Project Topics

Mutating for Fun

Develop a pipeline to identify mutations a given collection of isolates, strains generated from a mutagenesis experiment. Determine what genes are mutated. Summarize the types of mutations, size of them, etc. For example see the Lenski E. coli LTEE dataset with a guide to downloading data as part of this Data Carpentry tutorial.

For a plant project: e.g. some data from "Next-generation forward genetic screens: using simulated data to improve the design of mapping-by-sequencing experiments in Arabidopsis"

Or try a smaller genome example in bacteria. Do this for point mutations - can you identify the specific candidate changes based on analysis compared to a reference genome? Can you identify mutational biases - eg if the mutagen was UV vs EMS can you identify the mutational bias or pattern?

Transporterland

Cells use transporters to move metabolites, compounds, amino acids, and many molecules through cell membranes. Their affinities can vary and provide a means to maintain osmoregulation or available energy levels in cells. Develop software package to characterize at least one type of transporter family for a given proteome or set of query proteins provided to the tool.

This will be accomplished most likely by using Pfam HMM domain(s) but potentially also through structural analyses and identification of Transmembrane domains. As these cross the membrane they will have transmembrane domains. Provide a report of the number of transmembrane domains present, summarize the transporter classes found, their TM count, and other summary statistics. Different directions to explore might include doing a comparison among set of species to identify those transporter families or orthologs that are shared between species.

Quick links: * https://en.wikipedia.org/wiki/Membrane_transport_protein * https://en.wikipedia.org/wiki/Ion_transporter

Workflows

Develop a simple nextflow type workflow pipeline to deploy a set of analyses. This is more suited towards those with programming experience or who want to learn how to develop with this pipeline software.

Genome (RE)assembly

Reassembly of genomes with newer assemblers can sometimes yield improvements. Getting out possible plasmid, organelle (Chloroplast) or Mitochondrial genomes from organisms where the nuclear genome was only generated and deposited can sometimes give additional insight and datasets for comparison.

An area that has been ignored sometimes in creating genome assemblies is getting a full and corrected Mitochondrial genome. Develop a working project to take a list of [SRA](https://sra.ncbi.nih.gov) accessions, download the reads, do targeted assembly of MT with tools like NOVOPlasty (and use tools like AAFTF to simplify this). Consider also running tools like unicycler which can do improved assembly of MT and circularize.

Secondary Metabolistic Explosions

Develop pipeline for screening for secondary metabolite clusters in genomes of interest. Identify the active enzymes in the clusters, and process these to do a comparison on the similarities between the sequences.

(need to provide more updates and examples here)

From SNPs to Genes to Functions

Identify Genes with SNPs or SNVs based on previous work. Organize by which fall into Genes and those which have impacts on the gene either interrupting the coding sequence, changing amino acids, or splicing. Gather the function of these genes and test for enrichment of domains, GO function, and provide a report. Generalize this for data from GFF annotated genomes and VCF files from existing datasets or describe pipeline that can produce the necessary input.