Stage 3:

Reproducing Science

Project 4a:

The Genome aggregation database (GnomAD) is a reference population database that is a powerful tool for understanding the biological function of genetic variation. The paper describing GnomAD was published in 2022 and they reported ~200K clinically relevant human variants from 190K individuals. In 2024, they now have ~800+K individuals with over 800M variants.

Your task is to reproduce **figure 2** of their 2022 paper using the new data available on their 2024 paper. Compare and contrast and report the new things they have found.

2022 Paper: https://onlinelibrary.wiley.com/doi/10.1002/humu.24309
Data Source: https://gnomad.broadinstitute.org/
NB: You do not have to use exactly the same color set

Project 4b: Global Salmonella Concord genomics

Antimicrobial resistant Salmonella enterica serovar Concord (S. Concord) is known to cause severe gastrointestinal and bloodstream infections in patients. You are presented with over 200 samples from ethiopia. Your task is to reproduce **figure 2** of this paper using any 50 random samples from the paper; i.e. resolve the phylogenetic relationship between this samples and their antimicrobial resistance profiles.

Paper: https://www.nature.com/articles/s41467-023-38902-x
Data Source: Accession number of samples on SRA/ENA

Project 4C: Genetic Interaction Study using CRISPRi-Seq
Genetic interaction studies are important for our understanding of the molecular mechanism of how genomes work and by extension, the entire cellular machinery. The baker's yeast is a very important microorganism for global food security and understanding its genetics will help us know how to engineer it for improved taste and fermentative abilities. CRISPRi-Seq enables high throughput study of the importance of numerous genes across different environmental conditions. In this paper, the authors generated multiple double gene knock-down mutants across different conditions to study the effect of these genes on bacterial fitness.
Your task is to reproduce **figure 1** of this paper using all input 17000 strains.
Paper:
https://genome.cshlp.org/content/29/4/668.full?sid=a057d3b6-606c-441d-8ea0-c33bbef9c2d2#sec-15
Data Source: SRA

Project 4D:

A CRISPRi-Seq screen for functional assessment of BRCA1 mutants.
Genetic mutations in BRCA1, which is crucial for the process of DNA repair and maintenance of genomic integrity, are known to significantly increase the risk of breast and ovarian cancers. There are over 700 reported variants of BRCA1 on ClinVar; which mutations have the most significant consequences? Your task is to answer this question and reproduce **figure 1** of this paper using all input mutants.
Paper: https://www.nature.com/articles/s41388-019-0968-2#Fig1
Data Source:
https://www.ncbi.nlm.nih.gov/sra?linkname=bioproject_sra_all&from_uid=529534

Tutorials
1: Complete this Whole Exome Sequencing tutorial:
https://training.galaxyproject.org/training-material/topics/variant-analysis/tutorials/exome-seq/tutorial.html

Data Source: https://zenodo.org/records/3054169

2: Complete Prokaryotic Genome Annotation with this tutorial:
https://genomics.sschmeier.com/ngs-annotation/#annotation-with-prokka

3: Complete this Variant Calling Tutorial:
https://genomics.sschmeier.com/ngs-variantcalling/

4: Complete the Phylogeny Tutorial (Use IQ-TREE):
https://genomics.sschmeier.com/ngs-orthology/

5: Identification of AMR genes in an assembled bacterial genome:
https://training.galaxyproject.org/training-material/topics/genome-annotation/tutorials/amr-gene-detection/tutorial.html

Submission:

This is a teamwork output; i.e., we expect a single document that the entire team agrees and approves for submission. We expect you to submit a github link for the team containing the following:

- A one-slide presentation file containing a flowchart of the pipeline (To be submitted by wednesday). This flowchart will be defended by a representative of the team.
- A one-page scientific brief describing the project you chose under the following headings: Background/Introduction, Significance, Aims, Pipeline, Results in Brief, Discussion. (You are allowed to have an extra page only for the purpose of referencing)
- A 1-2 minutes video presentation describing the importance of the paper, the analysis, the results and how these can be translated for the betterment of human health.
- All codes that were used to analyze the datasets and generate the figures.