Tutorial: How to convert fasta file to a count matrix for sequencing data

The nf-core RNA-seq pipeline is a powerful tool for processing and analyzing RNA sequencing data. It uses popular tools like STAR, RSEM, HISAT2, and Salmon to perform tasks such as alignment, quantification, and quality control.

To get a read count matrix from sequencing fasta/fastq files, you can use the nf-core RNAseq pipeline. This pipeline is designed to analyze RNA sequencing data obtained from organisms with a reference genome and annotation. The pipeline takes a samplesheet and FASTQ files as input, performs quality control (QC), trimming and (pseudo-)alignment, and produces a gene expression matrix and extensive QC report.

Prerequisites

1. **Nextflow:** Nextflow is a workflow management system that is required to run the nf-core RNAseq pipeline. You can install Nextflow with the following the instruction
    - curl -s https://get.nextflow.io | bash
2. **Docker:** The pipeline is best run using Docker. Install Docker from here.
3. **NF-core**: Download the NF-core pipeline that you need for your sequencing data. For example, in case of bulk RNA seq, you can use:
    - nextflow pull nf-core/rnaseq
4. **Unix-like operating system** (Linux, macOS) and enough storage space to store the sequencing files

Step 1: Running the Pipeline

1. **Prepare the Samplesheet**: Prepare your input files: The nf-core RNAseq pipeline requires a samplesheet and FASTQ files as input. You will need to create a samplesheet with information about the samples you would like to analyze. It must be a comma-separated file with 4 columns, and a header row as shown below:

```
1. sample,fastq_1,fastq_2,strandedness
2. CONTROL_REP1,AEG588A1_S1_L002_R1_001.fastq.gz,AEG588A1_S1_L002_R2_001.fastq.gz,auto
3. ...
```

2. **Run the pipeline**: Replace <SAMPLESHEET> with the path to your samplesheet file and <OUTDIR> with the desired output directory.

```
1.  nextflow run nf-core/rnaseq --input <SAMPLESHEET> --outdir <OUTDIR> --genome GRCh38 -profile
    docker
```

3. **Updating the Pipeline:** To ensure you are running the latest version of the pipeline, regularly update the cached version

```
1.  nextflow pull nf-core/rnaseq
```

4. **Reproducibility**: Specify a pipeline version when running the pipeline to ensure that a specific version of the pipeline code and software is used. For example:

```
1.  nextflow run nf-core/rnaseq -r 3.12.0 --input <SAMPLESHEET> --outdir <OUTDIR> --genome GRCh37 -
    profile docker
```

5. **Conclusion**: The nf-core RNA-seq pipeline provides a robust and flexible solution for processing and analyzing RNA sequencing data. By following this tutorial, you can easily generate a read count matrix from your sequencing fasta/fastq files. For more detailed information, consult the official documentation:
    a. Usage Guide
    b. Output Documentation
    c. nf-core RNA-seq GitHub Repository