Analysis of Genes of Interest Using Online Resources

Case Study 1: ACTIN2 (ACT2) in Arabidopsis thaliana (AT3G18780).
This gene is well-characterized and encodes a protein from the actin family, involved in the cellular cytoskeleton. *Arabidopsis thaliana* is a widely used plant model in plant biology and genetics, with a relatively small genome for a eukaryote (about 135 Mb), making it easier to manipulate for bioinformatics analyses.

Case Study 2: BRCA1 in humans, a breast cancer predisposition gene.
Available resources are extensive for this gene as it is widely studied by the community (homework assignment).

---

1. Sequence Search and Retrieval on NCBI

- Access the Nucleotide database on NCBI (https://www.ncbi.nlm.nih.gov/nucleotide/).
  - Using keywords (Gene Name and Organism) in the advanced search, find the coding sequence (CDS) of the studied gene. What do you notice?
  - Note the accessions and sizes of the different mRNA of this gene in bp.
  - What is the taxonomic identifier of the species? To which order does it belong? What is the taxonomic class?
- Access the NCBI Gene database and search for the gene.
  - On which chromosome is the gene located? What is the physical position on the chromosome? What is the accession number of the chromosome?
  - How many exons does the gene contain?
  - How many scientific articles discuss this gene (cross-reference)?
  - Check accessions of mRNA associated with this gene
  - Retrieve the FASTA sequence of the gene.
- Retrieve the FASTA sequence of the protein using its accession number.
- Retrieve the FASTA sequence of the corresponding chromosome and its GFF annotation using its accession number.

---

2. Search for Available Genomes

- How many complete genomes are available for this species? How many genomes have a chromosome assembly status? How many are annotated?
- Retrieve the reference genome in FASTA format, the GFF annotation file, and the predicted protein file.

---

3. BLAST Analysis and Identification of Orthologous Sequences

- Go to the NCBI BLAST page (https://blast.ncbi.nlm.nih.gov/Blast.cgi).
    - Use the BLASTN tool to search for orthologous sequences of the gene in other phylogenetically close species.
    - Analyze the results to identify species with similar sequences and analyze sequence differences in 8 closely related species of your choice.
    - Retrieve similar sequences from the 8 closely related species.
    - Display the distance tree from the BLAST results based on this gene.

4. Ensembl and web browser

- Open the web browser and go to Ensembl Genomes
- Search for "Arabidopsis thaliana" in the search bar.
- Identification of the ACT2 Gene:
    - In the "Gene" section, search for "ACT2."
    - Select the ACT2 gene from the results. Q1: What is the gene ID? Q2: What is/are the location(s) of the gene?
- Analysis of Paralogs:
    - On the ACT2 gene page (actin2), go to the "Paralogs" section.
    - Identify and list the paralogues of the ACT2 gene.
    - Download the DNA sequences of the paralogues in FASTA format.
    - Compare the paralog sequences with that of the ACT2 gene using a sequence alignment tool (e.g., Clustal Omega). Q3: What is a paralogous gene?
- Analysis of Orthologs:
    - On the ACT2 gene page, go to the "Orthologs" section.
    - Identify and list the orthologs of the ACT2 gene in different species.
    - Download the DNA sequences of the orthologs in FASTA format.
    - Compare the ortholog sequences with that of the ACT2 gene using a sequence alignment tool (e.g., Clustal Omega). Q4: What is an orthologous gene? Q5: What is a homologous gene?
- Analysis of Genetic Variations:
    - Explore the genetic variants associated with the ACT2 gene and its paralogues and orthologs using the tools available on Ensembl.
    - Note any significant SNPs (Single Nucleotide Polymorphisms) and other variations.
- Data Visualization:
    - Use the visualization tools in Ensembl to examine the genomic data of the ACT2 gene, its paralogues, and orthologs.
    - Create screenshots of the visualizations for inclusion in the report.

---

5. Primer Design with Primer Blast

- Enter the accession number of the gene and run the primer design analysis, then observe the results.

Change the maximum amplicon size to 2000 or other parameters and observe the changes in the results.

---

6. Navigation in a Genome Browser

- Go to a Genome Browser such as the TAIR Genome Browser (JBrowse).
- Search for the gene using genomic coordinates.
- Explore the gene features in the genomic context and its neighboring genes. How many exons/introns/UTRs are there? Are there identified SNPs/indels/SSRs in the exons?

By selecting visualization tracks (EST, RNA-Seq), can you see transcriptomic evidences confirming the gene structure?

---

7. Search for Associated RNA-Seq Data on ENA

- Access the ENA (European Nucleotide Archive) database (https://www.ebi.ac.uk/ena).
- Search for RNA-Seq experiments associated with *Arabidopsis thaliana* using the keywords "Arabidopsis thaliana RNA-Seq"
- Explore available datasets, filtering by sample type or experimental condition (e.g., stress response, development, mutants).
- Download data from an RNA-Seq experiment of interest for further analysis.

---

8. Study of the Gene-Derived Protein with Uniprot, Protein Motif Search with Pfam/Interpro, and Functional Characterization with Gene Ontology (GO)

- Search for the protein in the Uniprot database using keywords (species, gene name). What is its accession number?
- Observe the Gene Ontology terms associated with this protein.
- In which compartment will the protein be located?
- Observe information related to protein motifs and domains (Pfam/Interpro). What protein motif does it contain? What is the Interpro motif accession number? Check if the protein sequence contains other domains by entering the sequence directly into PFAM.

Bonus (if there’s time or for those who want to continue)

9. Exploration of the Protein's 3D Structure with PDB

- Access the PDB (Protein Data Bank) database (https://www.rcsb.org/).
- Search for a 3D structure of the protein of interest. Although the specific structure of the protein may not be available, search for homologous structures (e.g., actin from yeast or mammals).
- Visualize the 3D structure.

---

10. Search for Signaling Pathways with KEGG

To which metabolic pathway does the gene contribute?

---

11. SNP Search and Genetic Diversity Analysis

- Download VCF (Variant Call Format) files containing SNPs for *Arabidopsis thaliana*. These files are often available in databases like 1001 Genomes (http://1001genomes.org/), which lists genetic variations within different cultivars and accessions of *Arabidopsis*.
- Use the VCF subset tool to generate the VCF corresponding only to the gene region.
- Use tools like SNiPlay (https://sniplay.southgreen.fr/cgi-bin/analysis_v3.cgi) to analyze SNPs present in the gene region.

---

12. Use of a Local Genome Browser (IGV) to Import Data

- Download and install IGV.
- Import the reference genome, its structural annotation (GFF), and integrate the VCF file.
- What SNPs are identified in the exons gene?

Recommended Materials and Software:

Web browser to access databases (NCBI, ENA, TAIR).
Bioinformatics software such as IGV.