Assignment: RNA-Seq Contaminant Detection Pipeline Development using Nextflow

Objective:

To develop a Nextflow pipeline that processes two RNA-Seq samples and performs both global and localized contaminant detection using the Fastv tool. The pipeline should downsample the original datasets and provide detailed analyses in parallel for each sample.

Dataset Information:

You will be working with two samples from an RNA-Seq dataset. These samples can be downloaded from the following links:

- **Sample 1**: *P. nigrescens exposed HEKa RNA (Rep 1)*
  - Read 1: SRR25233843_1.fastq.gz
  - Read 2: SRR25233843_2.fastq.gz
- **Sample 2**: *Media control HEKa RNA (Rep 1)*
  - Read 1: SRR25233831_1.fastq.gz
  - Read 2: SRR25233831_2.fastq.gz

Task Breakdown:

1. **Downsampling**:
   - Use the tool **Seqtk** to randomly downsample 1 million reads from both R1 and R2 of each sample.
     - **Seqtk tool**: https://github.com/lh3/seqtk
     - The resulting downsampled FASTQ files should be used for further analysis.
2. **Pipeline Development**:
   - Develop a **Nextflow** pipeline that takes the paired-end downsampled reads (R1+R2) as input and performs two main analyses for each sample using the **Fastv** tool.
     - **Fastv tool**: https://github.com/OpenGene/fastv

Required Analyses:

a) Global Contaminant Detection:

- Compare the downsampled 1M FASTQ files of each sample against the entire **bacterial + viral contaminant k-mer database**.

- The analysis should generate an output in both **HTML** and **JSON** formats.
- **Database link**: http://opengene.org/microbial.kc.fasta.gz

b) Localized Detection (Specific to *P. nigrescens*):

- Perform a localized comparison of the FASTQ files against the k-mer and genome fasta files of ***P. nigrescens***.
- **K-mer file**: P. nigrescens k-mer
- **Genome fasta file**: P. nigrescens genome

Pipeline Requirements:

- The pipeline should be designed to run analyses for both samples **in parallel**.
- Provide comprehensive **documentation** that includes:
    - Step-by-step instructions on how to launch the pipeline.
    - Clear explanations of the parameters required for Fastv and other tools.
    - Description of the expected output formats (HTML, JSON).

Deliverables:

- Nextflow pipeline script.
- Downsampled FASTQ files for each sample.
- Output files (HTML and JSON) for both global and localized analyses for each sample.
- Pipeline documentation.

Additional Notes:

- Ensure that the pipeline can be easily extended or modified for additional samples or analyses.
- Focus on creating a modular and well-documented pipeline to ensure ease of use and reproducibility.