Project Name:
Single-Cell Clustering of Retinal Bipolar Neurons

Background:
Single-cell RNA sequencing (scRNA-seq) enables fine-grained resolution of cell types based on transcriptomic profiles. In this project, we aim to reproduce parts of the study *"Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics"* (Shekhar et al., *Nature Neuroscience*, 2016), which identified subtypes of bipolar cells in the mouse retina using scRNA-seq. Understanding retinal neuron subtypes is critical for deciphering visual information processing and may inform regenerative or therapeutic strategies for vision disorders.

Goal:
Reproduce the clustering and cell type identification analyses shown in the original paper. Specifically, we aim to replicate:

- The dimensionality reduction and clustering of single cells using UMAP/t-SNE (e.g., Fig. 2 of the paper)
- Marker gene identification for each cluster
- Visualization of marker expression with violin and heatmaps

We may optionally explore different clustering parameters or pipelines to compare robustness.

Dataset:

- **Source**: Sequence Read Archive (SRA)
- **Accession**: SRP058181 (BioProject PRJNA274464)
- **Type**: Single-cell RNA-seq (mouse retina)
- **Size**: Approx. ~10–20 GB (downsampled if necessary)

How to Start (Tools Needed):

- fastqc: read quality check

- kallisto | bustools: pseudoalignment and quantification
- scanpy (Python) or Seurat (R): dimensionality reduction, clustering, marker gene detection
- matplotlib, seaborn, pandas: data wrangling and plotting
- Optionally: bioconductor, scran for normalization

Stretch Goal(s):

- Compare different clustering resolutions or algorithms
- Evaluate cluster robustness and alignment with known cell type markers
- Identify potentially novel or uncharacterized subpopulations

---

Example:

Project:	
Background (paragraph)	The reference human genome sequence is inarguably the most important and widely used resource in the fields of human genetics and genomics. However, it's use introduces innate biases towards the sequence variants present in the reference. A collection of many human genomes, organized into the human pangenome reference, can reduce the number of missing variants for downstream analysis. However, the effect of introducing missing variants has not been quantified in many common analysis including GWAS, Functional Analysis, and Pathogenicity. Nor is it well understood how these effects vary by genome coverage or by population group.
Goal	Can we replicate results from HPRC initial paper determining that we can recover more variants using the pangenome.
Dataset	- Human Pangenome Reference subset down to 2 regions. (1 highly variable and 1 without much variation) - 1000 Genomes reads over those regions with simulated phenotype information
Suggest way to start (tools)	1. Align reads a. to pangenome using **graphaligner** or **giraffe** b. to linear reference genome with **bwa-mem** or other 2. Call variants a. from pangenome set using **vg** b. from linear set using **deep variant** or **broad tool** 3. Downstream analysis a. Use **pandas** to determine patterns in types of variants b. Use **plink2** to run GWAS using both sets of variants c. Pathogenicity something
Stretch Goal	1. Determine how these effects change by coverage and population group 2. Quantify the effect the additional variants have on downstream analysis including GWAS and determine if patterns in pathogenicity or variant type emerge.