript_id "OX45703
## OX457036.1   gth exon    104626  104645  .   -   2   gene_id "OX457036.1_g_gene3_mRNA3"; transcript_id "OX4
## OX457036.1   gth CDS 104696  104750  .   -   0   gene_id "OX457036.1_g_gene3_mRNA3"; transcript_id "OX45703
## OX457036.1   gth exon    104696  104750  .   -   0   gene_id "OX457036.1_g_gene3_mRNA3"; transcript_id "OX4
## OX457036.1   gth CDS 104904  105745  .   -   2   gene_id "OX457036.1_g_gene3_mRNA3"; transcript_id "OX45703
## OX457036.1   gth exon    104904  105745  .   -   2   gene_id "OX457036.1_g_gene3_mRNA3"; transcript_id "OX4
```

```
computeFlankingRegion.pl bonafide.gtf
```

Output van computeFlankingRegion.pl:

Total length gene length (including introns): 5412279. Number of genes: 1090. Average Length: 4965.39357798165 The flanking_DNA value is: 2482 (the Minimum of 10 000 and 2482)

```
gff2gbSmallDNA.pl bonafide.gtf genome.fa 2482 bonafide.gb
```

Bonafide.gb wordt in de volgende pipeline gebruikt om redundantie te verwijderen.

5.4 Protocol 6.Verwijderen van Redundant Genstructuren (protocol 6)

Voor NCBI Blast, controleer de link en stel het Path in naar de Blast uitvoerbare bestanden.

```
wget ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/
```

export PATH=$PATH$:HOME/ncbi-blast-2.16.0+

Maak gebruik van de opgegeven commandoregel om het GTF-bestand van de trainingsgenstructuur te transformeren naar een FASTA-bestand dat de eiwitsequentie omvat.

```
gtf2aa.pl genome.fa bonafide.f.gtf prot.aa
```

Inspecteer prot.aa :

```
head lumbricus/protocol2/data_processing/Redundancy/prot.aa
```

```
## bash: cannot set terminal process group (2732574): Inappropriate ioctl for device
## bash: no job control in this shell
## >OX457036.1_t_gene395_mRNA512
## ESLLPRCCPAGRGGGSQDSIAHARCFDRRITFSMMTLVGLGKEGLKRRKGGMDGERDLNWLEGGMGGEVQNWRVIGIERRY*
## >OX457036.1_t_gene508_mRNA640
## MEESRPVTPAQPSRPPSSMEVLLEAIQTNAKSTHDAMTSIQSSLQLNARDTQEAIATVELNVLAVQSNVREEISSVKSYVRDTQDAISSVQSNVSDAISSVQLNVRE
## >OX457036.1_t_gene532_mRNA668
## MTCLRRIEGVTRRERIRNTEIHNRLKIQRDIVDRIQIRRMRYFGHVVRMQSGRYPKVALQGYVHGKRRRGRPRKRWMDVAEEDCLRMGLTVGEATRRAQDRDDWRLS
## >OX457036.1_t_gene891_mRNA1213
## GKGRVNGCCFWRIRSGKLVREISTFCDIEFCEFCKFGRDSFEVSCRGGKMASLEEELIPEFGDVRDIPSDTLRLVSETYGEEVEDVSRSQVRRMAMKPLSPKLGSAA
## >OX457036.1_t_gene296_mRNA397
## WSEEPEEGDGVWLVMIVEELQKIGIHEADHSMVDHIRNEEVLKLAGSRYLEYIIMGRRGRLAGHILRLPKERIARTAIKWVPEGGKRRRGRPRNTWRRTFKGDLERM
```

Voer een Blast uit van alle eiwitsequenties uit de vorige stap met elkaar en toon alleen de eiwitsequenties die minder dan 80% identiek zijn aan een andere sequentie in de groep.

```
aa2nonred.pl prot.aa prot.nr.aa
```

```
head lumbricus/protocol2/data_processing/Redundancy/prot.nr.aa
```

```
grep ">"  lumbricus/protocol2/data_processing/Redundancy/prot.nr.aa | wc -l
```

```
## bash: cannot set terminal process group (2732574): Inappropriate ioctl for device
## bash: no job control in this shell
## >OX457036.1_t_gene753_mRNA930
## SIVGAATEVYNRMSSDFLPTPTKSHYIFNLRDLSKCIQESKQVFRLFCHEALRVFHDRLTTSEDKMSFYAILAEIAPKFFNENADAQSFLKHPIIFGDFIKVAAPRE
```

```
## >OX457036.1_t_gene803_mRNA1030
## SQKSRASATIVVCDLDHMMIRLPHFTAKRSVEPFQSTEEQVLGRIRSFPQGSSGGPDGLRPQHLSDLVNCVEIGSELIFAITGLVNLLLKGECPEDIRPVLFGGTLM
## >OX457036.1_t_gene103_mRNA135
## MSINFAQRIQMPGIERVHGVTKVRNEFNILGYSVSFRYVISVFEDRIPYRLRKEIRLTGIRNAVDIGSSENANCLYVSDYEEKCVRKITRERDGGHKIIKWLITAYR
## >OX457036.1_t_gene500_mRNA632
## IMRAEIQGRLNRGRQKKSWMDMIQQDMEFLGLRKEEVRDRTTWRQRIRINGLKYVYVYGHVSVNMKDIIIEHRLTVAELHFLKRAEILDRREKPLDVERKRQTETET
## >OX457036.1_t_gene503_mRNA635
## MCEVAEYFENGELVIFDDSDPAPSYADEMESDEMDDSKSDFPEAECAMALLELAQSFGLVSSLNSFGHINDETGLRNAATEPSNVPLNNTAENLASTADARQHFSAF
## 602
```

Daarna hebben we 602 niet-redudante eiwitten om mee verder te gaan:

```
grep ">"  lumbricus/protocol2/data_processing/Redundancy/prot.nr.aa | wc -l
```

```
## bash: cannot set terminal process group (2732574): Inappropriate ioctl for device
## bash: no job control in this shell
## 602
```

```
cat bonafide.gb | perl -ne 'if(m/\/gene=\"(\S+)\"/){ \
print "\"".$1."\"\n";}' | sort -u > traingenes.lst
```

regel 1: syntaxisfout bij onverwacht token '('

Dit leverde een syntaxisfout op, waarna alle perl -ne regex werden vervangen door Python regex, die werden uitgevoerd in de IDE.

```
import re
import subprocess


# Read from the file 'bonafide.gb'
with open('bonafide.gb', 'r') as file:
    content = file.read()
```

```
# Find all unique gene names
gene_names = set(re.findall(r'/gene="(\S+)"', content))


# Writing unique gene names to a file
with open('traingenes.lst', 'w') as f:
    for gene in sorted(gene_names):
        f.write(f'"{gene}"\n')
```

De uitvoer bevat de strings die als transcriptnamen worden gebruikt in het bonafide.gtf-bestand, waaruit bonafide.gb oorspronkelijk is gemaakt, met aanhalingstekens.

```
head lumbricus/protocol2/data_processing/Redundancy/traingenes.lst
```

```
## "OX457036.1_t_gene1000_mRNA1441"
## "OX457036.1_t_gene1001_mRNA1448"
## "OX457036.1_t_gene1002_mRNA1452"
## "OX457036.1_t_gene1003_mRNA1454"
## "OX457036.1_t_gene1004_mRNA1455"
## "OX457036.1_t_gene1006_mRNA1462"
## "OX457036.1_t_gene1007_mRNA1463"
## "OX457036.1_t_gene1008_mRNA1464"
## "OX457036.1_t_gene1009_mRNA1468"
## "OX457036.1_t_gene100_mRNA132"
```

Hierna volgt een reeks scripts/opdrachten die alleen bedoeld zijn om een lijst te verkrijgen van niet-redudante genen en hun bijbehorende loci in GeneBank.Dit is voornamelijk een bewerking voor tekstbestanden

```
grep -oE '(OX457036[A-Za-z1-9._]{1,})\w+'  prot.nr.aa > nonred.lst
```

```
head lumbricus/protocol2/data_processing/Redundancy/nonred.lst
```

```
## OX457036.1_t_gene753_mRNA930
## OX457036.1_t_gene803_mRNA1030
## OX457036.1_t_gene103_mRNA135
## OX457036.1_t_gene500_mRNA632
## OX457036.1_t_gene503_mRNA635
## OX457036.1_t_gene504_mRNA636
## OX457036.1_t_gene573_mRNA720
## OX457036.1_t_gene618_mRNA766
## OX457036.1_t_gene384_mRNA500
## OX457036.1_t_gene496_mRNA628
```

Isoleer de genen in traingenes.lst van bonafide.gtf:

```
grep -f traingenes.lst -F bonafide.gtf > bonafide.f.gtf
```

```
head lumbricus/protocol2/data_processing/Redundancy/bonafide.f.gtf
```

```
## OX457036.1  gth CDS 51009   54860   .  -  0  gene_id "OX457036.1_g_gene1_mRNA1"; transcript_id "OX45703
## OX457036.1  gth exon    51009   54860   .  -  0  gene_id "OX457036.1_g_gene1_mRNA1"; transcript_id "OX4
## OX457036.1  gth CDS 82883   86734   .  -  0  gene_id "OX457036.1_g_gene2_mRNA2"; transcript_id "OX45703
## OX457036.1  gth exon    82883   86734   .  -  0  gene_id "OX457036.1_g_gene2_mRNA2"; transcript_id "OX4
## OX457036.1  gth CDS 104626  104645  .  -  2  gene_id "OX457036.1_g_gene3_mRNA3"; transcript_id "OX45703
## OX457036.1  gth exon    104626  104645  .  -  2  gene_id "OX457036.1_g_gene3_mRNA3"; transcript_id "OX4
## OX457036.1  gth CDS 104696  104750  .  -  0  gene_id "OX457036.1_g_gene3_mRNA3"; transcript_id "OX45703
## OX457036.1  gth exon    104696  104750  .  -  0  gene_id "OX457036.1_g_gene3_mRNA3"; transcript_id "OX4
## OX457036.1  gth CDS 104904  105745  .  -  2  gene_id "OX457036.1_g_gene3_mRNA3"; transcript_id "OX45703
## OX457036.1  gth exon    104904  105745  .  -  2  gene_id "OX457036.1_g_gene3_mRNA3"; transcript_id "OX4
```

```
grep -oE '(OX457036[A-Za-z1-9._]{1,})\w+'  prot.nr.aa > nonred.lst
```

```
head lumbricus/protocol2/data_processing/Redundancy/nonred.lst
```

```
## OX457036.1_t_gene753_mRNA930
## OX457036.1_t_gene803_mRNA1030
## OX457036.1_t_gene103_mRNA135
## OX457036.1_t_gene500_mRNA632
## OX457036.1_t_gene503_mRNA635
## OX457036.1_t_gene504_mRNA636
## OX457036.1_t_gene573_mRNA720
## OX457036.1_t_gene618_mRNA766
## OX457036.1_t_gene384_mRNA500
## OX457036.1_t_gene496_mRNA628
```

In nonred.lst gaan we nu een niet-redundante subset van genen vinden.

Voor het filteren van het bestand bonafide.gb hebben we een lijst met loci-namen nodig in plaats van genenamen.

```
cat bonafide.gb | perl -ne '
if ( $_ =~ m/LOCUS\s+(\S+)\s/ ) {
$txLocus = $1;
} elsif ( $_ =~ m/\/gene=\"(\S+)\"/ ) {
$txInGb3{$1} = $txLocus
}
if( eof() ) {
foreach ( keys %txInGb3 ) {
print "$_\t$txInGb3{$_}\n";
}
}' > loci.lst
```

```
Unrecognized character \xE2; marked by <-- HERE after <-- HERE near column 1 at -e line 1.

cat: write error: Broken pipe

./test.sh: line 2: syntax error near unexpected token `('

./test.sh: line 2: `if ( $_ =~ m/LOCUS\s+(\S+)\s/ ) {'
```

Deze commando van het protocol veroorzaakte een fout en is vervangen. Het is nu locilist.py (scripts, protocol2).

```
import re


txInGb3 = {}

txLocus = ""


with open("bonafideOrtho.gb.db") as file:

    for line in file:

        if re.search(r'LOCUS\s+(\S+)\s', line):

            txLocus = re.search(r'LOCUS\s+(\S+)\s', line).group(1)

        elif re.search(r'/gene="(\S+)"', line):

            gene = re.search(r'/gene="(\S+)"', line).group(1)

            txInGb3[gene] = txLocus


with open("loci.lst", "w") as output_file:

    for key in txInGb3.keys():

        output_file.write(f"{key}\t{txInGb3[key]}\n")
```

en nonred.loci.py (scripts, protocol2)):

```
import subprocess


with open('nonred.lst', 'r') as f:

    patterns = f.read().splitlines()
```

```
with open('loci.lst', 'r') as f:
    loci = f.read().splitlines()

matched_loci = [locus.split('\t')[1] for locus in loci if any(pattern in locus for pattern in patterns)]

with open('nonred.loci.lst', 'w') as f:
    f.write('\n'.join(matched_loci))
```

wat nonred.loci.lst en loci.lst (met 2 kolommen) produceert:

```
head lumbricus/protocol2/data_processing/Redundancy/nonred.loci.lst
```

```
## OX457036.1_102144-115856
## OX457036.1_161655-167728
## OX457036.1_180282-185623
## OX457036.1_225887-235418
## OX457036.1_345964-351295
## OX457036.1_411769-417637
## OX457036.1_418604-428585
## OX457036.1_428586-437296
## OX457036.1_468333-473965
## OX457036.1_488481-495418
```

```
head lumbricus/protocol2/data_processing/Redundancy/loci.lst
```

```
## OX457036.1_t_gene1_mRNA1 OX457036.1_48527-57342
## OX457036.1_t_gene2_mRNA2 OX457036.1_80401-89216
## OX457036.1_t_gene3_mRNA3 OX457036.1_102144-115856
## OX457036.1_t_gene4_mRNA4 OX457036.1_138781-147529
## OX457036.1_t_gene5_mRNA5 OX457036.1_161655-167728
```

```
## OX457036.1_t_gene6_mRNA6 OX457036.1_180282-185623
## OX457036.1_t_gene7_mRNA7 OX457036.1_225887-235418
## OX457036.1_t_gene8_mRNA8 OX457036.1_321850-327440
## OX457036.1_t_gene9_mRNA9 OX457036.1_345964-351295
## OX457036.1_t_gene10_mRNA10   OX457036.1_389861-394620
```

```
filterGenesIn.pl nonred.loci.lst bonafide.gb > bonafide.f.gb
```

Deze commando haalt enkel de laatste locus uit de bonafide.gb. Het doel is om alle unieke loci uit de bonafide.gb te verzamelen, niet alleen de laatste.

Om alle unieke loci te krijgen, moeten we dit in een loop zetten (protocol2, scripts, bonafide.nonred.f.py).

```
import re
origfilename ="bonafideRED.gb"
goodfilename ="nonred.loci.lst"


goodlist = {}
with open(goodfilename, 'r') as goodfile:
     for line in goodfile:
          goodlist[line.strip()] = 1




with open(origfilename, 'r') as origfile:
    content = origfile.read().split('\n//\n')
    for gendaten in content:
          match = re.match(r'^LOCUS +(\S+) .*', gendaten)
          if match:
              genname = match.group(1)
              if genname in goodlist:
                  with open('bonafide.filtered.nonred.gb', 'a') as f2:
                      f2.write( gendaten+ '\n'+'//'+'\n')
```

```
                    f2.close()
```

```
grep -c LOCUS lumbricus/protocol2/data_processing/Redundancy/bonafide.f.nonred.gb
```

```
## 602
```

Na deze fase zijn er 602 met niet-redudante loci in Bonafide.

5.5 Trainingsset van Proteins.Etrain

We hebben in de vorige sectie 602 niet-redundante genstructuren ontdekt die kunnen dienen om een nieuwe soort te ontwikkelen.

Creëer een nieuwe species

```
new_species.pl --species=wormNonredEP
```

```
etraining --species=wormNonredEP bonafide.gb &> bonafide.out
```

Check for stop-codons:

```
grep -c "Variable stopCodonExcludedFromCDS set right" bonafide.out
```

49

We moeten 49 stopcodons uitfilteren.Bad List:

```
etraining --species=wormNonredEP    bonafide.gb 2>&1\
| grep "in sequence" \
|  sed -E 's/.*n sequence (\\S+):.*/\\1/' \
| sort -u   > bad.pre.list
```

```
grep -oE  "in sequence.*(OX457036.[1-9A-Za-z_0-]{1,})\w+" \
bad.pre.list\
| grep -oE  "(OX457036.[1-9A-Za-z_0-]{1,})\w+"> bad.list
```

```
head  ~/lumbricus/protocol2/data_processing/bad-list/bad.list
```

```
## bash: cannot set terminal process group (2732574): Inappropriate ioctl for device
## bash: no job control in this shell
## OX457036.1_80264327-80269533
## OX457036.1_3169142-3174603
## OX457036.1_3169142-3174603
## OX457036.1_82306032-82311964
## OX457036.1_83519819-83526493
## OX457036.1_85356189-85367403
## OX457036.1_3254876-3258078
## OX457036.1_87513969-87519492
## OX457036.1_3258079-3261568
## OX457036.1_92067632-92073579
```

Vervolgens fitler bad.list uit bonafide.gb:

```
perl filterGenes.pl  bad.list  bonafide.filtered.nonred.gb  \
> bonafide.filtered.gb
```

```
grep -c LOCUS bonafide.gb  bonafide.filtered.gb
```

bonafide.gb:602 bonafide.filtered.gb:373

```
ln -s bonafide.filtered.gb  bonafide.gb
```

test.gb is een klein bestand dat dient voor training. Train.gb is een groot bestand dat gebruikt wordt om de training te evalueren.

```
randomSplit.pl bonafide.gb 200
mv bonafide.gb.test test.gb
mv bonafide.gb.train train.gb
```

```
etraining --species=wormNonredEP  train.gb &> etrain.out
```

```
tail -6 etrain.out | head -3
```

tag: 97 (0.26) taa: 102 (0.273) tga: 174 (0.466)

Je moet deze waarden corrigeren in je wormNonredEP_parameters.cfg in config map

```
augustus --species=wormNonredEP test.gb > test.out
```

Eerst werd er een test gedaan op het model voordat het geoptimaliseerd werd, waarbij redudante structuren werden verwijderd. Deze test gaf een gevoeligheid en specificiteit van 0.01.

Na het toepassen van het protocol voor het verwijderen van redundante genstructuren, nam de specificiteit toe met 0,3 tot 0,5 punten.

```
*******      Evaluation of gene prediction     *******

-----------------------------------------------\
                | sensitivity | specificity |
-----------------------------------------------|
nucleotide level |       0.942 |       0.762 |
-----------------------------------------------/
```

	#pred / unique	#anno / unique	TP	FP = false pos. part	FP = false pos. ovlp	FP = false pos. wrng	FN = false neg. part	FN = false neg. ovlp	FN = false neg. wrng	sensitivity	specificity
exon level	1884	1580	813	1071		767				0.515	0.432
	1884	1580		436	104	531	456	145	166		

transcript	#pred	#anno	TP	FP	FN	sensitivity	specificity
gene level	454	373	88	366	285	0.236	0.194

Zie lumbricus/protocol2/test/test.out voor meer informatie.

6 Pipiline met eiwitten van kortere evolutionaire afstand

Voor de eiwitten van kortere evolutionaire afstand is proteoom van wormen geselecteerd. Het proteoom komt van UniProt, dat zowel het proteoom van Lumbricus Terrestris als dat van Eisenia Fetida omvat. Twee van de Fasta-bestanden die we van UniProt hebben gekregen, zijn in één bestand samengevoegd.

Proteome Lumbcricus Terrestris: https://www.uniprot.org/uniprotkb?query=%28taxonomy_id%3A6397%29

Proteome Eisenia Fetida: https://www.uniprot.org/uniprotkb?query=%28taxonomy_id%3A6393%29

Samengevat Lumbricus en Eisenia:

lumbcricus ->protocol2.2 -> data_processing -> merged_6393_and_6397.fa

Fasta files, afkomstig van Uniprot: Eisienia: lumbricus ->protocol2.2 -> data_raw-> uniprotkb_taxonomy_id_6393_2024_12_29.fasta

Lumbricus : lumbricus ->protocol2.2 -> data_raw-> uniprotkb_taxonomy_id_6397_2024_12_29.fasta.

Eerste step is de ProtHints:

```
../bin/prothint.py ../OX457036.1.fasta ../merged_6393_and_6397.fa
```

Het programma ProtHints wordt gebruikt om hints voor te bereiden (ProtHints installatie vond plaats in protocol 2). In deze fase wordt het bestand prothint_augustus.gff aangemaakt. Voorbeeld prothint_augustus.gff :

```
OX457036.1  ProtHint    start   33409650    33409652    2   -   0   src=P;mult=2;pri=4;al_score=0.433058;
OX457036.1  ProtHint    intron  34198705    34199175    2   +   .   src=P;mult=2;pri=4;al_score=0.38446;
OX457036.1  ProtHint    intron  34199278    34199565    2   +   .   src=P;mult=2;pri=4;al_score=0.26901;
OX457036.1  ProtHint    intron  37878497    37880236    2   +   .   src=P;mult=1;pri=4;al_score=0.488541;
OX457036.1  ProtHint    intron  37880480    37881139    2   +   .   src=P;mult=1;pri=4;al_score=0.474112;
OX457036.1  ProtHint    stop    37881166    37881168    2   +   0   src=P;mult=1;pri=4;al_score=0.429752;
```

Je kunt hints gelijk toepassen in augustus.

```
augustus --species=caenorhabditis
--predictionStart=2000000 --predictionEnd=3000000\
OX457036.1.fasta
--extrinsicCfgFile=extrinsic.cfg
--hintsfile=prothint_augustus.gff
> augustus.extrinistics.hints.gff
```

For de extrinsic.cfg zee:

https://github.com/nextgenusfs/augustus/blob/master/config/extrinsic/cgp.extrinsic.cfg

Voorbeeld extrinsic.cfg:

```
# source of extrinsic information:
# M manual anchor (required)
# P protein database hit
# E EST/cDNA database hit
```

```
# C combined est/protein database hit
# D Dialign
# R retroposed genes
# T transMapped refSeqs
# W wiggle track coverage info from RNA-Seq

[SOURCES]
M RM E W P

#
# individual_liability: Only unsatisfiable hints are disregarded. By default this flag is not set
# and the whole hint group is disregarded when one hint in it is unsatisfiable.
# 1group1gene: Try to predict a single gene that covers all hints of a given group. This is relevant for
# hint groups with gaps, e.g. when two ESTs, say 5' and 3', from the same clone align nearby.
#
[SOURCE-PARAMETERS]



#   feature        bonus         malus   gradelevelcolumms
#       r+/r-
#
# the gradelevel colums have the following format for each source
# sourcecharacter numscoreclasses boundary    ...  boundary    gradequot  ...  gradequot
#

[GENERAL]
      start      1        0.8  M    1  1e+100  RM  1     1    E 1    1    W 1    1      P       1     1e3
       stop      1        0.8  M    1  1e+100  RM  1     1    E 1    1    W 1    1  P   1     1e3
        tss      1          1  M    1  1e+100  RM  1     1    E 1    1    W 1    1  P   1   1
        tts      1          1  M    1  1e+100  RM  1     1    E 1    1    W 1    1  P   1   1
```

```
        ass      1   0.95 0.1  M    1  1e+100  RM  1     1    E 1    1    W 1    1  P   1     100
        dss      1   0.95 0.1  M    1  1e+100  RM  1     1    E 1    1    W 1    1  P   1     100
   exonpart      1  .992 .985  M    1  1e+100  RM  1     1    E 1    1    W 1    1.02   P   1   1
       exon      1        0.9  M    1  1e+100  RM  1     1    E 1    1    W 1    1  P   1     1e4
 intronpart      1          1  M    1  1e+100  RM  1     1    E 1    1    W 1    1  P   1   1
     intron      1        .34  M    1  1e+100  RM  1     1    E 1  1e6    W 1    1  P   1     100
    CDSpart      1     1 .985  M    1  1e+100  RM  1     1    E 1    1    W 1    1  P   1     1e5
        CDS      1          1  M    1  1e+100  RM  1     1    E 1    1    W 1    1  P   1   1
    UTRpart      1     1    1  M    1  1e+100  RM  1     1    E 1    1    W 1    1  P   1   1
        UTR      1          1  M    1  1e+100  RM  1     1    E 1    1    W 1    1  P   1   1
     irpart      1          1  M    1  1e+100  RM  1     1    E 1    1    W 1    1  P   1   1
nonexonpart      1          1  M    1  1e+100  RM  1     1.15 E 1    1    W 1    1  P   1   1
  genicpart      1          1  M    1  1e+100  RM  1     1    E 1    1    W 1    1  P   1   1
```

Tijdens deze stap wordt er een gff-annotatiebestand geproduceerd. Voorbeeld Augustus gff van protein Hints:

```
# start gene g10
OX457036.1  AUGUSTUS    gene    2072765 2073299 0.59    +   .   g10
OX457036.1  AUGUSTUS    transcript  2072765 2073299 0.59    +   .   g10.t1
OX457036.1  AUGUSTUS    tss 2072765 2072765 .   +   .   transcript_id "g10.t1"; gene_id "g10";
OX457036.1  AUGUSTUS    5'-UTR  2072765 2072799 0.99    +   .   transcript_id "g10.t1"; gene_id "g10";
OX457036.1  AUGUSTUS    start_codon 2072800 2072802 .   +   0   transcript_id "g10.t1"; gene_id "g10";
OX457036.1  AUGUSTUS    single  2072800 2073033 0.93    +   0   transcript_id "g10.t1"; gene_id "g10";
OX457036.1  AUGUSTUS    CDS 2072800 2073033 0.93    +   0   transcript_id "g10.t1"; gene_id "g10";
OX457036.1  AUGUSTUS    stop_codon  2073031 2073033 .   +   0   transcript_id "g10.t1"; gene_id "g10";
OX457036.1  AUGUSTUS    3'-UTR  2073034 2073299 0.6 +   .   transcript_id "g10.t1"; gene_id "g10";
OX457036.1  AUGUSTUS    tts 2073299 2073299 .   +   .   transcript_id "g10.t1"; gene_id "g10";
# protein sequence = [MYKLVDETSKLAWLLCLMRMLSQKYYVSSMLMLANSRASLLPLLIAYNELISRDDELSCYRFLHSCDMFILTFFRRS]
# Evidence for and against this transcript:
# % of transcript supported by hints (any source): 0
```

```
# CDS exons: 0/1
# CDS introns: 0/0
# 5'UTR exons and introns: 0/1
# 3'UTR exons and introns: 0/1
# hint groups fully obeyed: 0
# incompatible hint groups: 0
# end gene g10
```

De proteïnesequentie in de gff is gereed om te gebruiken met de Blast-web of API-versie. Voor het evidence at protein level is het nuttig om GenomeThreader (https://genomethreader.org/download.html) in te zetten om specifieke eiwitten van Lumbricus Terrestris te vinden. Commando voor GenomeThreader 1.7.1:

```
gth  -genomic soft.masked.chromosome1.OX457036.1.fasta -protein merged_6393_and_6397.fa
```

GenomeThreader geeft de positie van het gevonden eiwit in de gegeven dna-sequentie, evenals de identificatie van het eiwit "Q8MWU7" (protocol2.2-> data_processing-> output.gth).

```
$ Arguments: -genomic soft.masked.chromosome1.OX457036.1.fasta -protein merged_6393_and_6397.fa
****************************************************************************
Protein Sequence: file=merged_6393_and_6397.fa, description=tr|Q8MWU7|Q8MWU7_9ANNE Hox01 (Fragment) OS=Eisenia

  1  HFNRYLTRRR RIEIAHALCL TERQI

Genomic Template: file=soft.masked.chromosome1.OX457036.1.fasta, strand=-, from=4635697, to=4635026, descripti

Predicted gene structure:

 Exon  1  4635397  4635323 (  75 n);  Protein      1     25 (  25 aa); score: 0.878

MATCH   OX457036.1- tr+ 0.878   75  1.000   P
```

```
PGS_OX457036.1-_tr+ (4635397  4635323)


Alignment (genomic DNA sequence = upper lines):


CACTTCAACA AATATCTGAC GAGGAGACGA AGGATCGAGA TCTCGCACCA GCTGTGCTTG     4635338
 H  F  N   K  Y  L  T   R  R  R   R  I  E   I  S  H  Q   L  C  L
 |  |  |   +  |  |  |   |  |  |   |  |  |   |  +  |     |  |  |
 H  F  N   R  Y  L  T   R  R  R   R  I  E   I  A  H  A   L  C  L          20



ACGGAACGAC AGGTG                                                     4635323
 T  E  R   Q  V
 |  |  |   |  +
 T  E  R   Q  I                                                           25



*******************************************************************************
```

Gegevens met betrekking tot chromosoom 1 voorspellingen komen voort uit bewijs van eiwitten (file GenomeThreader.output.gth, data_processing, protocol2.2).

```
Predicted:


strand=- from=4635697 to=4635026
Protein ID : Q8MWU7|
DNA-binding transcription factor activity, RNA polymerase II-specific


Predicted:


strand=- from=19847071 to=19841778
```

```
A0A088BZ25_EISFE 26S proteasome non-ATPase regulatory subunit 4 (Fragment) OS=Eisenia fetida


Predicted:

strand=- from=23688885  to=23683840

G3LY18_EISFE Superoxide dismutase  OS=Eisenia fetida


Predicted:

strand=-  from=23686573  to=23683870

B9TY06_LUMTE Superoxide dismutase  OS=Lumbricus terrestris


Predicted:

strand=- from=23686573  to=23684768

B9TY04|B9TY04_LUMRU Superoxide dismutase (Fragment) OS=Lumbricus rubellus




Predicted:

strand=+  from=37878052  to=37880778

Q9GRJ1  CALM_LUMRU Calmodulin OS=Lumbricus rubellus




Predicted:


strand=+ from=37879936 to=37880745

Q2I6A7_EISFE Calmodulin (Fragment) OS=Eisenia fetida




Predicted:


strand=+  from=39296646 to=39299946

V9VGQ0_LUMRU glutathione gamma-glutamylcysteinyltransferase OS=Lumbricus rubellus OX=35632
```

```
Predicted:

strand=+  from=44647234  to=44650203

P92182 ACT1_LUMTE Actin-1 OS=Lumbricus terrestris


Predicted:


strand=+ from=44647234 to=44650203

P91754|ACT_LUMRU Actin  OS=Lumbricus rubellus


Predicted:


strand=+ from=44648761 to=44649954

E9KJS6_9ANNE Beta-actin  OS=Lumbricus friendi


Predicted:


strand=+ from=51521291  to=51525245

Q2I743_LUMTE Extracellular hemoglobin linker L2 subunit OS=Lumbricus terrestris


Predicted:

strand=- from=55530224 to=55501373

Q0G8J7_LUMTE High-affinity serotonin transporter protein  OS=Lumbricus terrestris


Predicted:

V9GWR0_LUMTE Peroxidasin OS=Lumbricus terrestris

strand=+ from=57431103 to=57470825


Predicted:

strand=+ from=61318513 to=61319184
```

```
Q8MWS8_9ANNE Hox20  OS=Eisenia andrei


Predicted:

strand=- from=66406471 to=66405222

Q2I6A6_EISFE HSp60 (Fragment) OS=Eisenia fetida


Predicted:

strand=- from=7768149  to=77676755

Q2I741_LUMTE Extracellular hemoglobin linker L4 subunit OS=Lumbricus terrestris


Predicted:


strand=- from=77703849 to=77700628

P08924|GLB1_LUMTE Extracellular globin-1 OS=Lumbricus terrestris


Predicted:

strand=- from=87203704 to=87077023

P92182|ACT1_LUMTE Actin-1 OS=Lumbricus terrestris


Predicted:


strand=+ from=90736653 to=90737447

Q2I6A1_EISFE Ubiquitin (Fragment) OS=Eisenia fetida


Predicted:

 strand=+ from=90736725, to=90737447

P84589|UBIQ_LUMTE Ubiquitin (Fragment) OS=Lumbricus terrestris


Predicted:

strand=+ from=91986226 to=91991793
```

```
A0A143Y4B3_9ANNE Myeloid Differentation primary response protein MyD88


Predicted:
strand=+ from=98669240 to=98672183
Q2I699_EISFE Protein kinase C2 (Fragment) OS=Eisenia fetida
```

Visualisatie van protein evidence voor Chromosome I OX457036.1 Lumbricus Terrestris:

https://alenagrrr3.github.io/OX457036.1.html/OX457036.1.proteins.Lumbricus

Op dit moment hebben we 20 eiwitten gevonden op het eerste chromosoom, die afkomstig zijn uit het proteoom van Lumbricus Terrestris en Eisenia Fetida. Daarnaast zijn we geïnteresseerd in de andere genen, dus we keren terug naar onze gff annotatie. Dit keer passen we ons eigen model toe, dat is gebaseerd op eiwitten en ontwikkkeld, protolol, 2,6 en 7. Verkrijgen van gff annotatie:

```
augustus --species=protsLumter sequence.fasta  --predictionStart=2000000 --predictionEnd=3000000 --hintsfile=p
```

Gff3 file augustus.prothints.gff3 is te vinden: protocol2.2-> gff->augustus.prothints.gff3 Je kunt de webversie van Blast of Api Blast gebruiken om eiwitten te vinden. We zullen de Api-versie gebruiken. Om dit te doen, verwijderen we eerst alle overbodige tekens:

```
with open('augustus.prothints.gff3', 'r') as infile, open('parsed.gff3', 'w') as outfile:
    temp = infile.read().replace("#", "")
    outfile.write(temp)
```

```
 Predicted genes for sequence number 1 on both strands
 start gene g1
OX457036.1  AUGUSTUS    gene    2000789 2003917 1   +   .   ID=g1
OX457036.1  AUGUSTUS    transcript  2000789 2003917 1   +   .   ID=g1.t1;Parent=g1
OX457036.1  AUGUSTUS    start_codon 2000789 2000791 .   +   0   Parent=g1.t1
OX457036.1  AUGUSTUS    CDS 2000789 2003917 1   +   0   ID=g1.t1.cds;Parent=g1.t1
```

```
OX457036.1  AUGUSTUS    stop_codon  2003915 2003917 .   +   0   Parent=g1.t1
 protein sequence = [MEESRPVTPAQPSRPPSSMEILLEAIQTNARSTHEAIQTNAKSSQEAMQAHAKSTHDAMTSIQSSLQLNARETQEAIA
 TVEFNVLAVQSNVSEAISSVQSNVREEIREEISAVRDNVREALTEMVSRLERLEASPVPKPAVDSNPGYLTAITPADAPYHSTIGLGETLGARPKDFT
 QPGILRRSDRLAGRPPISYREYGSRKDWPPFLGWDSNPEVTSSCPPSISRARPQQHAVPSGEDPEVATPGMPIGAGVTIGPSQWGQISSRDFGDDRLE
 EETDYARTGEMAISFERRKEREFADNVEIMSDDGNVDIEISKIMESRPKTVKIIDNRMHAAQQPEFRNFEVNKPIANKLSREGGDRVNPVPLASSSLP
 LVEYPRPDPQWRMQASLAHVDQQRVEMAPSRVDFTQPASVMPTYQSMPDCVLDGRATSTMVGRDYARPDLPPGPAAMATVDWMQPYARPDHSNMPWRY
 APTYTPSAFYGSSDNRVWVDPLRCLDPLRPSDAVFPRWPTVVESARMSSCLGAQSFGFSRDAELPRCQAVMKSTPLERNETTADNKATSAVGETNALA
 PTYVSVGPIKIVPTTTVATQTTGDWELPSISSKGTVKELEPTASATAKEGEVKQSSTSPPKPTQFLKLGSFSGKTDVETFLRKFSVCARNNRWTDEER
 LNQLVVSLVEPATHLLSESNADSLDTWTALVQRLRERYGNAEQQALFQTQLSTRKQKADEDMGALVDDVRRLTSRAYPGSSTVHSEAIAVRAFLDALR
 DRTLALKIREREPKSLDEAYKVAMRLDGYQKAEDGGHEQHERRYGRVNAVKEEDDSEMRVLKRQVEQIMRQMERQPVSAPQYRDNNRGSWTGQNGSNG
 NGWSNANRNNWRNNQRCSICNRTGHWSSVCRYRQAAEQEDGPSARRCFECSAMDHIARFCPLRQQQQAPPSDNLGQPVGDNHDNPSVGRVFAVRSAPA
 PSRDQRNGQNQERRSRRSPSSSRRRDSGRNSPREERRCFYCDDSSHLLRSSPLRNGPPEDRIWRMLGEPTGLRPPENLTSAG]
```

Voor de volgende stap zullen we de CDS-coördinaten en eiwitsequentie uit ons gff3-bestand knippen.

```
cds_coords = []


content = open("parsed.gff3", 'r').read()
pattern_a = r'CDS.*\s+\d+\w+'
matches_a = re.findall(pattern_a, content)
cds_coords.extend(matches_a)
print(cds_coords)
```

```
CDS\t2000789\t2003917
```

```
protein_seq=[]


content = open("parsed.gff3", 'r').read()
pattern_b="protein sequence =.*[A-Za-z\s\]]{1,}\]"
```

```
matches_b=re.findall(pattern_b, content)
protein_seq.extend(matches_b)


print(protein_seq)
```

```
protein sequence = [MEESRPVTPAQPSRPPSSMEILLEAIQTNARSTHEAIQTNAKSSQEAMQAHAKSTHDAMTSIQSSLQLNARETQEAIA
 TVEFNVLAVQSNVSEAISSVQSNVREEIREEISAVRDNVREALTEMVSRLERLEASPVPKPAVDSNPGYLTAITPADAPYHSTIGLGETLGARPKDFT
 QPGILRRSDRLAGRPPISYREYGSRKDWPPFLGWDSNPEVTSSCPPSISRARPQQHAVPSGEDPEVATPGMPIGAGVTIGPSQWGQISSRDFGDDRLE
 EETDYARTGEMAISFERRKEREFADNVEIMSDDGNVDIEISKIMESRPKTVKIIDNRMHAAQQPEFRNFEVNKPIANKLSREGGDRVNPVPLASSSLP
 LVEYPRPDPQWRMQASLAHVDQQRVEMAPSRVDFTQPASVMPTYQSMPDCVLDGRATSTMVGRDYARPDLPPGPAAMATVDWMQPYARPDHSNMPWRY
 APTYTPSAFYGSSDNRVWVDPLRCLDPLRPSDAVFPRWPTVVESARMSSCLGAQSFGFSRDAELPRCQAVMKSTPLERNETTADNKATSAVGETNALA
 PTYVSVGPIKIVPTTTVATQTTGDWELPSISSKGTVKELEPTASATAKEGEVKQSSTSPPKPTQFLKLGSFSGKTDVETFLRKFSVCARNNRWTDEER
 LNQLVVSLVEPATHLLSESNADSLDTWTALVQRLRERYGNAEQQALFQTQLSTRKQKADEDMGALVDDVRRLTSRAYPGSSTVHSEAIAVRAFLDALR
 DRTLALKIREREPKSLDEAYKVAMRLDGYQKAEDGGHEQHERRYGRVNAVKEEDDSEMRVLKRQVEQIMRQMERQPVSAPQYRDNNRGSWTGQNGSNG
 NGWSNANRNNWRNNQRCSICNRTGHWSSVCRYRQAAEQEDGPSARRCFECSAMDHIARFCPLRQQQQAPPSDNLGQPVGDNHDNPSVGRVFAVRSAPA
 PSRDQRNGQNQERRSRRSPSSSRRRDSGRNSPREERRCFYCDDSSHLLRSSPLRNGPPEDRIWRMLGEPTGLRPPENLTSAG]
```

In de volgende fase moeten we een database opzetten waarin elke(of meerdere)CDS coördinaat(en) gekoppeld is aan een specifieke eiwitsequentie. Je kunt ook de eiwitsequentie kopiëren en een blast uitvoeren op de website als alternatief: https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome

Maak een databasebestand aan met eiwitten en hun CDS-coördinaten. Splits het hele GFF-bestand per “eind gen”, zodat elke gendata één geninformatie bevat.

```
import re
readtxtfile = r'parsed.gff3'
with open(readtxtfile) as fp:
    txtrawdata = fp.read()
    genes = re.split(r'end\sgene.*',txtrawdata)


print(genes[-1])
```

Doorloop de lijst met genen.

```
str_coords=[]

str_proteins=[]




for gendata in genes:

    cds_coords = []

    protein_seq=[]


    pattern_a = r'CDS.*\s+\d+\w+'

    matches_a = re.findall(pattern_a, gendata)

    cds_coords.extend(matches_a)


    pattern_b="protein sequence =.*[A-Za-z\s\]]{1,}\]"

    matches_b=re.findall(pattern_b, gendata)

    protein_seq.extend(matches_b)


    delimiter = " " # Define a delimiter

    joined_coords=delimiter.join(cds_coords)


    str_coords.append(joined_coords)

    str_proteins.append(protein_seq)




dict={"cd_coords":str_coords, "protein_seq":str_proteins}

df = pd.DataFrame(dict)
```

Exporteer het dataframe naar een bestand.

```
outfile = open("proteinsdb.fa", 'w')


for index, row in df.iterrows():

    print( ">",  row["cd_coords"],  "\n", row['protein_seq'])

    outfile.write( ">" + str(row["cd_coords"]) +  "\n" + str(row['protein_seq'])+ "\n" )

outfile.close()
```

Momenteel beschikken we over een bestand waarvan we enkel de overbodige tekens moeten verwijderen.

```
>CDS    2374823 2375022 CDS 2378236 2378761

protein sequence = [MHSSMQTNARDTQEAMQSHAREAQEAMQSHARETQEAISVVQSNVTEEISAIQSNITEEISAIQSNITEEISTVRSDV

REEISAVYSNVREVLTEVVTRIERLEESPVPRSVVGLNPGLRSPASIMADVPHQSTIRLAESSGARPKDFTHLGTLRRSERLAHKDPISYRELGSRDE

DSGNKVYVVLVRELEIMTSGSKVVNRTGHLGKLVAKDMATVPYDGAEQALVRFLNHISSRGRGIR]

>CDS    2381370 2384417

protein sequence = [MHSSMQTNARDTQEAMQSHAREAQEAMQSHARETQEAISVVQSNVTEEISAIQSNITEEISAIQSNITEEISTVRSDV

REEISAVYSNVREVLTEVVTRIERLEESPVPRSVVGLNPGLRSPASIMADVPHQSTIRLAESSGARPKDFTHLGTLRRSERLAHKDPISYRELGSRDG

WQSFHRLDLNPRVPSSYPRSISQDHDHDPQQHAASLPYDDPELATPGMTVGAGVTIGTRSGVLVQIGSRDFGDDSLEEEADDVGRGELAMSFERRMER

KETERRLREFADNVEIMSDEGSVDIEISRMMEARPKTVKIIDNRMQTALQPEFRDLEISKPIANQLPREGGDKANPVPLAFSSLPHVEYPRPNPQWRM

QASLADVGRQRAETATNLGDILQSTSGRPTYRSMADSAFDGRTTSTLVGRGYARPDLRPAPVVMATVDRMQPYVLPDHSYLSRMSAPTHAPSAFYENI

DNQVWLDSFRPSDDIFPRWPPIVESARMSSCLGAQSFGFGRDVELPRSSLLKPTLPVRIETTGVDKTPASSNVSVGLIRTVPMTTVATQTTGDLELPS

TSLGGTAKEPEPTASATPKVGEVKPSSTPAPKPKQWLKLGSYSGKTDVEIFLRRFSVCAKNNGWSEEEKLNQLVVALVEPATNLLSETNADSLDTWTA

LVQRLRERYGNAEQQALFQTQLSTRKQKPDEDMGSVVDDIRRLTSRAYPGSSTVHSEAIAVRAFLDALRDRTLALKIREREAKSLDEAYKVAMRLDGY

QKAEAGGHEQHERRQGRINAIKEEDAKVKDSELKDLRWELEQMKRQVERQPLAAPQVRYHNSGPWRGQNGSTGNGWYNANGNNGNWRNNQRGYYDYRP

EHWNSAGGYRQAVRPEAGRSGRRCYECGQVTDHIARFCPLRQQQAARNDNQGQPGSDTPCNPIVGRVNAVQSSAPAPSGNQGNQMLPSSTRRRDDGRN

SPGEERRCFYCDSRFHLLRGCPLRNGSSDDHNWRTLEEPTGSHPSGTWHHMGIYP]

>CDS    2414960 2415253 CDS 2417218 2417280

protein sequence = [MNALCTVERRVTSQCTQDRTAGRDPKEEKKNNSNNNNINNNNNNDNDNESDDDDDDNDNDNDNDNDYNNDNDNDYDNS

SLIPLAVETMDPINKDDLAFVSEIRRRLTKLSGYLRGKDF]

>CDS    2478928 2483715
```

Onnodige tekens verwijderen

```
with open('proteinsdb.fa', 'r') as infile, open('temp.fa', 'w') as outfile:

     temp = infile.read().replace("protein sequence = [", "")

     outfile.write(temp)
```

Het bestand is gereed om blasten uit te voeren, nu de extra tekens zijn weggehaald. Elke eiwit heeft één of meer cds en de coördinaten daarvan worden gebruikt als unieke identificatiecode.

```
>CDS    2007959 2008723
MEESRPVTPAQPSRPPSSMEILLEAIQTNARSTHEAIQTNAKSSQEAMQAHAKSTHDAMTSIQSSLQLNARETQEAIA
TVEFNVLAVQSNVSEAISSVQSNVREEIREEISAVRDNVREALTEMVSRLERLEASPVPKPAVDSNPGYLTAITPADAPYHSTIGLGETLGARPKDFT
QPGILRRSDRLAGRPPISYREYGSRKDWPPFLGWDSNPEVTSSCPPSISRACPQQHAVPSGEDPEVATPGRPIGAGVT
>CDS    2035683 2035737 CDS 2036822 2037021
MTNLNSGLCPRSFTGRLPSKKTNNIPTQSCYNGMTFNINNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNKAHL
>CDS    2087020 2087265
MTCLRKIEGVTRKDRIRNAVIYDRLNIKQDIVDKVRNRRMRYFGHVTRMGNEIYPKIGSTDMYMGKDQGETKEKMDRH
DKR
>CDS    2108840 2109797 CDS 2112970 2113319
MGTSHVTATMATDVPLEEQFFYPGARELRAFHDWYIDVHGYLATVVCVFGIVANLLNIVVLTRRNMISPTNCILTALA
LSDGLTMLAYLPFALRFYVLYGTGVSPERNSLEAIRFMLYVFILYPRHGSIWCSLRRSKLAVFIVNVVTIIVCIPNFVTIRVQGSRSDPQLHDNNHQI
NQNYHLHQHTNHSSGNHSIGNHSSGNHSIGNHSSGNHSSGNHSSGNHSSGNHSSGTHSISPLIWIVEFKEDSDRDILVKNLNFWIQALLVKLVPCVGL
TILSLLLVRTMKVAEQRRLKLRCNNPGNTTTDTGRRSKEETSSRDRKTNRTTRMLLVVVVLFLITEFPQGIINLLSGIASTNTNFVIEVYMTLGDLLD
ILALINNGINFILYCTMSKQFRDTFVCVFLPSCGARTASRGGVAAGGGGGMVTTTTATRTTVE
>CDS    2205956 2206438
```

Het bestand dat alle eerder genoemde stappen bevat, is te vinden in protocol2.2, in de map scripts, met de naam parse-proteins.py. De database van de aangemaakte proteïnen vind je in identification-p onder prediction, en heet proteinsdb.fa.

6.1 Gen-identificatie (proteine niveau)

Blast wordt uitgevoerd met deze query, waarbij er maximaal één eiwit per fractie wordt gebruikt. Deze query werkt op dezelfde manier als de Standard Protein BLAST web-versie die je gebruikt om vergelijkbare eiwitten in de database te blasten. Er is ook een gedetailleerde uitleg over de functie van qblast, in het hoofdstuk hieronder, nucleotide identificatie.

```
from Bio.Blast import NCBIWWW
from Bio.Blast import NCBIXML


genomic="blast/fraction1-proteins.fa"
sequence_data = open(genomic).read()
result_handle = NCBIWWW.qblast("blastp", "nr", sequence_data, hitlist_size=5, alignments=50)
result_handle
```

De volledige query staat beschreven in het bestand blast-p.py, bij de protocol2.2-> scripts. Blast doet verschillende alignments, identificeert eiwitten en beoordeelt ze, met aandacht voor een Hsp_evalue, vergelijkbare p-waarde. We hebben vooral belangstelling voor deze drie waarden: Hit_id, Hit_def en Hsp_evalue.

```
<Hit>
  <Hit_num>1</Hit_num>
  <Hit_id>ref|XP_042643872.1|</Hit_id>
  <Hit_def>kinesin-like protein KIF11 isoform X1 [Tyto alba]</Hit_def>
  <Hit_accession>XP_042643872</Hit_accession>
  <Hit_len>1175</Hit_len>
  <Hit_hsps>
    <Hsp>
      <Hsp_num>1</Hsp_num>
      <Hsp_bit-score>48.521</Hsp_bit-score>
      <Hsp_score>114</Hsp_score>
      <Hsp_evalue>0.049464</Hsp_evalue>
```

```
      <Hsp_query-from>23</Hsp_query-from>
      <Hsp_query-to>116</Hsp_query-to>
      <Hsp_hit-from>691</Hsp_hit-from>
      <Hsp_hit-to>784</Hsp_hit-to>
      <Hsp_query-frame>0</Hsp_query-frame>
      <Hsp_hit-frame>0</Hsp_hit-frame>
      <Hsp_identity>31</Hsp_identity>
      <Hsp_positive>54</Hsp_positive>
      <Hsp_gaps>8</Hsp_gaps>
      <Hsp_align-len>98</Hsp_align-len>
      <Hsp_qseq>LEAIQTNARSTHEAIQTNAKSSQEAMQ----AHAKSTHDAMTSIQSSLQLNARETQEAIATVEFNVLAVQSNVSEAISSVQSNVREEIREEISA
      <Hsp_hseq>LKKLQEETTSVFAQLQNDCENLKEEVEMTRLAHTKSTAELMSSLQSQLDLFARETQKNLT----NVLTKNGSLKTAITAVQENIHLKTTDLVSS
      <Hsp_midline>L+ +Q    S    +Q + ++ +E ++    AH KST + M+S+QS L L ARETQ+ +     NVL    ++  AI++VQ N+  +  + +
    </Hsp>
  </Hit_hsps>
</Hit>
```


Alle xml-bestanden per CDS bevinden zich in de map xml, identificatie-p, met een naam die unieke CDS-coördinaten bevat zoals CDS2007959-2008723.xml. De details over het parsen van het xml-resultaat zijn te vinden in het hoofdstuk dat gaat over nucleotidenidentificatie.

We zijn vooral geïnteresseerd in het vergelijken van twee anotaties van verschillende bronnen, eiwit en mRna. En om te zien of er overlap is als er overlappingen zijn.

In de genomische browser Jbrowser zijn overlappende genstructuren weergegeven, zoals te zien is in de screenshots die zijn gekozen voor vergelijking. De bovenste track laat annotaties zien die afkomstig zijn uit de mRNA-pijplijn, en de onderste track toont annotaties uit de proteïnepijplijn.

G5 Toont 2 CDS overeenkomsten, groen en blauw.

inzoomen:

https://raw.githubusercontent.com/alenagrrr3/OX457036.1.html/refs/heads/main/intersectiong5.svg

Informatie over dit gen:

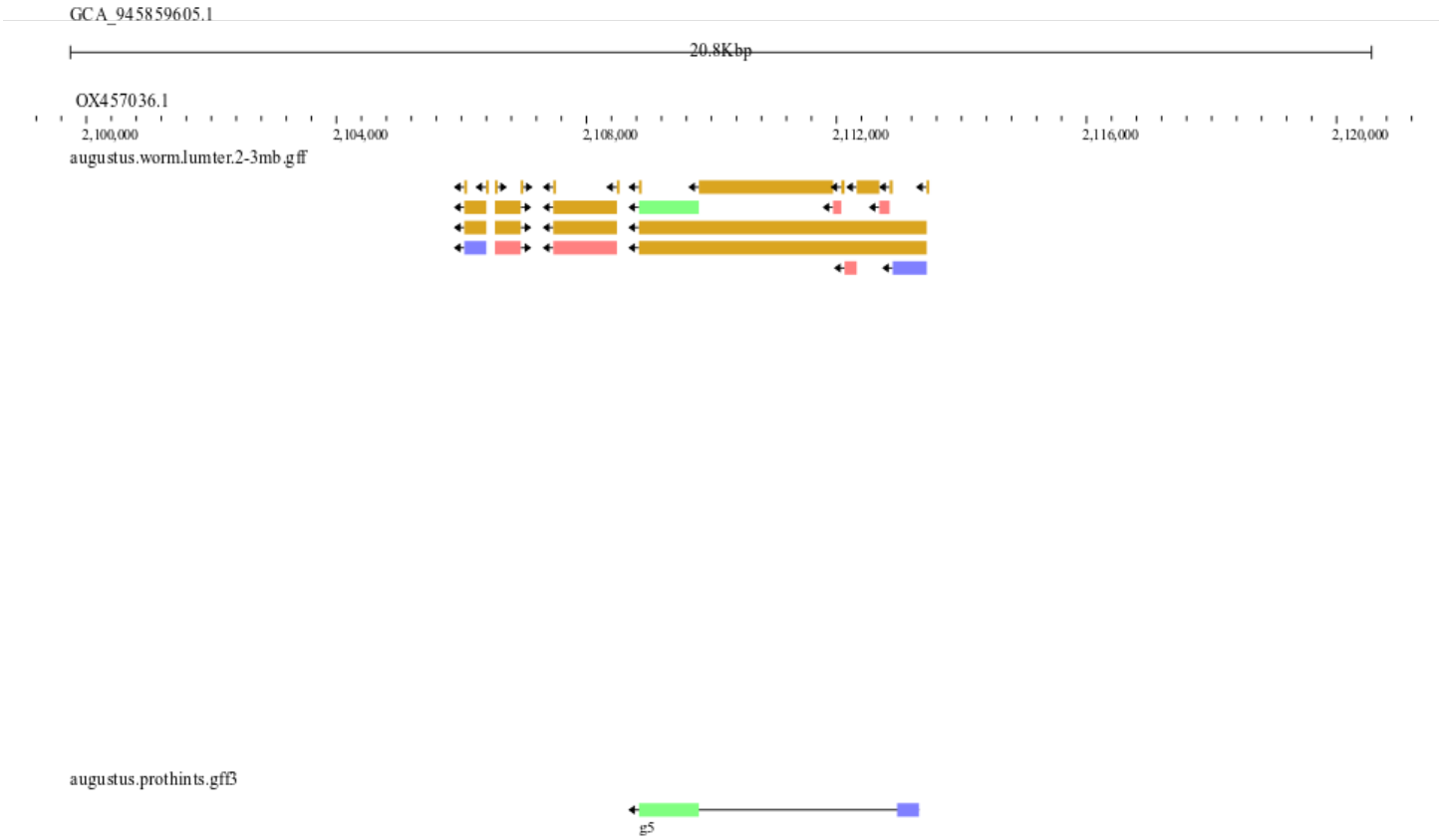

GCA_945859605.1
20.8Kbp
OX457036.1
2,100,000
2,104,000
2,108,000
2,112,000
2,116,000
2,120,000
augustus.worm.lumter.2-3mb.gff
augustus.prothints.gff3
g5

Figure 2: G5.

g5,coördinaten OX457036.1:2108840-2109808, in de nucleotidenlijn werd er een voorspelling voor verkregen : Candidozyma auris strain BA03 chromosome; 1 eval; CP157508.1 in de eiwitlijn, werd er een voorspelling voor verkregen: endonuclease-reverse transcriptase, GFN75565.1

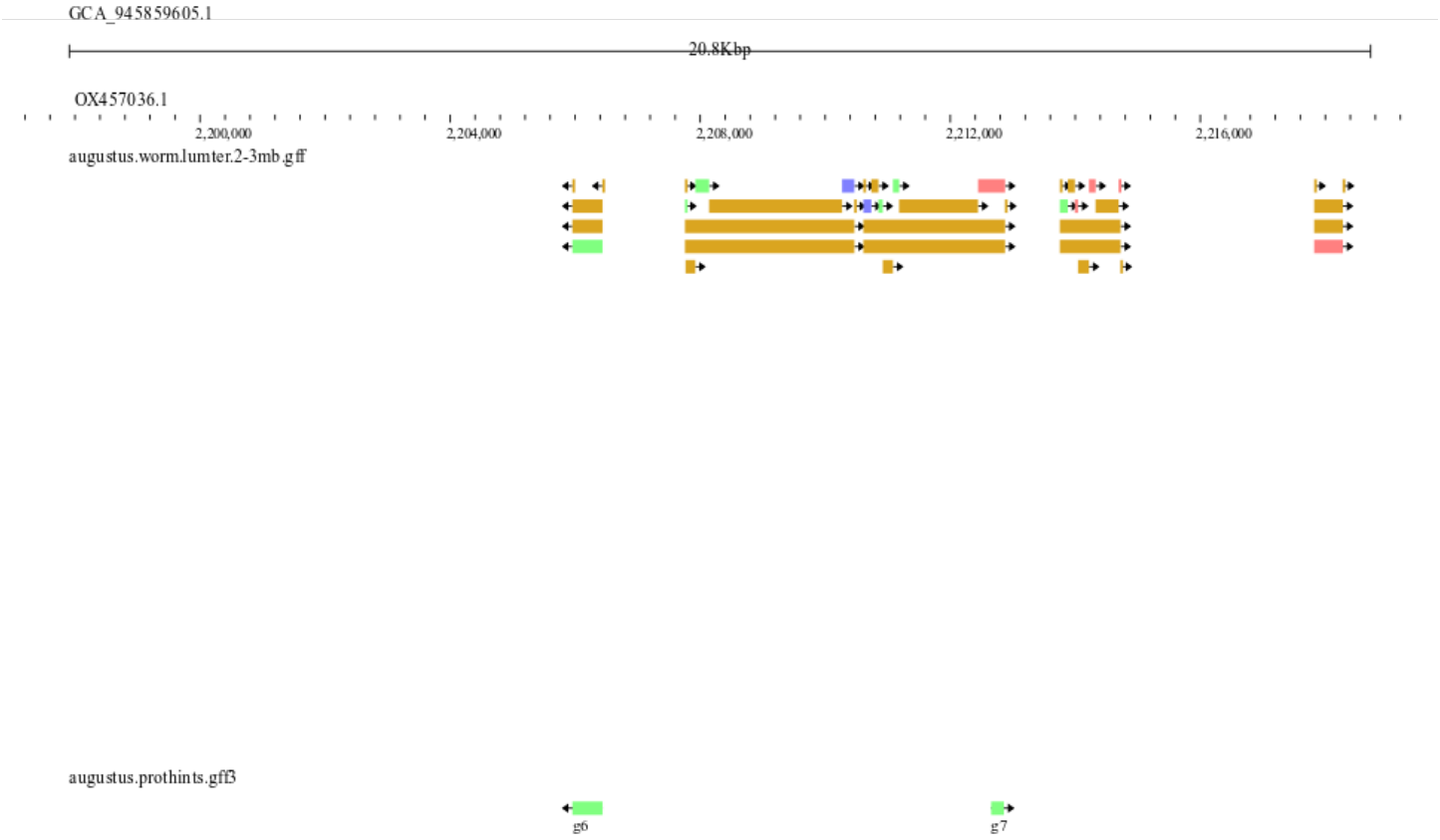


Figure 3: G6-G7.

g6 coördinaten OX457036.1:2,205,956-2,206,438(-) ,voor dit gen geen Hits gevonden. CDS , in groen, werd geïdentificeerd in zowel nucleotde pipline als proteine pipline G7 CDS is geïdentificeerd als bifunctional 4-hydroxy-2-oxoglutarate aldolase, WP_117560622.1

inzoomen:

https://raw.githubusercontent.com/alenagrrr3/OX457036.1.html/refs/heads/main/lumterAM182481.1g6-7.svg

G8, G9: Twee CDS g8, en g9, in blauwe kleur, op beide annotaties

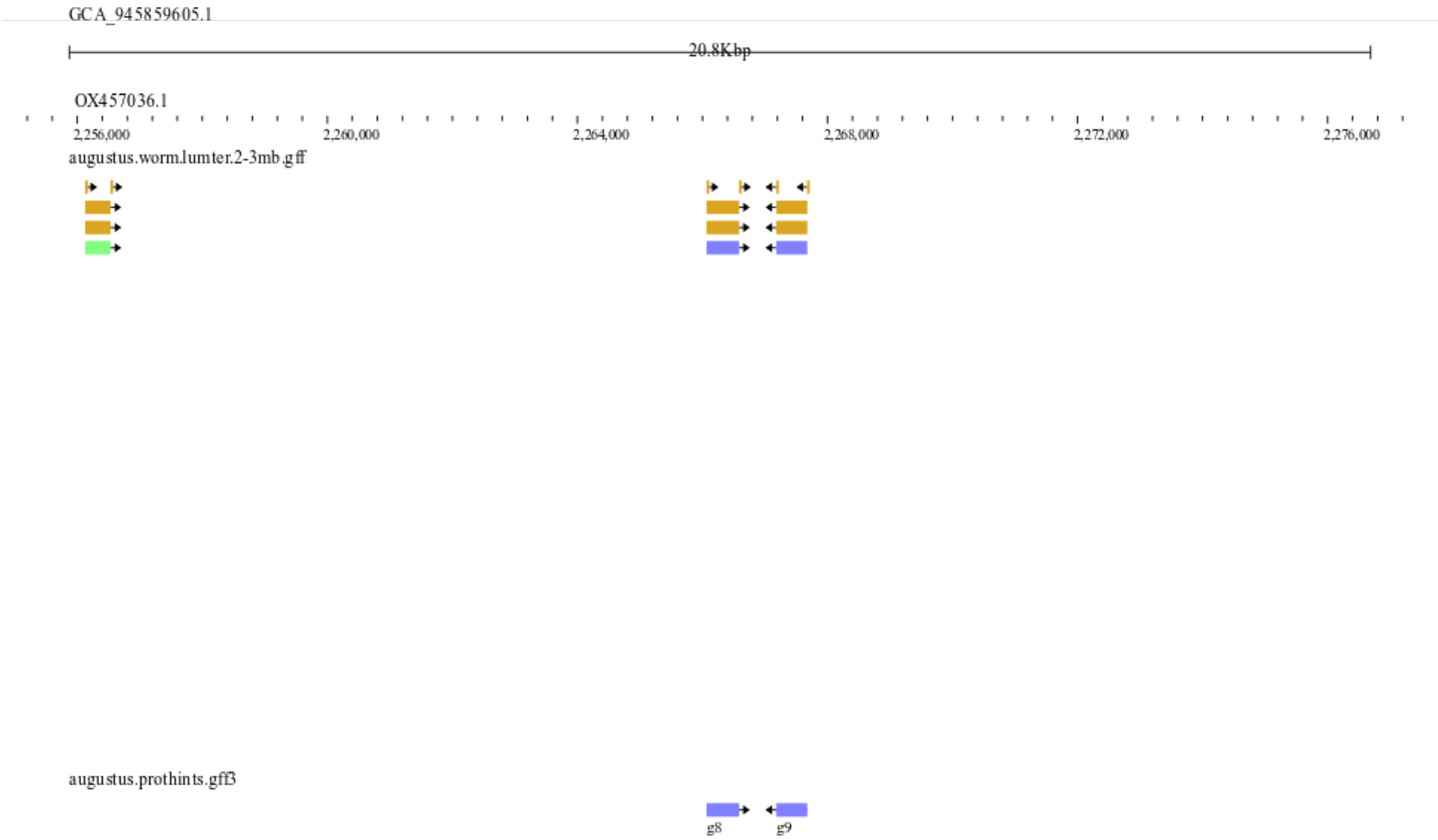


Figure 4: G8-G9.

Voor g8 zijn er geen its op nucleotide niveau gevonden. Op de proteine niveau BLAST toonde dit resultaat voor g8 emb CAI2738958.1 naamloos eiwitproduct, CDS coördinaten 2266065-2266586

Voor g9 is deze prediction gevonden op nucleotide niveau:Earthworm (L.terrestris) extracellular globin chain c gene, complete cds, Query ID lcl|Query_2713981. Bij de proteïne-identificatie werd het volgende resultaat verkregen: MAG,hypothetical protein DMG74_22350

inzoomen:

https://raw.githubusercontent.com/alenagrrr3/OX457036.1.html/refs/heads/main/g8g9.svg

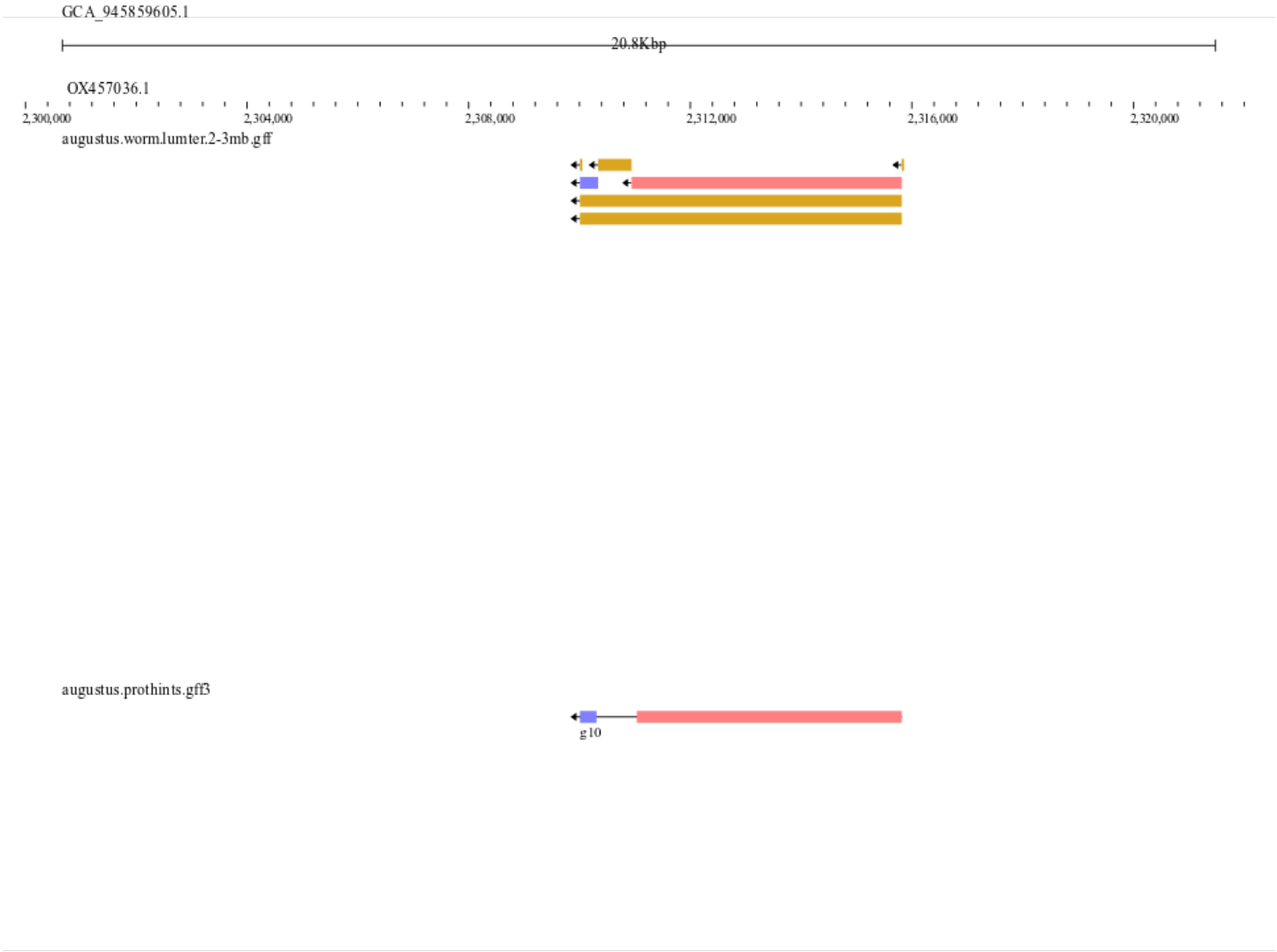
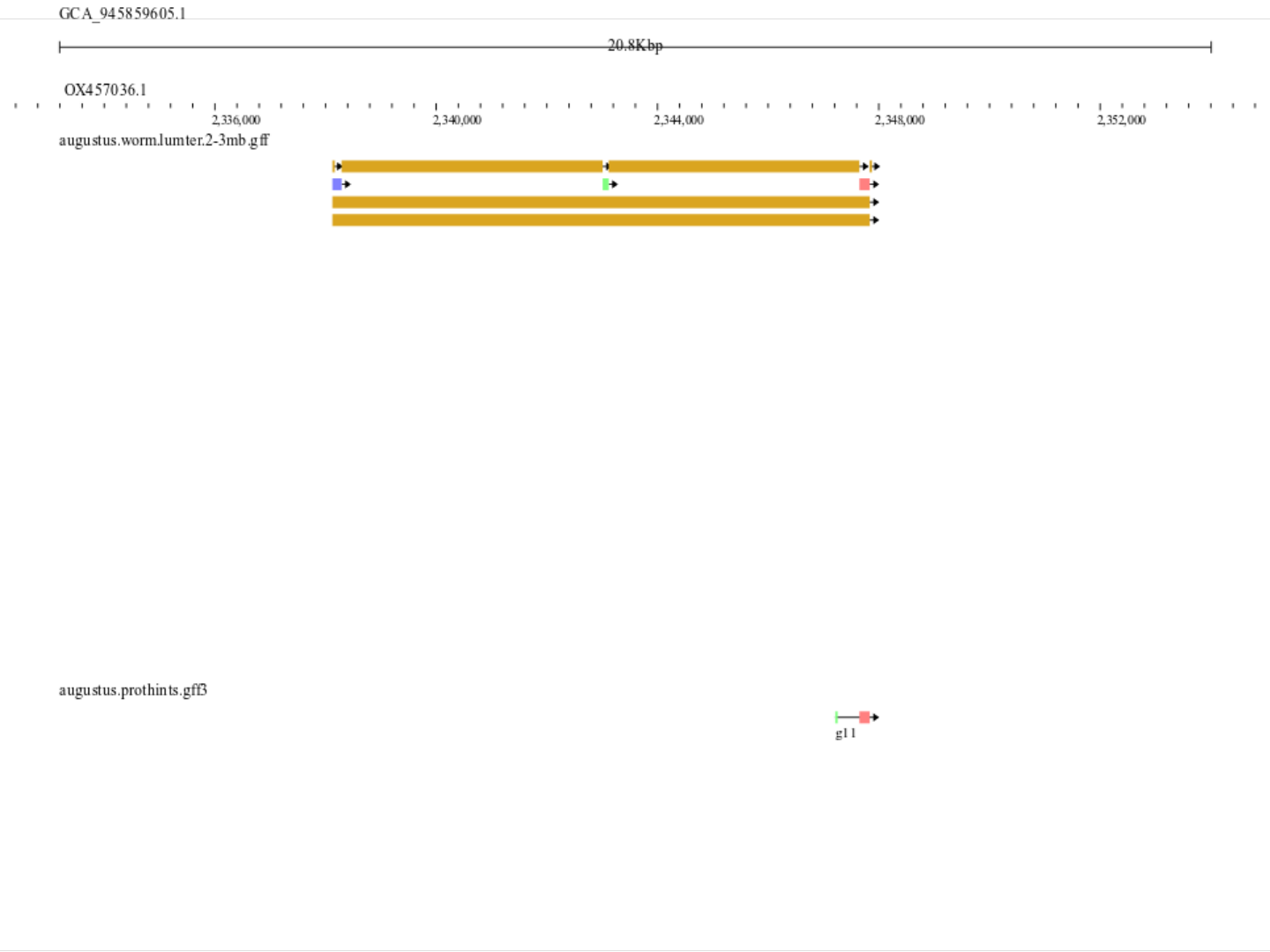


Figure 5: G10.

Hier zijn de 2 CDS van dit gen roze en blauw, die overeenkomen met de bovenste en onderste annotatie. Voor g10 is er geen significante overeenkomst gevonden op nucleotide niveau. Bij de proteine niveau, voor gen10: uncharacterized protein XP_038057353.1

inzoomen:

https://raw.githubusercontent.com/alenagrrr3/OX457036.1.html/refs/heads/main/g10.svg



G11

Hier afgebeeld zijn cds, in roze op de bovenste en onderste annotatie, de identiteit van dit gen is onbekend

G12,G13,G14

Genen 12, 13 en 14 hebben een volledige structurele overlap

Ondanks de volledige overlap in genomische structuur, is de identificatie van deze genen onbekend

G13: gb KAG8287034.1 hypothetical protein J6590_047011

G13 : Pipistrellus nathusii genome assembly, Query_3407871

G14: hypothetical protein protein emb|VDI23710.1

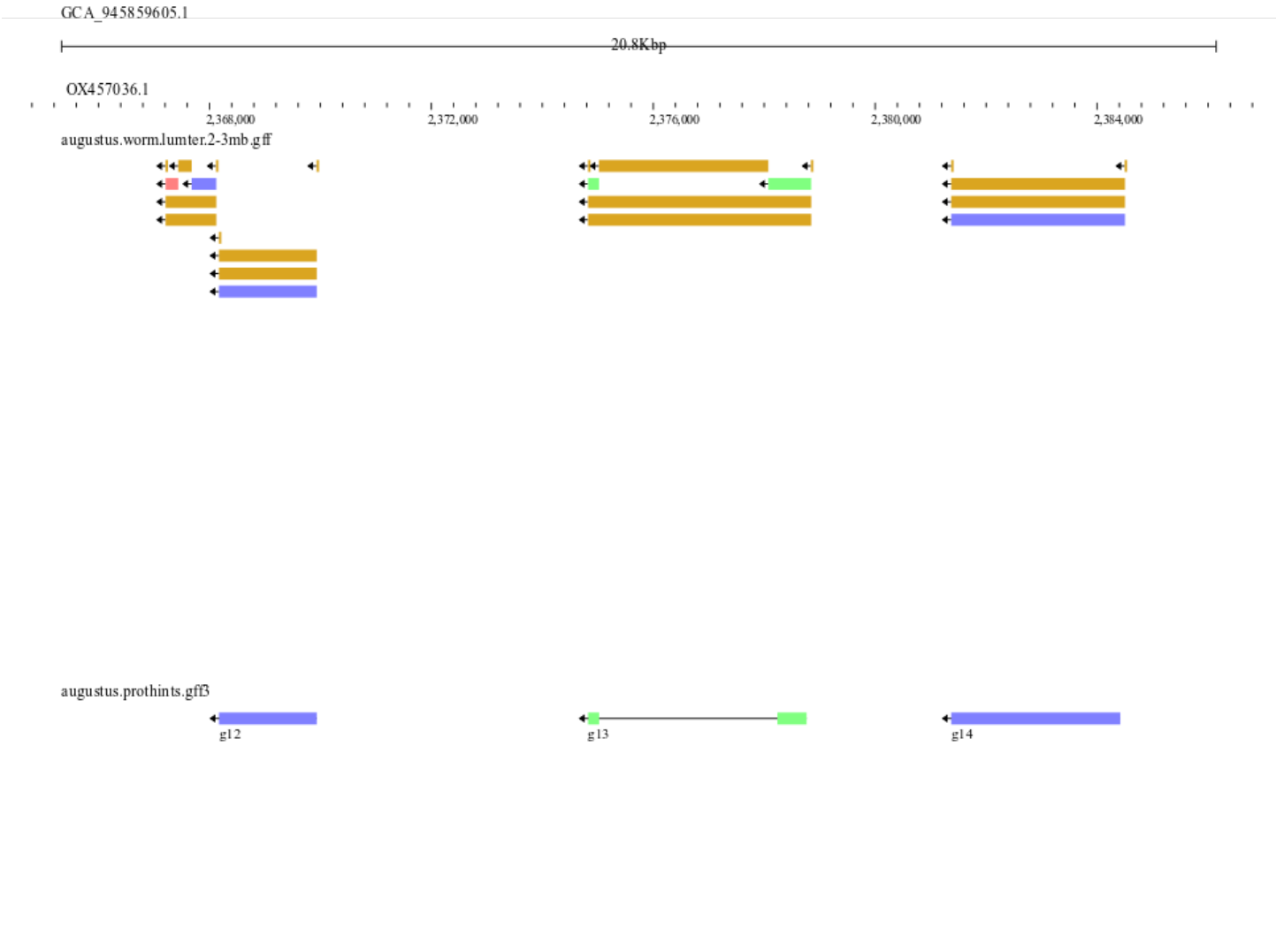


Figure 6: G12G13G14.

G14: Argiope bruennichi uncharacterized LOC129975135 XM_056088086.1

inzoomen

https://raw.githubusercontent.com/alenagrrr3/OX457036.1.html/refs/heads/main/g12g13g14.svg

G16: predicted: Gadus chalcogrammus uncharacterized LOC130405754 (LOC130405754), transcript variant X1, mRNA;XM_056610933.1; OX457036.1:2478928-2483715 G16: hypothetical protein FSP39_018360

inzoomen:

https://raw.githubusercontent.com/alenagrrr3/OX457036.1.html/refs/heads/main/g16.svg

Voor extra zoekopdrachten naar overeenkomsten in de genomische structuur van deze twee pipilinen zijn de extracten van de genomische browser beschikbaar:

https://raw.githubusercontent.com/alenagrrr3/OX457036.1.html/refs/heads/main/g17g18g19.svg

https://raw.githubusercontent.com/alenagrrr3/OX457036.1.html/refs/heads/main/g20-g23.svg

https://raw.githubusercontent.com/alenagrrr3/OX457036.1.html/refs/heads/main/g24g25.svg

https://raw.githubusercontent.com/alenagrrr3/OX457036.1.html/refs/heads/main/g26.svg

https://raw.githubusercontent.com/alenagrrr3/OX457036.1.html/refs/heads/main/g27.svg

https://raw.githubusercontent.com/alenagrrr3/OX457036.1.html/refs/heads/main/g28g29.svg

https://raw.githubusercontent.com/alenagrrr3/OX457036.1.html/refs/heads/main/g30.svg

https://raw.githubusercontent.com/alenagrrr3/OX457036.1.html/refs/heads/main/g31.svg

https://raw.githubusercontent.com/alenagrrr3/OX457036.1.html/refs/heads/main/g32g33.svg

https://raw.githubusercontent.com/alenagrrr3/OX457036.1.html/refs/heads/main/g34.svg

https://raw.githubusercontent.com/alenagrrr3/OX457036.1.html/refs/heads/main/g35g36.svg

https://raw.githubusercontent.com/alenagrrr3/OX457036.1.html/refs/heads/main/g37.svg

https://raw.githubusercontent.com/alenagrrr3/OX457036.1.html/refs/heads/main/g38.svg

Voor zowel de eiwitlijn als de nucleotidenlijn geldt dat er in 1 megabase ongeveer 30 genen te vinden zijn, wat gelijkstaat aan 1% van het chromosoom. Dit betekent dat er ongeveer 2000-3000 genen op elk chromosoom aanwezig zijn.

6.2 Gen-identificatie (nucleotide niveau)

Dit hoofdstuk bespreekt voorspelling op nucleotidenniveau. Alle voorspellingen zijn gebaseerd op een DNA-fragment van 1 mb. De exacte locatie is aangeduid als 2000000-3000000. (2-3 mb) van chromosoom 1. De predictor is toegepast op het eigen lumtermodel (zie, protocol 1, model) dat in deel 1 is ontwikkeld. Alle stappen voor identificatie zijn vastgelegd in prediction.xlsx (map identification-n).

```
augustus --species=lumter lumter.fasta --predictionStart=2000000 --predictionEnd=3000000 --gff3=on
```

Voor het identificeren van genen hebben we de qblast() functie gebruikt uit de Bio.Blast.NCBIWWW module van Biopython. De qblast functie heeft verschillende opties die vergelijkbaar zijn met de parameters die je kunt instellen op de BLAST webpagina. Wij hebben nucleotide blast ("blastn", "nt") gebruikt. Deze functie is bedoeld om nucleotidesequenties te vinden die vergelijkbaar zijn met die van andere organismen, en deze gegevens zijn beschikbaar in de NCBI-database. Hulp voor de qblast functie:

```
from Bio.Blast import NCBIWWW

help(NCBIWWW.qblast)
```

Some useful parameters:

```
 - program        blastn, blastp, blastx, tblastn, or tblastx (lower case)
 - database       Which database to search against (e.g. "nr").
 - sequence       The sequence to search.
 - ncbi_gi        TRUE/FALSE whether to give 'gi' identifier.
 - descriptions   Number of descriptions to show.  Def 500.
 - alignments     Number of alignments to show.  Def 500.
 - expect         An expect value cutoff.  Def 10.0.
 - matrix_name    Specify an alt. matrix (PAM30, PAM70, BLOSUM80, BLOSUM45).
 - filter         "none" turns off filtering.  Default no filtering
 - format_type    "HTML", "Text", "ASN.1", or "XML".  Def. "XML".
 - entrez_query   Entrez query to limit Blast search
 - hitlist_size   Number of hits to return. Default 50
 - megablast      TRUE/FALSE whether to use MEga BLAST algorithm (blastn only)
```

```
 - short_query    TRUE/FALSE whether to adjust the search parameters for a
                  short query sequence. Note that this will override
                  manually set parameters like word size and e value. Turns
                  off when sequence length is > 30 residues. Default: None.
 - service        plain, psi, phi, rpsblast, megablast (lower case)

This function does no checking of the validity of the parameters
and passes the values to the server as is.  More help is available at:
https://ncbi.github.io/blast-cloud/dev/api.html
</p>
```

Eerst hebben we het ruwe GFF-bestand voorbereid voor de Blast API door alle spaties en het '#' symbool te verwijderen. Om de gencoördinaten te krijgen, maakten we gebruik van een regex-patroon.

```
pattern_a = r'gene.*\s+(OX457036.*AUGUSTUS\sgene.*g\d+)'
```

Voor het ophalen van de coderingssequentie uit het GFF-bestand maakten we gebruik van een andere regex.

```
pattern_b = r"coding sequence =.*[actg\s\]]{1,}".
```

Nadat je het GFF-bestand hebt geparsed, is het klaar voor gebruik met de Blast API. Elke coderingssequentie heeft een unieke identificatie die de start- en eindcoördinaten bevat: genomisch OX457036.1:2000789-2003917

Voor meer details kun je de scripts bekijken, vooral scripts->parse-nucleotide.py, deel identification-n.

```
head lumbricus/identification-n/prediciton/genome.fa.gff
```

```
## bash: cannot set terminal process group (2732574): Inappropriate ioctl for device
## bash: no job control in this shell
## >genomic OX457036.1:2000789-2003917
## atggaggagtctaggccagtcactcccgctcagccttctaggcccccttcttctatggagatattgctcgaggcaatac
## aaactaatgctaggtccactcatgaagcaatacagactaacgctaagtcttcacaagaggctatgcaagcgcatgctaagtcaactcatgatgctatg
## acttctatacagtcgtctttgcaactgaatgccagagagacgcaagaggcgattgccacggtggagtttaatgtcctggcagtgcaatcaaatgttag
```

```
## cgaagctatttcctcagtgcaatcaaatgtaagagaggagataagagaagagatctcggctgtaagagataatgtcagggaagcgctgacggaaatgg

## tatcacgattggaaaggctagaggcgtcgccggtacccaagcctgctgtggattcgaaccctggttacctcaccgctattacccctgcggacgcgcca

## taccactcgaccatcggcctgggggaaactttgggtgctaggcctaaagatttcacgcaacctggtatattgcggagaagtgatagattggctggtag

## gccgccaatttcatatagggagtacggtagtcgaaaagactggccgcctttcctgggttgggattcgaacccagaagtcacctcctcctgtcctccct

## ctatctctagagctcgtccacagcagcacgcggtcccatcaggcgaggatccggaagtggcgactccggggatgccgataggggcgggcgttacaatt

## ggtcccagccagtggggtcaaattagttctagagattttggtgatgataggttagaagaggaaactgactatgctagaacaggcgaaatggcaatttc
```

De blast-query's via Bio.Blast.NCBIWWW.qblast zijn uitgevoerd en de resultaten zijn teruggegeven in XML-formaat (voor meer informatie, zie: blast.py).

```
from Bio.Blast import NCBIWWW
from Bio.Blast import NCBIXML



genomic="genome.fa"


sequence_data = open(genomic).read()


sequence_data


result_handle = NCBIWWW.qblast("blastn", "nt", sequence_data, hitlist_size=5, alignments=50)



with open('reults.xml', 'w') as save_file:
   blast_results = result_handle.read()
   save_file.write(blast_results)
```

Voor de blast-analyse is het bestand genome.fa opgedeeld in drie verschillende fracties, wat resulteerde in 3 xml-bestanden (identificatie->xml). Elke DNA-sequentie die je invoert in nucleotide BLAST krijgt een bepaald aantal hits, en het geeft ook wat statistieken over die hits.

Een voorbeeld van een hit: .

```
<Iteration_hits>
<Hit>
  <Hit_num>1</Hit_num>
  <Hit_id>gi|11071239|emb|AJ299434.1|</Hit_id>
  <Hit_def>Lumbricus rubellus mt2A gene for metallothionein 2A, exons 1-4</Hit_def>
  <Hit_accession>AJ299434</Hit_accession>
  <Hit_len>7302</Hit_len>
  <Hit_hsps>
    <Hsp>
      <Hsp_num>1</Hsp_num>
      <Hsp_bit-score>85.143</Hsp_bit-score>
      <Hsp_score>93</Hsp_score>
      <Hsp_evalue>7.19655e-12</Hsp_evalue>
      <Hsp_query-from>70</Hsp_query-from>
      <Hsp_query-to>246</Hsp_query-to>
      <Hsp_hit-from>306</Hsp_hit-from>
      <Hsp_hit-to>490</Hsp_hit-to>
      <Hsp_query-frame>1</Hsp_query-frame>
      <Hsp_hit-frame>1</Hsp_hit-frame>
      <Hsp_identity>131</Hsp_identity>
      <Hsp_positive>131</Hsp_positive>
      <Hsp_gaps>8</Hsp_gaps>
      <Hsp_align-len>185</Hsp_align-len>
      <Hsp_qseq>AGATTGAACATCAAACAGGATATAGTTGACAAAGTGCGGAATAGAAGAATGCGATACTTTGGACATGTGA------CAAGAATGGGGAACGAAA
      <Hsp_hseq>AGACTGAATATTCAACATGATATAATACACAAGATCCAAAGTAAACGACTACGCTACTTTGGCCACGTATATATATCCAGAATGAGGGATGAGA
      <Hsp_midline>||| |||| ||  |||| |||||| |  |||| | |  | || | || | || |||||||| || ||        | |||||| || | |
    </Hsp>
  </Hit_hsps>
</Hit>
```

De XML-resultaten van de blast-uitvoer laten zien hoe goed de Alignment overeenkomt, samen met de eval-waarde. De gevonden

Hits worden bewaard met het NCBI-referentienummer, zoals “ref XM_003731435.1”, of het Ensemble-referentienummer, zoals “emb OE003277.1”. Zodra je de XML-resultaten hebt, is de eerste stap om ze te parseren. De XML-resultaten zijn geparsed en gesorteerd op coördinaten en e-waarde (sort-blast-by-coords.py, sort-blast-by-pval.py).

```
import os

cwd = os.getcwd()

print(cwd)



import sys

from Bio.Blast import NCBIXML

OUT = open("sorted_by_coordinates.fraction3.txt", 'w')

OUT.write("Query Name\tQuery Length\tAlignment ID NCBI\teValue\n")

result_handle = open("blast.results.fraction3.xml")

blast_records = NCBIXML.parse(result_handle)

for rec in blast_records:

        for alignment in rec.alignments:

            for hsp in alignment.hsps:

                fields = [rec.query_id, rec.query[:100], str(rec.query_length), alignment.hit_id,

                          alignment.accession, str(hsp.expect)]

                OUT.write("\t".join(fields) + "\n")

OUT.close()

print('Done')
```

```
sorted_by_coordinate <- read_excel("lumbricus/identification-n/prediction.xlsx", sheet = 6 )



sorted_by_p <- read_excel("lumbricus/identification-n/prediction.xlsx", sheet = 5 )



# sorted by coordinates

head(sorted_by_coordinate )
```

```
## # A tibble: 6 x 6

##   `Query Name`  `Query Length`        `Alignment ID NCBI` eValue Column1    `_1`
```

```
##   <chr>         <chr>                              <dbl> <chr>  <chr>     <dbl>
## 1 Query_1234140 genomic OX457036.1:2~                459 gi|26~ XM_063~ 6.18e-5
## 2 Query_1234140 genomic OX457036.1:2~                459 gi|26~ XM_063~ 3.20e-2
## 3 Query_1234140 genomic OX457036.1:2~                459 gi|26~ XM_062~ 4.75e-1
## 4 Query_1234140 genomic OX457036.1:2~                459 gi|26~ XM_062~ 4.75e-1
## 5 Query_1234140 genomic OX457036.1:2~                459 gi|26~ XM_062~ 4.75e-1
## 6 Query_1234141 genomic OX457036.1:2~                408 gi|28~ OZ0783~ 4.46e-6
```

```
# sorted by p-val
head(sorted_by_p)
```

```
## # A tibble: 6 x 2
##   Column1 Column2
##   <chr>   <chr>
## 1 <NA>    <NA>
## 2 query:  genomic OX457036.1:2108840-2109808
## 3 match:  gi|2739567124|gb|CP157508.1| Candidozyma auris strain BA03 chromosome~
## 4 query:  genomic OX457036.1:2108840-2109808
## 5 match:  gi|2739567124|gb|CP157508.1| Candidozyma auris strain BA03 chromosome~
## 6 query:  genomic OX457036.1:2108840-2109808
```

Eerst moeten we naar alle voorspellingen kijken, ook naar de voorspellingen met ongunstige eval-waarden (vergelijkbaar met p-waarden). Alle voorspellingen: .

```
all_predictions <- read_excel("lumbricus/identification-n/prediction.xlsx", sheet = 1 )


all_predictions
```

```
## # A tibble: 89 x 5
##    `OX457036.1:2000789-2003917` AUGUSTUS gene  predicted:not satisfac~1 `185403`
##    <chr>                        <chr>    <chr> <chr>                   <chr>
##  1 OX457036.1:2007959-2008723   AUGUSTUS gene  predicted:not satisfact~ 1852
```

```
##  2 OX457036.1:2039309-2039692   AUGUSTUS gene  predicted:not satisfact~ 881419
##  3 OX457036.1:2062296-2062562   AUGUSTUS gene  predicted:not satisfact~ 0
##  4 OX457036.1:2087020-2087265   AUGUSTUS gene  predicted: Lumbricus ru~ 7.38114~
##  5 OX457036.1:2089471-2089899   AUGUSTUS gene  predicted:Lampetra plan~ 8.69409~
##  6 OX457036.1:2090721-2091137   AUGUSTUS gene  predicted:not satisfact~ 965729
##  7 OX457036.1:2106048-210639    AUGUSTUS gene  predicted:Mus musculus ~ 6.12276~
##  8 OX457036.1:2106538-2106948   AUGUSTUS gene  predicted:not satisfact~ 640374
##  9 OX457036.1:2107471-2108487   AUGUSTUS gene  predicted:not satisfact~ 3.24628~
## 10 OX457036.1:2108840-2109808   AUGUSTUS gene  predicted: Candidozyma ~ 1.27857~
## # i 79 more rows
## # i abbreviated name: 1: `predicted:not satisfactory p-value`
```

In deze fase hadden we voorspellingen(Hits) voor 92 genen op een 1mb chromosoom (tussen 2mb en 3mb), zelfs met enkele genen die niet zo'n goede eval-waarden hadden.

```
colnames(all_predictions ) <- c("id","source","feature", "predicted", "eval")
```

```
all_predictions $eval <- parse_number(all_predictions $eval)
```

```
df.f.pavlue <-  all_predictions %>% filter(eval<= 1e-4) %>%  filter(eval!=0)
head(df.f.pavlue)
```

```
## # A tibble: 6 x 5
##   id                          source   feature predicted                     eval
##   <chr>                       <chr>    <chr>   <chr>                        <dbl>
## 1 OX457036.1:2087020-2087265 AUGUSTUS gene    predicted: Lumbricus rub~ 7.38e- 7
## 2 OX457036.1:2089471-2089899 AUGUSTUS gene    predicted:Lampetra plane~ 8.69e-99
## 3 OX457036.1:2106048-210639  AUGUSTUS gene    predicted:Mus musculus c~ 6.12e-10
## 4 OX457036.1:2108840-2109808 AUGUSTUS gene    predicted: Candidozyma a~ 1.28e-22
## 5 OX457036.1:2108840-2109808 AUGUSTUS gene    predictied: Phaeodactylu~ 2.82e-13
## 6 OX457036.1:2112894-2113442 AUGUSTUS gene    predicted:  Ixodes scapu~ 1.27e-14
```

```
write.table( df.f.pavlue, +

                "lumbricus/identification-n/prediction/df.filtered.txt",sep="\t")
```

```
predictions <-read.table("lumbricus/identification-n/prediciton/df.filtered.txt")
```

In de daaropvolgende fase hebben we een eval, evaluatiedrempel van 1e-4 ingesteld, wat redelijk mild is.Na het filteren van de voorspellingen met ongunstige eval-waarden, hebben we 32 voorspellingen gevonden die betrekking hebben op 32 genen voor een 1 Mb segment van het eerste chromosoom, wat 1% van het totale chromosoom is. De uiteindelijke voorspelling voor het fragment dat we onderzoeken, is als volgt.

predicition:

```
table7 <-  predictions   %>%  select(V7)

table7 %>%

  kable("html") %>%

  kable_styling(font_size = 7)
```

V7

predicted: Lumbricus rubellus mt2A gene for metallothionein 2A, exons 1-4;AJ299434.1;

predicted:Lampetra planeri genome assembly, chromosome: 62; emb OZ078387.2

predicted:Mus musculus chromosome 8, clone RP23-339I14, complete sequence;AC121136.11

predicted: Candidozyma auris strain BA03 chromosome; 1 eval; CP157508.1

predictied: Phaeodactylum tricornutum CCAP 1055/1 predicted protein partial mRN;XM_002176960.1

predicted: Ixodes scapularis G-protein coupled receptor dmsr; XM_029969893.4

predicted :Melanogrammus aeglefinus genome assembly, chromosome: 10; emb OZ180142.1

predicted : Earthworm (L.terrestris) extracellular globin chain c gene, complete cds; gb J05161.1 LUMHBC

predicted:Zymobacter palmae IAM14233 DNA, complete genome;dbj|AP018933.1

predicted: Hylaeus volcanicus uncharacterized LOC128877144 (LOC128877144), transcript variant X5, mRNA;XM_054124195.1

predicted:Mus musculus BAC clone RP23-95F15 from chromosome 1, complete sequence;AC165443.5

predicted:4_Tte_b3v08;emb|OE003277.1

predicted:Earthworm (L.terrestris) extracellular globin chain c gene, complete cds;J05161.1 LUMHBC

predicted: XM_009033761.1| Helobdella robusta hypothetical protein mRNA

predicted:XM_069820523.1| PREDICTED: Periplaneta americana carbonic anhydrase beta (CAHbeta), transcript variant X3, mRNA

preddicted:Loxodonta africana zinc finger protein 252-like (LOC100666328), transcript variant X4, mRNA

predicted:gb|KX592814.1| Bos taurus isolate Dominette_000065F genomic sequence

predicted: gb|J05161.1|LUMHBC Earthworm (L.terrestris) extracellular globin chain c gene, complete cds

predicted:ref|XM_637462.1| Dictyostelium discoideum AX4 hypothetical protein (DDB_G0277655) mRNA, complete cds

predicted:Rattus norvegicus uncharacterized LOC134482949 (LOC134482949), ncRNA

Melanogrammus aeglefinus genome assembly, chromosome: 13

predicted:PREDICTED: Portunus trituberculatus putative uncharacterized protein DDB_G0271982 (LOC123514901), partial mRNA

predicted:emb|LN021320.1| Spirometra erinaceieuropaei genome assembly S_erinaceieuropaei ,scaffold SPER_scaffold0020968

predicted: gb|L12688.1|LUMBT Earthworm DNA sequence

prediction:emb|OZ078459.1| Lampetra fluviatilis genome assembly, chromosome: 56

emb|OZ180149.1| Melanogrammus aeglefinus genome assembly, chromosome: 17

predicted: ef|NC_043824.1|;Passiflora obovata chloroplast, complete genome;gb|MK694931.1|

predicted:ref|XM_005559078.4;Macaca fascicularis piggyBac transposable element derived 4 (PGBD4), mRNA

predicted:emb|OE179951.1| 2_Tcm_b3v08

predicted:emb;BX544872.8;Zebrafish DNA sequence from clone DKEY-58L12 in linkage group 3, complete sequence

predicted:XM_023356947.1;Centruroides sculpturatus uncharacterized LOC111615539 (LOC111615539), mRNA

predicted:XM_066083420.1| PREDICTED: Magallana gigas retrovirus-related Pol polyprotein from transposon 412 (LOC105343682), mRNA

For more details,see Voor meer informatie, kijk in de map identification-n, prediction.xlsx, sheet “df_fitlered”.

6.3 Visualisatie

6.4 GenViz

Voor het voorbereiden van de data kun je de volgende bestanden bekijken: genviz-features.py, map visualisatie en GenomeViz. De genen die zijn gevonden, worden weergegeven in grafieken, met speciale aandacht voor de eerste 2-3 megabases van chromosoom 1 (coördinaten 2000000-3000000).

Om te scrollen door de features, kun je de webversie gebruiken:

https://alenagrrr3.github.io/2-3mb-terrsetris/

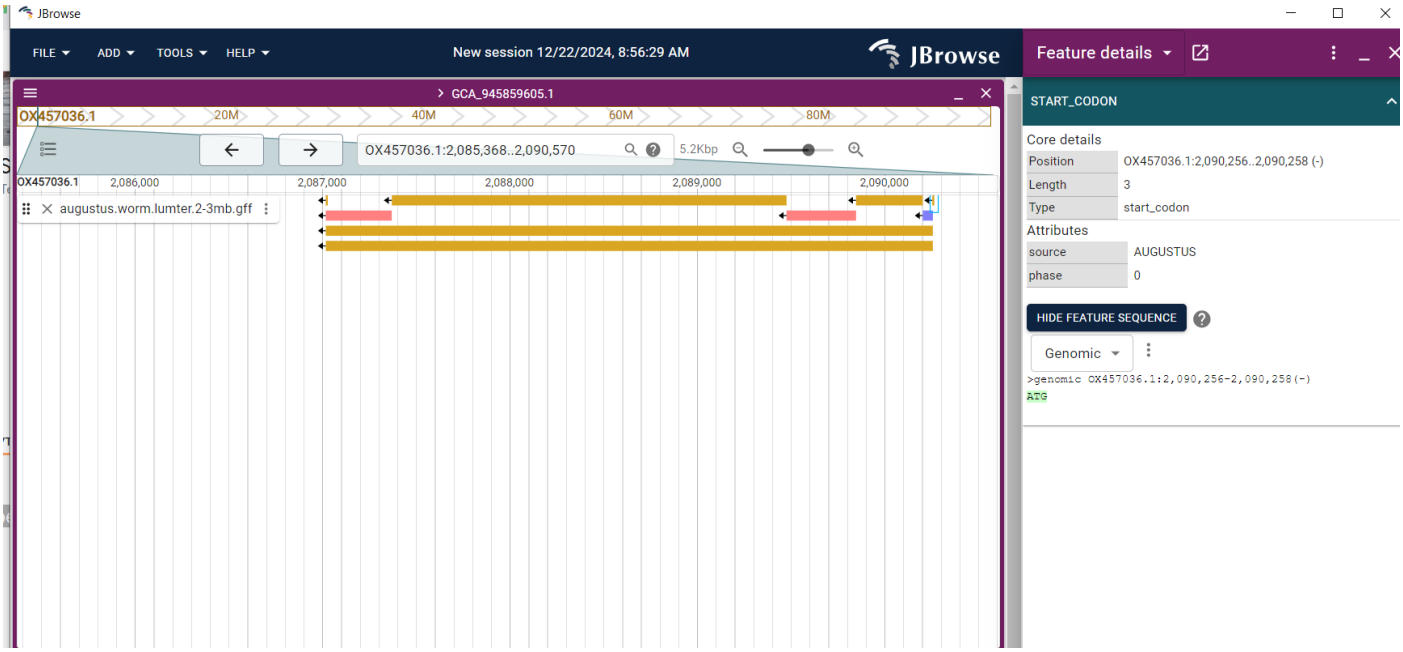
De totale representatie van het chromosoom /OX457036.1.

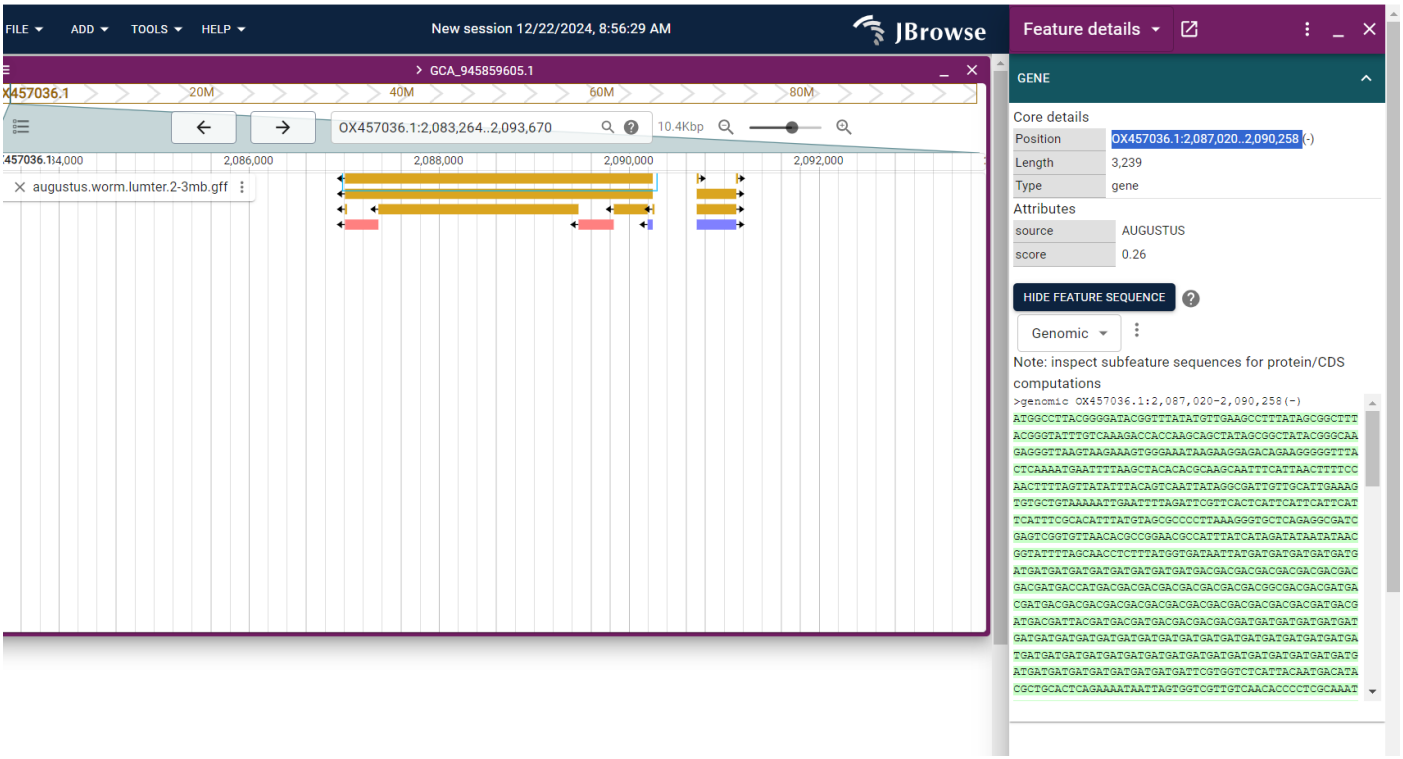
https://alenagrrr3.github.io/OX457036.1.html/

6.5 JBrowse

Het gen met de coördinaten OX457036.1:2,087,020 - 2,090,258 is geïdentificeerd als het mt2A-gen voor metallothioneïne 2A van Lumbricus rubellus, inclusief exons 1-4; AJ299434.1. is onderzocht in de in Jbrowser (“JBrowse | JBrowse” n.d.)

Gene 5, with intron, Cds, and transctipt:



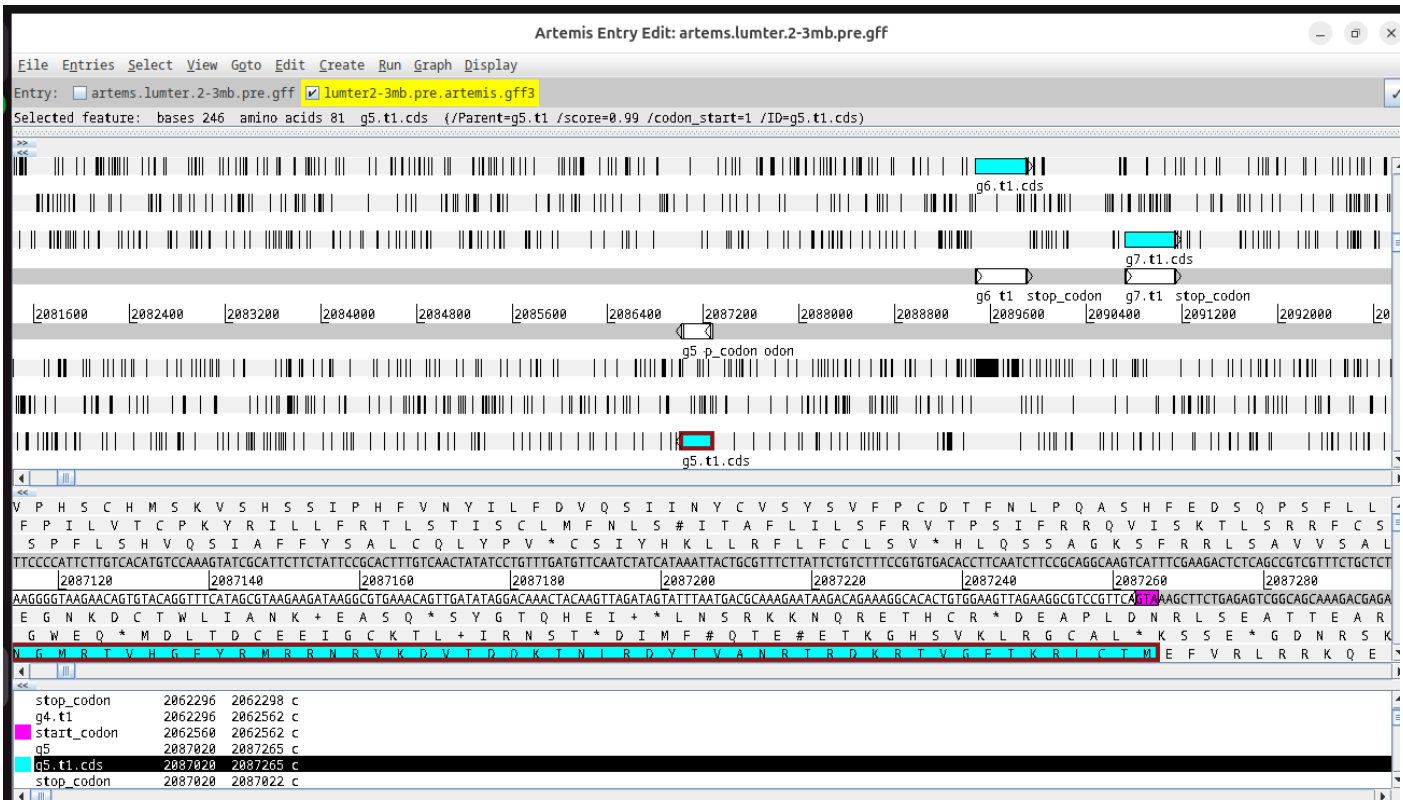


To zoom in, you can ues the link:

https://raw.githubusercontent.com/alenagrrr3/OX457036.1.html/refs/heads/main/lumterAM182481.1-gene5.svg

6.6 Artemis

```
gen "g5" (OX457036.1:2,087,020 - 2,090,258 )   in Artemis Browser  met startcodon en CDS (minus streng):
```



To zoom in, you can ues the link:

https://raw.githubusercontent.com/alenagrrr3/OX457036.1.html/refs/heads/main/artemis-g5-startcodon.webp

7 Conclusie en Discussie

RNA-Seq supported annotatie.

Het werken met zulke grote genomen is vaak moeilijk vanwege beperkte computermiddelen, zoals rekentijd en de hoeveelheid geheugen die nodig is om genoomgegevens te verwerken. De structurele annotatie van het eerste chromosoom is uitgevoerd met Augustus-pijplijnen op een Linux-systeem met 4 cores, elk met een snelheid van 799.986 MHz. De totale tijd die nodig, was 22 dagen, exclusief de tuning van Augustus en ontwikkeling van het model. Dit is ook rekening houdend met het feit dat de RNA-seq alignment drie keer is gedaan met drie verschillende RNA-seq splice-aware aligners, STAR, TopHat and Minimap.Splice-aware aligners houden in dat je niet alleen exons ontvangt, maar ook splice junctions tussen exonen in genregio's als resultaat. Vervolgens worden er intronshints van gemaakt. filterGenemark.pl filtert de genen uit die zijn opgenomen in introns-files.En alleen de genen die niet in introns zitten blijven over. De transcripten van de transcriptoomassemblages van Lumbcricus terrestris werden als basis gebruikt. 1975 genmodellen voor de chromosoom-1 die ondersteund worden door RNA-Seq alignementen worden opgeslagen in het bestand bonafide.gtf (protocol1/data_processing/bonafide.gtf).

Proteïne supported annotatie

Door Uniprot-eiwitten te gebruiken als uitgangspunt, zijn er genen verkregen voor de chromosoom-1 die voor deze eiwitten coderen: Superoxide dismutase, Calmodulin, Protein kinase C2,Myeloid Differentation primary response protein, Actin,serotonin transporter,Peroxidasin, ExtracellularHemoglobinLinker, Ubiquitin. In het bijzonder zijn de volgende eiwitten van groot belang voor de eukaryote cel.

Superoxide dismutase B9TY04 ("Superoxide Dismutase - Lumbricus Rubellus (Humus Earthworm) | UniProtKB | UniProt" n.d.), is een enzym van de klasse oxidoreductases dat de dismutatie van superoxide in zuurstof en waterstofperoxide katalyseert. Het speelt een cruciale rol in de antioxidantverdediging van vrijwel alle cellen die op de een of andere manier in contact komen met zuurstof.

Calmodulin Q9GRJ1 ("Calmodulin - Lumbricus Rubellus (Humus Earthworm) | UniProtKB | UniProt" n.d.) is een multifunctioneel intermediair calciumbindend proteïne dat tot expressie komt in alle eukaryote cellen. Het is een intracellulair doelwit van de secundaire Ca2+ boodschapper en Ca2+-binding is nodig voor calmoduline-activatie.

Protein kinase C2 Q2I699 ("Protein Kinase C2 - Eisenia Fetida (Red Wiggler Worm) | UniProtKB | UniProt" n.d.) enzym dat eiwitfosforylering uitvoert en zo deelneemt aan celsignaleringscascades.

Myeloid Differentation primary response protein A0A143Y4B3, ("EaMyD88 - Myeloid Differentation Primary Response Protein MyD88 - Eisenia Andrei | UniProtKB | UniProt" n.d.), Een adaptoreiwit dat signalen voor Toll-like receptor (TLR)-betrokkenheid door het plasmamembraan stuurt.

Actin P92182 · ACT1_LUMTE ("ACT1 - Actin-1 - Lumbricus Terrestris (Common Earthworm) | UniProtKB | UniProt" n.d.), een globulair eiwit dat de basis vormt van het cytoskelet van eukaryote cellen en het hoofdbestanddeel is van spiervezels, waar het samenwerkt met myosine om spiersamentrekkingen te produceren.

serotonin transporter ("SERT - High-affinity Serotonin Transporter Protein - Lumbricus Terrestris (Common Earthworm) | UniProtKB | UniProt" n.d.) is een intracellulair eiwit. De serotonine transporter behoort tot de familie van monoamine transporter eiwitten.

Peroxidasin V9GWR0 · V9GWR0_LUMTE ("Peroxidasin - Lumbricus Terrestris (Common Earthworm) | UniProtKB | UniProt" n.d.) Peroxidasine is een eiwit dat peroxidase en extracellulaire matrixmotieven combineert. Peroxidase katalyseert radioactieve jodering, oxidatie en dityrosinevorming in aanwezigheid van waterstofperoxide. Peroxidazine heeft functies in extracellulaire matrixconsolidatie, fagocytose en afweer.Verhoogde expressie van dit gen wordt waargenomen in de visuele organen, vet en endometrium.

Extracellular hemoglobin linker L4 subunit, ("Extracellular Hemoglobin Linker L4 Subunit - Lumbricus Terrestris (Common Earthworm) | UniProtKB | UniProt" n.d.), hemoglobine is een eiwit in rode bloedcellen dat moleculaire zuurstof transporteert.

In de regenworm Lumbricus terrestris vormen de linker subeenheden een kerncomplex met D(6)-symmetrie waaraan 12 hemoglobinedodecameren zich binden om het hele complex te vormen.(Royer et al. 2006)

Ubiquitin P84589 · UBIQ_LUMTE ("Ubiquitin - Lumbricus Terrestris (Common Earthworm) | UniProtKB | UniProt" n.d.), een klein (8,5 kDa) geconserveerd eukaryotisch eiwit dat betrokken is bij de regulatie van intracellulaire afbraak van andere eiwitten en bij de wijziging van hun functies. Het is aanwezig in bijna alle weefsels van meercellige eukaryoten, evenals in eencellige eukaryotische organismen. Het op ubiquitine gebaseerde signaleringssysteem is betrokken bij de regulering van belangrijke cellulaire processen, waaronder celcyclus, apoptose, reparatie, transcriptie, intracellulaire signaaltransductie, immuunrespons en andere. Een speciale rol van dit systeem is het handhaven van eiwithomeostase in de cel, wat wordt uitgevoerd door gerichte afbraak van defecte of kortlevende intracellulaire eiwitten gelabeld met ubiquitine in een speciale moleculaire machine - proteasoom.

Tijdens het vergelijken van de genomische structuren in de genomic browser, verkregen uit twee verschillende bronnen, de eiwitlijn en de RNA-seq lijn, merkten we dat de meeste structuren overeenkwamen. Toch waren er veel genen die nog niet goed geïdentificeerd waren, of slechts als "ongekarakteriseerd eiwit" of "hypothetisch eiwit" werden gekarakteriseerd. Aangezien er niet veel ondersteunende gegevens zijn voor deze soort, die gebruikt kunnen worden voor genvoorspelling en kwaliteitscontrole, blijven veel genomische structuren onontdekt.

8 Bijlage

Map struture:

```
fs::dir_tree("lumbricus")
```

```
## lumbricus
## +-- DESCRIPTION
## +-- NAMESPACE
## +-- bib
## +-- docs
## |   +-- PVA_Regenwormproject.html
## |   +-- PVA_regenwormproject.Rmd
## |   +-- docs.pdf
```

```
## |   \-- pva_feedback.v.1.0_lajsa_alena_merged.docx
## +-- identification-n
## |   +-- gff
## |   |   +-- OX457036.1.gff3
## |   |   \-- lumter2-3mb.gff3
## |   +-- prediciton
## |   |   +-- df.filtered.txt
## |   |   \-- genome.fa.gff
## |   +-- prediction.xlsx
## |   +-- scripts
## |   |   +-- blast.py
## |   |   +-- parse-nucleotide.py
## |   |   +-- parse-proteins.py
## |   |   +-- soort-blast-by-coords.py
## |   |   \-- sort-blast-by-pval.py
## |   \-- xml
## |       +-- blast.results.fraction2.xml
## |       +-- blast.results.fraction3.xml
## |       \-- results.fraction1.xml
## +-- identifictation-p
## |   +-- predicition
## |   |   +-- prediction.xlsx
## |   |   \-- proteinsdb.fa
## |   +-- viz
## |   |   +-- g10.png
## |   |   +-- g11.png
## |   |   +-- g12g13g14.png
## |   |   +-- g6-7.png
## |   |   +-- g8g9.png
## |   |   +-- intersectiong5.png
```

```
## |   |   \-- intersectiong5.svg
## |   \-- xml
## |       +-- CDS2000789-2003917.xml
## |       +-- CDS2007959-2008723.xml
## |       +-- CDS2087020-2087265.xml
## |       +-- CDS2108840-2109797CDS2112970-2113319.xml
## |       +-- CDS2205956-2206438.xml
## |       +-- CDS2212655-2212858.xml
## |       +-- CDS2266065-2266586.xml
## |       +-- CDS2267178-2267675.xml
## |       +-- CDS2310009-2310308CDS-2311036-2315817.xml
## |       +-- CDS2347211-2347231CDS2347648-2347833.xml
## |       +-- CDS2368170-2369933.xml
## |       +-- CDS2374823-2375022CDS2378236-2378761.xml
## |       +-- CDS2381370-2384417.xml
## |       +-- CDS2414960-2415253CDS2417218-2417280.xml
## |       +-- CDS2478928-2483715.xml
## |       +-- CDS2564359-2565530CDS2565843-2569281
## |       +-- CDS2577242-2577559.xml
## |       +-- CDS2579897-2580625.xml
## |       +-- CDS2591913-2592509.xml
## |       +-- CDS2596779-2597135.xml
## |       +-- CDS2597827-2599023CDS2600521-2600871
## |       +-- CDS2605011-2605567CDS2606116-2606311.xml
## |       +-- CDS2654466-2654759.xml
## |       +-- CDS2668036-2669283.xml
## |       +-- CDS2753195-2753644.xml
## |       +-- CDS2784545-2784997.xml
## |       +-- CDS2794556-2795188.xml
## |       +-- CDS2825999-2831233.xml
```

```
## |       +-- CDS2845298-2847028.xml
## |       +-- CDS2890565-2891302.xml
## |       +-- CDS2938553-2938625CDS2944301-2944450CDS2944934-2945007CDS2945591-2945668.xml
## |       +-- CDS2961015-2962238.xml
## |       \-- CDS2976735-2976952CDS2979221-2981863CDS2981906-2982863.xml
## +-- lumbricus.Rproj
## +-- protocol1
## |   +-- data_processing
## |   |   +-- GeneMarkES
## |   |   |   +-- genemark.average_gene_length.out
## |   |   |   +-- genemark.f.good.gtf
## |   |   |   +-- genemark.gtf
## |   |   |   \-- hmm.model
## |   |   |       \-- gmhmm.mod
## |   |   +-- TOPHAT
## |   |   |   +-- accepted_hits.bam
## |   |   |   +-- align_summary.txt
## |   |   |   +-- igv
## |   |   |   |   +-- exon-intron.png
## |   |   |   |   +-- exon-introns.svg
## |   |   |   |   +-- exon_ids.bed
## |   |   |   |   +-- exontranscripts.png
## |   |   |   |   +-- igv_snapshot.svg
## |   |   |   |   +-- igv_snapshot_bed_vs_juncions.png
## |   |   |   |   +-- junction_vs_bam_10kb.png
## |   |   |   |   +-- junction_vs_bam_11kb.png
## |   |   |   |   +-- junction_vs_bam_11kb2.png
## |   |   |   |   +-- junction_vs_bam_2.8.bp.png
## |   |   |   |   \-- transctipts_ids.bed
## |   |   |   +-- introns.gff
```

```
## |   |   |   +-- introns_by_gmh_with_gtf.gff
## |   |   |   +-- junctions.bed
## |   |   |   \-- transcripts.gtf
## |   |   \-- bonafide
## |   |       +-- bonafide.gb
## |   |       +-- bonafide.gtf
## |   |       +-- bonafide.unique.gb
## |   |       +-- etrain.out
## |   |       \-- test.out
## |   +-- data_raw
## |   |   +-- masked
## |   |   |   \-- chromosome1.fasta.masked
## |   |   \-- soft-masked
## |   |       \-- soft.masked.chromosome1.OX457036.1.fasta
## |   +-- model
## |   |   +-- lumter
## |   |   |   +-- lumter_exon_probs.pbl
## |   |   |   +-- lumter_igenic_probs.pbl
## |   |   |   +-- lumter_intron_probs.pbl
## |   |   |   +-- lumter_metapars.cfg
## |   |   |   +-- lumter_metapars.cgp.cfg
## |   |   |   +-- lumter_metapars.utr.cfg
## |   |   |   +-- lumter_parameters.cfg
## |   |   |   \-- lumter_weightmatrix.txt
## |   |   \-- wormET0
## |   |       +-- wormET0_exon_probs.pbl
## |   |       +-- wormET0_igenic_probs.pbl
## |   |       +-- wormET0_intron_probs.pbl
## |   |       +-- wormET0_metapars.cfg
## |   |       +-- wormET0_metapars.cgp.cfg
```

```
## |   |       +-- wormET0_metapars.utr.cfg
## |   |       +-- wormET0_parameters.cfg
## |   |       \-- wormET0_weightmatrix.txt
## |   +-- refs
## |   |   +-- README.GeneMark-ET
## |   |   \-- refs
## |   +-- scrips
## |   |   +-- bed_to_gff.pl
## |   |   +-- filterGenemark.pl
## |   |   +-- step1.sh
## |   |   +-- step2.sh
## |   |   +-- step3.sh
## |   |   +-- step4.sh
## |   |   \-- step5.sh
## |   \-- test
## |       +-- test.gb
## |       +-- test.out
## |       \-- train.gb
## +-- protocol2
## |   +-- data_processing
## |   |   +-- Bonafid
## |   |   |   +-- bonafide.gb
## |   |   |   +-- bonafide.gtf
## |   |   |   \-- etrain.out
## |   |   +-- ProtHints
## |   |   |   +-- augustus.hints.prots.orthodb.arthropoda.2-3mb.gff
## |   |   |   +-- extrinsic.cfg
## |   |   |   +-- gth.concat.aln
## |   |   |   +-- prothint.gff
## |   |   |   +-- prothint_augustus.gff
```

```
## |   |   |   +-- run.cfg
## |   |   |   \-- seed_proteins.faa
## |   |   +-- Redundancy
## |   |   |   +-- bonafide.f.gb
## |   |   |   +-- bonafide.f.gtf
## |   |   |   +-- bonafide.f.nonred.gb
## |   |   |   +-- loci.lst
## |   |   |   +-- nonred.loci.lst
## |   |   |   +-- nonred.lst
## |   |   |   +-- prot.aa
## |   |   |   +-- prot.nr.aa
## |   |   |   \-- traingenes.lst
## |   |   \-- bad-list
## |   |       +-- bad.list
## |   |       +-- bad.pre.list
## |   |       +-- bonafide.f.nonred.gb
## |   |       \-- inseq
## |   +-- data_raw
## |   |   \-- transcriptome.refs
## |   +-- filter
## |   |   +-- before.png
## |   |   +-- bonafide.filtered.nonred.gb
## |   |   +-- export.hist.on.wormEP
## |   |   +-- filterGenes.pl
## |   |   +-- filtered.gb
## |   |   +-- prot.out.png
## |   |   \-- test.out
## |   +-- model
## |   |   +-- protsLumter
## |   |   |   +-- protsLumter_exon_probs.pbl
```

```
## |   |   |   +-- protsLumter_igenic_probs.pbl
## |   |   |   +-- protsLumter_intron_probs.pbl
## |   |   |   +-- protsLumter_metapars.cfg
## |   |   |   +-- protsLumter_metapars.cgp.cfg
## |   |   |   +-- protsLumter_metapars.utr.cfg
## |   |   |   +-- protsLumter_parameters.cfg
## |   |   |   \-- protsLumter_weightmatrix.txt
## |   |   \-- wormNonredEP
## |   |       +-- wormNonredEP_exon_probs.pbl
## |   |       +-- wormNonredEP_igenic_probs.pbl
## |   |       +-- wormNonredEP_intron_probs.pbl
## |   |       +-- wormNonredEP_metapars.cfg
## |   |       +-- wormNonredEP_metapars.cgp.cfg
## |   |       +-- wormNonredEP_metapars.utr.cfg
## |   |       +-- wormNonredEP_parameters.cfg
## |   |       \-- wormNonredEP_weightmatrix.txt
## |   +-- refs
## |   |   \-- refs
## |   +-- resources
## |   |   \-- ncbi-blast-2.16.0+
## |   |       +-- ChangeLog
## |   |       \-- bin
## |   |           +-- blast_vdb_cmd
## |   |           +-- blastn_vdb
## |   |           +-- blastp
## |   |           +-- makeprofiledb
## |   |           \-- rpsblast
## |   +-- scripts
## |   |   +-- bonafide.nonred.f.py
## |   |   +-- create_train_list.py
```

```
## |   |   +-- createbonafidef.py
## |   |   +-- locilst.py
## |   |   +-- nonred.loci.py
## |   |   +-- step1.sh
## |   |   +-- step2.sh
## |   |   \-- step3.sh
## |   \-- test
## |       +-- test.gb
## |       +-- test.out
## |       \-- train.gb
## +-- protocol2.2
## |   +-- data_processing
## |   |   +-- GenomeThreader.output.gth
## |   |   +-- extrinsic.cfg
## |   |   +-- merged_6393_and_6397.fa
## |   |   \-- prothint_augustus.gff
## |   +-- data_raw
## |   |   +-- soft.masked.chromosome1.OX457036.1.fasta
## |   |   +-- uniprotkb_taxonomy_id_6393_2024_12_29.fasta.gz
## |   |   \-- uniprotkb_taxonomy_id_6397_2024_12_29.fasta.gz
## |   +-- gff
## |   |   \-- augustus.prothints.gff3
## |   \-- scripts
## |       +-- blast-p.py
## |       +-- parse-proteins.py
## |       \-- stap1.sh
## \-- visualization
##     +-- GenomeViz
##     |   +-- custom_bopython-feature.png
##     |   +-- genviz-features.py
```

```
##     |   +-- index.html
##     |   \-- terr.png
##     +-- IGV
##     |   +-- OX457036.1.fasta
##     |   +-- bam
##     |   |   +-- accepted_hits.bam
##     |   |   \-- accepted_hits.bam.bai
##     |   \-- bed
##     |       +-- exon_ids.bed
##     |       \-- tranasctips_ids.bed
##     +-- artemis
##     |   +-- artemis-g5.png
##     |   \-- lumter.artemis.track.gff3
##     \-- jbrowser
##         +-- 2087020.png
##         +-- cds.png
##         +-- intron.png
##         +-- start_codon.png
##         \-- transcript1.png
```

References

"ACT1 - Actin-1 - Lumbricus Terrestris (Common Earthworm) | UniProtKB | UniProt." n.d. Accessed January 2, 2025. https://www.uniprot.org/uniprotkb/P92182/entry.

"Augustus." n.d. Bioinformatics Notebook. Accessed November 25, 2024. https://rnnh.github.io/bioinfo-notebook/docs/augustus.html.

"Augustus/Docs/RUNNING-AUGUSTUS.md at Master · Gaius-Augustus/Augustus." n.d. GitHub. Accessed November 25, 2024. https://github.com/Gaius-Augustus/Augustus/blob/master/docs/RUNNING-AUGUSTUS.md.

Baum, Dr Julia. n.d. "Ever Thought about Earthworms?" African Wildlife Economy Institute. Accessed November 25, 2024. https://www0.sun.ac.za/awei/articles/ever-thought-about-earthworms.

"Bioinformatics and Other Bits - Creating a Local RefSeq Protein Blast Database." n.d. Accessed November 28, 2024. https://dmnfarrell.github.io/bioinformatics/local-refseq-db.

"Bioinformatics Web Server - University of Greifswald." n.d. Accessed December 18, 2024. https://bioinf.uni-greifswald.de/bioinf/partitioned_odb11/.

Blaxter, Mark L., David Spurgeon, and Peter Kille. 2023a. "The Genome Sequence of the Common Earthworm, Lumbricus Terrestris (Linnaeus, 1758)." *Wellcome Open Research* 8 (October): 500. https://doi.org/10.12688/wellcomeopenres.20178.1.

———. 2023b. "The Genome Sequence of the Common Earthworm, Lumbricus Terrestris (Linnaeus, 1758)." *Wellcome Open Research* 8 (October): 500. https://doi.org/10.12688/wellcomeopenres.20178.1.

Brůna, Tomáš, Katharina J Hoff, Alexandre Lomsadze, Mario Stanke, and Mark Borodovsky. 2021. "BRAKER2: Automatic Eukaryotic Genome Annotation with GeneMark-EP+ and AUGUSTUS Supported by a Protein Database." *NAR Genomics and Bioinformatics* 3 (1): lqaa108. https://doi.org/10.1093/nargab/lqaa108.

Buchfink, Benjamin, Chao Xie, and Daniel H Huson. 2015. "Fast and Sensitive Protein Alignment Using DIAMOND." *Nature Methods* 12 (1): 59–60. https://doi.org/10.1038/nmeth.3176.

"Calmodulin - Lumbricus Rubellus (Humus Earthworm) | UniProtKB | UniProt." n.d. Accessed January 2, 2025. https://www.uniprot.org/uniprotkb/Q9GRJ1/entry.

colauttilab.github.io/. n.d. "De Novo Assembly Tutorial." Accessed November 30, 2024. https://colauttilab.github.io/NGS/deNovoTutorial.html.

"De Novo Assembly Tutorial." n.d. Accessed December 29, 2024. https://colauttilab.github.io/NGS/deNovoTutorial.html.

"Download GenomeThreader." n.d. Accessed January 1, 2025. https://genomethreader.org/download.html.

"EaMyD88 - Myeloid Differentation Primary Response Protein MyD88 - Eisenia Andrei | UniProtKB | UniProt." n.d. Accessed January 2, 2025. https://www.uniprot.org/uniprotkb/A0A143Y4B3/entry.

ebi.ac.uk. n.d. "ENA Browser." Accessed November 25, 2024. https://www.ebi.ac.uk/ena/browser/view/PRJEB59400.

"ENA Browser." n.d. Accessed December 18, 2024. https://www.ebi.ac.uk/ena/browser/view/ERR10851549.

Erxleben, Anika, and Björn Grüning. 12:19:56 +0000. "Genome Annotation." Text. Galaxy Training Network; Galaxy Training Network. 12:19:56 +0000. https://translated.turbopages.org/proxy_u/en-ru.ru.dd5ab9ec-67446c58-5ab2c0cb-74722d776562/https/training.galaxyproject.org/training-material/topics/genome-annotation/tutorials/genome-annotation/tutorial.html.

"Extracellular Hemoglobin Linker L4 Subunit - Lumbricus Terrestris (Common Earthworm) | UniProtKB | UniProt." n.d. Accessed January 2, 2025. https://www.uniprot.org/uniprotkb/Q2I741/entry.

"Gaius-Augustus/BRAKER." (2018) 2024. Gaius-Augustus. https://github.com/Gaius-Augustus/BRAKER.

"Gene Cluster Visualizations in R." n.d. Accessed November 27, 2024. https://nvelden.github.io/geneviewer/.

"Genome Annotation / Tutorial List." 13:32:22 +0000. Text. Galaxy Training Network; Galaxy Training Network. 13:32:22 +0000. https://training.galaxyproject.org/training-material/topics/genome-annotation/.

"Genometools/Genomethreader." (2019a) 2024. GenomeTools. https://github.com/genometools/genomethreader.

———. (2019b) 2024. GenomeTools. https://github.com/genometools/genomethreader.

"Home · TransDecoder/TransDecoder Wiki." n.d. Accessed November 28, 2024. https://github.com/TransDecoder/TransDecoder/wiki.

"Hox20 - Eisenia Andrei | UniProtKB | UniProt." n.d. Accessed January 2, 2025. https://www.uniprot.org/uniprotkb/Q8MWS8/entry.

"Index of /Genomes." n.d. Accessed November 28, 2024. https://ftp.ncbi.nlm.nih.gov/genomes/.

"Index of /Genomes/MapView." n.d. Accessed November 28, 2024. https://ftp.ncbi.nlm.nih.gov/genomes/MapView/.

"Index of /Releases/Dfam_3.8/Families/FamDB/." n.d. Accessed December 18, 2024. https://www.dfam.org/releases/Dfam_3.8/families/FamDB/.

"JBrowse | JBrowse." n.d. Accessed November 26, 2024. https://jbrowse.org/jb2/.

Leung, Maxwell C. K., Phillip L. Williams, Alexandre Benedetto, Catherine Au, Kirsten J. Helmcke, Michael Aschner, and Joel N. Meyer. 2008. "Caenorhabditis Elegans: An Emerging Model in Biomedical and Environmental Toxicology." *Toxicological Sciences* 106 (1): 5–28. https://doi.org/10.1093/toxsci/kfn121.

"LumbriBASE." n.d. Accessed November 30, 2024. http://xyala2.bio.ed.ac.uk/Lumbribase/lumbribase_php/lumbribase.shtml.

ncbi.nlm.nih.gov. n.d. "The NCBI Eukaryotic Genome Annotation Pipeline." Accessed November 25, 2024. https://www.ncbi.nlm.nih.gov/refseq/annotation_euk/process/.

"Pcs-1a - Glutathione Gamma-Glutamylcysteinyltransferase - Lumbricus Rubellus (Humus Earthworm) | UniProtKB | UniProt." n.d. Accessed January 2, 2025. https://www.uniprot.org/uniprotkb/V9VGQ0/entry.

"Peroxidasin - Lumbricus Terrestris (Common Earthworm) | UniProtKB | UniProt." n.d. Accessed January 2, 2025. https://www.uniprot.org/uniprotkb/V9GWR0/entry.

Pilato, Giovanni. n.d. "The significance of musculature in the origin of the Annelida." Accessed November 30, 2024. http://ouci.dntb.gov.ua/en/works/ldperODl/.

"Protein Kinase C2 - Eisenia Fetida (Red Wiggler Worm) | UniProtKB | UniProt." n.d. Accessed January 2, 2025. https://www.uniprot.org/uniprotkb/Q2I699/entry.

Royer, William E., Hitesh Sharma, Kristen Strand, James E. Knapp, and Balaji Bhyravbhatla. 2006. "Lumbricus Ery-

throcruorin at 3.5 Å Resolution: Architecture of a Megadalton Respiratory Complex." *Structure* 14 (7): 1167–77. https://doi.org/10.1016/j.str.2006.05.011.

"Sanger-Pathogens/Artemis." (2009) 2024. Pathogen Informatics, Wellcome Sanger Institute. https://github.com/sanger-pathogens/Artemis.

"SERT - High-affinity Serotonin Transporter Protein - Lumbricus Terrestris (Common Earthworm) | UniProtKB | UniProt." n.d. Accessed January 2, 2025. https://www.uniprot.org/uniprotkb/Q0G8J7/entry.

Short, Stephen, Amaia Green Etxabe, Alex Robinson, David Spurgeon, and Peter Kille. 2023. "The Genome Sequence of the Red Compost Earthworm, Lumbricus Rubellus (Hoffmeister, 1843)." *Wellcome Open Research* 8 (August): 354. https://doi.org/10.12688/wellcomeopenres.19834.1.

Stanke, Mario. 2005. "Augustus Online." Service. Institute for Mathematics and Computer Science, University of Greifswald. February 4, 2005. https://bioinf.uni-greifswald.de/augustus/submission.php.

"Superoxide Dismutase - Lumbricus Rubellus (Humus Earthworm) | UniProtKB | UniProt." n.d. Accessed January 2, 2025. https://www.uniprot.org/uniprotkb/B9TY04/entry.

"(Taxonomy_id:6393) in UniProtKB Search (633) | UniProt." n.d. Accessed January 1, 2025. https://www.uniprot.org/uniprotkb?query=%28taxonomy_id%3A6393%29.

"(Taxonomy_id:6397) in UniProtKB Search (698) | UniProt." n.d. Accessed January 1, 2025. https://www.uniprot.org/uniprotkb?query=%28taxonomy_id%3A6397%29.

"The Genome Sequence of the Common ... | Wellcome Open Research." n.d. Accessed December 19, 2024. https://wellcomeopenresearch.org/articles/8-500.

"The NCBI Eukaryotic Genome Annotation Pipeline." n.d. Accessed December 29, 2024. https://www.ncbi.nlm.nih.gov/refseq/annotation_euk/process/.

"Ubiquitin - Lumbricus Terrestris (Common Earthworm) | UniProtKB | UniProt." n.d. Accessed January 2, 2025. https://www.uniprot.org/uniprotkb/P84589/entry.

University of Greifswald. n.d. "Bioinformatics Web Server - University of Greifswald." Accessed December 28, 2024. https://bioinf.uni-greifswald.de/bioinf/partitioned_odb11/.

Wang, Ying, Robyn Branicky, Alycia Noë, and Siegfried Hekimi. 2018. "Superoxide Dismutases: Dual Roles in Controlling ROS Damage and Regulating ROS Signaling." *Journal of Cell Biology* 217 (6): 1915–28. https://doi.org/10.1083/jcb.201708007.