Plan van Aanpak

Project Lumbricus terrestris

A shiny fish with a small white bug

Description automatically generated with medium confidence

Project : 'Lumbricus terrestris'

Naam projectleden: Laïsja Thomasia , Alena Savachenko

Opdrachtgever : Chris van Oevelen

Versie : 1.2

**Achtergrond**

Regenwormen zijn bodembewonende wormen die behoren tot de familie Lumbricidae. Regenwormen spelen een heel belangrijke rol in het ecosysteem, vooral in de bodem. Regenwormen zijn bekend om hun vermogen om organisch materiaal, zoals dode bladeren en plantenresten, af te breken en te verwerken tot humus, een voedzame stof die belangrijk is voor planten. Lumbricus terrestris, ook wel de gewone regenworm of zandworm genoemd, is een van de bekendste en grootste soorten regenwormen. Volwassen Lumbricus kunnen tussen de de 20 en 30 cm lang worden, en in uitzonderlijke gevallen zelfs tot 40 cm. Lumbricus terrestris hebben een een roze-bruine kleur en is voorzien van segmenten, met een duidelijk zichtbaar clitellum. Ze hebben ook een lichaam die cilindrisch is, met een glad oppervlak en segmenten die door een dunne slijmlaag verbonden zijn. in dit project wordt er een genoom annotatie uitgevoerd op Lumbricus terrestris. Genoomannotatie is het proces waarbij wetenschappers de verschillende functies van de genen en andere elementen binnen het genoom van een organisme identificeren en labelen. Het doel van genoomannotatie is om de informatie in het DNA te begrijpen, zodat onderzoekers kunnen bepalen welke stukken van het genoom coderen voor eiwitten, welke functies deze eiwitten hebben, en hoe ze bijdragen aan de werking van de cel of het organisme. Dit wordt uitgevoerd met de behulp van C.elegans als basemodel voor L. terrestris. Dit is omdat C. elegans goed geannoteerd is en als modelorganisme gebruikt kan worden voor de genoom annotatieprocess van Lumbricus terrestris. Doordat het genoom van L terrestris niet volledig geannoteerd is is het doel van dit onderzoek om een genoomannotatie uit te voeren op L. terrestris door C.elegans te gebruiken als basemodel. Dit wordt samen uitgevoerd met een positieve control om ons resultaten te verifieren. Er verwacht dat er veel genen homologie zal vertonen doordat L.terrestris en C.elegans een gemeenschappelijke evolutionaire oorsprong hebben. Een lage E waarde zou dit kunnen aantonen. Bij de IGV wordt er verwacht dat de exacte locaties van de homologe gen, exons en introns worden aangetoond.

**Positieve controle = Lumbricus rubellus**

L. rubellus en L. terrestris behoren tot de zelfde geslacht Lumbricus. Dit betekent dat hun genetische informatie vergelijkbare genen en genfamilies bevat, ondanks het verschil in chromosoomaantal. Dit maakt L. rubellus waardevol als referentie voor het identificeren van conservatieve genen in L. terrestris, omdat veel genen waarschijnlijk in beide soorten te vinden zullen zijn. Het genoom van rubellus is ook redelijk goed geanalyseerd en geannoteerd. Maar zit niet in AUGUSTUS, we kunnen dan onze eigen trainingset maken

**De Genoom annotatieproces van Lumbricus terrestris bestaat uit de volgende stappen:**

* Genoom data verkrijgen (FASTA bestanden van Lumbricus terrestris, C.elegans en positieve controle)
* Benodigde libraries installeren = Biostrings, ggplot2, tidyverse, BLAST, GenomicRanges, Rtracklayer (AUGUSTUS moet zelf op de systeem geinstalleerd zijn “UBUNTU”)
* FASTA-bestanden controleren (Biostrings)
* BLAST uitvoeren voor Alignment (BLASTR)
* Resultaten van BLAST ( kijken naar E waarde, hits en identiteit van de overeenkomstige genen)
* AUGUSTUS voor genoomannotatie (kijken naar sensitivity en specificity)
* AUGUSTUS visualiseren ( IGV of kan zelfs op R met de package RTracklayer)
* Genen verfijnen en valideren (het gebruik van Gene Ontology om de functies van de geannoteerde genen bepalen)

**Relevantie van het onderzoeksonderwerp**

Dit onderzoeksonderwerp is van wezenlijk belang, omdat er in de experimentele wetenschap en de geneeskunde op dit moment wordt gestreefd naar alternatieven voor het gebruik van dieren bij diverse laboratoriumexperimenten. Vanwege de belangrijke rol die diverse wormsoorten spelen als modellen in medisch en biologisch onderzoek, is het essentieel om hun genetisch materiaal te bestuderen. Dit project richt zich specifiek op de analyse van het genetisch materiaal van de soort Lumbricidae.(Leung et al. 2008). De soort Lumbricus Terresteris hoort tot de fylogenetische familie Annelida, Clitellata, Oligochaeta, Crassiclitellata, Lumbricina, Lumbricidae(Blaxter, Spurgeon, and Kille 2023). De soort ringwormen (Annelida) zijn de oudste evolutionaire groep. De musculatuur lijkt hier sterk op de dwarsgestreptte musculatuur van dieren. De soorten ringwormen (Annelida) zijn de oudste evolutionaire groep. In dit geval lijkt de musculatuur sterk op dwarsgestreepte spieren bij dieren(Giovanni 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. 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(colauttilab.github.io/ n.d.)De processen van genomassemblage en anatomie 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 2023(Blaxter et al. 2023). 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.

Workflows:

De analyse van het model en het opsporen van homologen worden uitgevoerd op genomisch niveau, evenals de eiwitanalyse.Algemeen gesproken blijven de algoritmen consistent, terwijl er verschillende benaderingen en hulpmiddelen worden ingezet voor de analyse van nucleotide- en eiwitsequenties.

Analyse op het niveau van de eiwitten.

Fasta source files:

<https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/949/752/735/GCA_949752735.1_wcLumTerr1.1/>

https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/945/859/605/GCA\_945859605.1\_wcLumRube1.1

Fasta per chromosome

<https://pmc.ncbi.nlm.nih.gov/articles/PMC10799228/#sec9> # Data availability

1. Het scannen van het fasta-bestand per chromosome in een of meerdere scanprogramma's, met behulp van basisprotocollen (augustus --species=species\_name input\_file.fna > output\_file.gff.De resultaten van verschillende scanners worden vergeleken op het gebied van genoomgrootte en het detecteren van noise. Dit gebeurt door de selectiviteit en specificiteit te wijzigen.Beschkbare augustus parameters voor de specfictiet: instellen de eigenschappen voor ExonModel, IntronModel, UtrModel, number of sample iteration, softmasking, noInFrameStop (not report transcripts with in-frame stop codons) , maxOverlap(maximum overlap of coding regions). De scanprogramma's te evalueren met behulp van een set positieve controles, en set van negatieve controles (set random DNA sequentes),
2. Het vergelijken van de verkregen sequenties met de beschikbare databases, zoals de Blast Rest Api en user defined scripts. In Augustus: Augustus *Protein* Profile eXtension (*PPX*).
3. Datavisualisatie met behulp van bio-informaticatools (Bio.Graphics GenomeDiagram)

References

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

colauttilab.github.io/. n.d. ‘De Novo Assembly Tutorial’. Retrieved 30 November 2024 (https://colauttilab.github.io/NGS/deNovoTutorial.html).

Giovanni Pilato. n.d. ‘The significance of musculature in the origin of the Annelida’. Retrieved 30 November 2024 (http://ouci.dntb.gov.ua/en/works/ldperODl/).

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. doi: 10.1093/toxsci/kfn121.

ncbi.nlm.nih.gov. n.d. ‘The NCBI Eukaryotic Genome Annotation Pipeline’. Retrieved 25 November 2024 (https://www.ncbi.nlm.nih.gov/refseq/annotation\_euk/process/).