Genome Data Download for Assembly Pipeline

This document outlines the data acquisition process for the genome assembly pipeline. It includes downloading raw sequencing data and reference genome files essential for downstream analysis.

Table of Contents

1. Data Acquisition

Tools Used: - **SRA Toolkit** (`prefetch`, `fastq-dump`)

Purpose: - Download sequencing reads from the NCBI Sequence Read Archive (SRA). - Convert `.sra` files into FASTQ format for downstream analysis.

```
#!/bin/bash
set -e

# Download SRA data
prefetch SRR9620862 --progress

# Convert SRA to FASTQ format
fastq-dump --split-files --outdir . SRR9620862/SRR9620862.sra
```

2. Reference Genome and Annotation Download

Tools Used: - **wget** (for downloading files from NCBI FTP servers)

Purpose: - Obtain the reference genome and annotation files for downstream alignment and annotation steps.

```
#!/bin/bash
set -e

# Create directory for reference genome
mkdir -p Reference_Genome
cd Reference_Genome

# Download reference genome and annotation files
wget ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/009/085/GCF_000009085.1_ASM908v1/GCF_000
wget ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/009/085/GCF_000009085.1_ASM908v1/GCF_000

cd ..
```

3. Execution Instructions

1. **Save the script** as `data.sh`.
2. **Make it executable**:

   ```
   chmod +x data_download.sh
   ```

3. **Run the script**:

   ```
   ./data_download.sh
   ```

This script ensures all necessary data is downloaded and formatted correctly for the genome assembly pipeline.