#!/bin/bash

# 1. Создание рабочей директории

mkdir -p terminal\_task/results

cd terminal\_task

# 2. Загрузка GFF-файла, если его нет

GFF\_URL="https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode\_human/release\_41/gencode.v41.basic.annotation.gff3.gz"

GFF\_GZ="gencode.v41.basic.annotation.gff3.gz"

GFF="gencode.v41.basic.annotation.gff3"

if [[ ! -f "$GFF\_GZ" ]]; then

echo "Скачивание GFF-файла..."

wget "$GFF\_URL"

fi

# 3. Разархивация

gunzip -kf "$GFF\_GZ"

# 4. Извлечение e-mail с русским доменом

echo "Поиск e-mail с русским доменом (пример):"

grep -E -o '[a-zA-Z0-9.\_%+-]+@[a-zA-Z0-9.-]+\.(рф|москва|рус|дети|онлайн|сайт|ком|орг)' "$GFF" | head

# 5. Извлечение IP-адресов (примерно, если файл содержит их)

echo "Поиск IP-адресов (пример):"

grep -Eo '([0-9]{1,3}\.){3}[0-9]{1,3}' "$GFF" | head

# 6. Поиск файлов с genome или GCA в имени

echo "Поиск файлов с 'genome' или 'GCA' в имени (в текущей директории):"

find . -type f \( -iname '\*genome\*' -o -iname '\*GCA\*' \)

# 7. Фильтрация генов по условиям

echo "Фильтрация генов по условиям..."

awk -F'\t' '

BEGIN { OFS="\t" }

$3 == "gene" {

if (match($9, /gene\_name=([^;]+)/, arr)) {

gene = arr[1]

chr = $1

start = $4

end = $5

strand = $7

if (strand == "-" && length(gene) > 3 && gene !~ /^[AZHА-Я]/) {

print gene, chr, start "-" end, strand

}

if (strand == "+" && gene ~ /^[AZHА-Я]/) {

print gene, chr, start "-" end, strand

}

}

}' "$GFF" > results/genes\_filtered.tsv

# 8. Извлечение транскриптов

echo "Извлечение транскриптов..."

awk -F'\t' '

$3 == "transcript" {

if (match($9, /gene\_name=([^;]+)/, arr)) {

print arr[1], $9

}

}' "$GFF" > results/all\_transcripts.tsv

# 9. Объединение: гены + транскрипты

echo "Объединение таблиц генов и транскриптов..."

join -t $'\t' -1 1 -2 1 <(sort results/genes\_filtered.tsv) <(sort results/all\_transcripts.tsv) > results/final\_gene\_table.tsv

# 10. Сводная таблица

echo "Генерация сводной таблицы..."

awk -F'\t' '

{

chrom = $2

len = split($3, coords, "-") ? coords[2] - coords[1] + 1 : 0

strand = $4

chr\_gene\_count[chrom]++

chr\_gene\_total\_len[chrom] += len

if (strand == "+") plus++

else if (strand == "-") minus++

}

END {

print "Хромосома\tГенов\tСредняя\_длина"

for (c in chr\_gene\_count) {

avg = chr\_gene\_total\_len[c] / chr\_gene\_count[c]

print c "\t" chr\_gene\_count[c] "\t" avg

}

print "\nГенов на + цепи: " plus

print "Генов на - цепи: " minus

}' results/final\_gene\_table.tsv > results/summary\_table.tsv

echo "Готово. Все результаты сохранены в: terminal\_task/results/"