Bash Program to separate DNA sequences by a single barcode and place them in individual folders/files.

This is a Bash program called **bash_barcodes.sh** that separates DNA sequences by a single barcode and puts them in folders/files named after the barcode. It is designed to be run from the Linux Ubuntu command line. Put the program in the same folder as the required files indicated below. Then navigate to the folder in question, provide the command to make it executable, and run it from the command line. See details on how to do this below.

The program requires two files:

- A text tab-delimited file with three columns indicating the (1) barcode sequence, (2) the reverse complement of the barcode sequence (to identify fragments that sequenced in the opposite direction), and the barcode name. In the example below, this file is called **1barcodes_f_b.txt**. No column headers should be included in the file. It can be created by putting the data in an Excel file and saving it as a tab-delimited text file.
- A file with the DNA sequences to be searched in FASTA format. In the example below, this file is called **1all.fasta.**

The program can be created in any text editor and saved with an .sh extension.

To run it, make it executable by navigating to the folder that it is in and running the following command from the command line:

- chmod +x bash_barcodes.sh

Then type the following command:

- ./bash_barcodes.sh

After it runs, you should see new folders created, each named after the barcode in the third column of the 1barcodes_f_b.txt file. Inside each folder, there will be one file with the sequences that contain the barcode in question in FASTA format.

The program is on the next page. Good luck!

Created by Windsor Aguirre using ChatGPT
Created: Sep/22/24

bash_barcodes.sh program:

```
#!/bin/bash

# Query file: 1barcodes_f_b.txt
# FASTA file: 1all.fasta

query_file="1barcodes_f_b.txt"
fasta_file="1all.fasta"

# Ensure the query file and fasta file exist
if [[ ! -f "$query_file" || ! -f "$fasta_file" ]]; then
  echo "Error: Query file or FASTA file not found!"
  exit 1
fi

# Read the query file line by line
while IFS=$'\t' read -r seq1 seq2 folder_name; do
  # Create a folder if it doesn't exist
  mkdir -p "$folder_name"

  # Grep the FASTA file for lines that contain either sequence 1 or sequence 2
  # Assume sequence lines don't start with '>' (FASTA headers start with '>')
  grep -E -B 1 "($seq1|$seq2)" "$fasta_file" > "$folder_name/matches_$folder_name.fasta"

  # Notify the user that the matches have been saved
  echo "Matches for $seq1 or $seq2 saved in $folder_name/matches_$folder_name.fasta"

done < "$query_file"
```

Note that the grep command adds an extra line and two dashes, "– ", between the sequences, so these should be removed to have the file in proper FASTA format. This can be done with the following two lines for each file if needed:

- sed 's/--//g' BCF1.fasta > BCF1_ed.fasta
- sed '/^$/d' BCF1_ed > BCF1_ed_b.fasta

Using the procedure, this has to be done for each file created. You could also write a program to do this for all of the new files, or add the code to the end of the program above.

Created by Windsor Aguirre using ChatGPT
Created: Sep/22/24