RNA-Seq Data Acquisition Script

This repository contains a Bash script (`data.sh`) that automates the initial data acquisition steps for an RNA-Seq pipeline. The script downloads RNA-Seq data from the SRA, retrieves the reference genome indices (for HISAT2), and downloads the GTF file for gene feature annotation (used in featureCounts).

Overview

The `data.sh` script performs the following steps: 1. **Download RNA-Seq Data from SRA:**
Uses the SRA Toolkit's `prefetch` command to download raw RNA-Seq data for both controlled and infected samples.
- Controlled samples: SRR11412215 and SRR11412216
- Infected samples: SRR11412230 and SRR11412229
2. **Download the Reference Genome Indices:**
Downloads the HISAT2 index for the human genome (GRCh38) from AWS S3 and extracts the tarball. 3. **Download the GTF File for FeatureCounts:**
Retrieves the Homo sapiens GTF annotation file (release 106) from Ensembl.

After executing these steps, the script prints out the total elapsed time.

Prerequisites

Before running the script, ensure you have the following installed: - **Conda or a compatible Bash environment - SRA Toolkit:** For the `prefetch` command. - **wget:** To download files. - **tar:** To extract the downloaded genome tarball.

You should also have appropriate permissions to write to the working directory (in this case, `/mnt/d/NGS/RNA_Seq`).

Usage

1. **Navigate to your project directory:** "'bash cd /mnt/d/NGS/RNA_Seq
2. **Run the script:**

```
bash data.sh
```

The script will:

Download the specified SRA files. Change into the Reference_Genome directory to download and extract genome indices. Download the GTF file for featureCounts. Return to the main directory and display the elapsed time.