PROJECT2-INSTRUCTIONS

May 15, 2024

1 Identifying trait-associated tandem repeats from selected African populations

1.1 1. Overview

Tandem repeats (TRs) comprise one of the major sources of genetic variation in humans. The need for accurate and reliable genotyping of TRs feeds into a broader goal of understanding population variation dynamics and the contribution of TRs to disease risk and complex traits. TRs are typically excluded from analyses such as genome-wide association studies, which focus instead on common single nucleotide polymorphisms. However, the recent proliferation of genotyping tools for tandem repeat analysis presents an opportunity to close this gap. This project aims to characterise TR variation within different subpopulations of the African cohort of the 1000 Genomes Project and identify TRs associated with expression of nearby genes. Each group will perform genotyping in this cohort to identify and catalogue tandem repeats for the African subpopulations. Groups will then identify TRs associated with expression of nearby genes. The significance of this work is to deepen our understanding of population variation dynamics based on tandem repeat analyses.

1.1.1 Dataset

The 1000 Genomes Project provides whole-genome sequencing data from individuals representing 26 distinct populations across the globe. For this project, we will look at subpopulations of the African cohort and compare the subpopulations to the European cohort. For a subset of samples, gene expression data is available.

1.1.2 TR Genotyping Tools

HipSTR, TRTools, and custom scripts for association analyses.

1.1.3 Expected Results

On successful implementation of the project, participants should be able to: * Perform STR genotyping and quality filtering across a population of individuals * Identify common and unique tandem repeats for each sub-population * Perform association tests between STR variation and complex traits (gene expression) * Explore the potential biological implications of the TR variations including impacts on gene expression

1.2 2. Getting started

Follow the instructions in `PROJECT2-SETUP.ipynb` to install the required tools for this project.

We will be working with the following datasets in the `~/public/project2-association` directory:

1.2.1 Data for performing STR genotyping

- BAM files and indices for 89 samples from African ancestry are located at `dataset/AFR_bam`
- BAM files and indices for 363 samples from European ancestry are located at `dataset/EUR_bam`
- The HipSTR reference file for GRCh38 is located at `dataset/hg38.hipstr_reference.bed`
- You can find the reference genome at: `~/public/genomes/Homo_sapiens_assembly38.fasta`

1.2.2 Data and script for performing gene expression association analyses

- Covariates to use in linear association for gene expression are saved in `dataset/covariates_all.csv`
- Normalized gene expression for African samples in Geuvadis dataset are saved in `dataset/AFR_normalized_and_filtered_hg38_chr11.csv`
- Normalized gene expression for European samples in Geuvadis dataset are saved in `dataset/EUR_normalized_and_filtered_hg38_chr11.csv`
- The script for association analysis is `project2_association_script.py`

1.3 3. Running HipSTR

We will first perform multi-sample genotyping using HipSTR on samples from each ancestry seperately. To save time, we are only analyzing a small region of the genome. We can first extract repeats from the HipSTR reference corresponding to the region where we have data:

```
# Restrict to reference STRs within chr11:57520464-57529754
cat ~/public/project2-association/dataset/hg38.hipstr_reference.bed | \
    awk '($1=="chr11" && $2 >=57520464 && $3 <= 57529754)' > hipstr_ref_small.bed
```

Now, we can run HipSTR on this small reference set of STRs:

```
# Create a text file with a list of BAMs to genotype
ls ~/public/project2-association/dataset/AFR_bam/*bam > AFR_bam_files.txt

# Run HipSTR. The command below assumes HipSTR is somewhere on your $PATH (e.g. $HOME/bin)
HipSTR --bam-files     AFR_bam_files.txt \
         --fasta         ~/public/genomes/Homo_sapiens_assembly38.fasta \
         --regions       hipstr_ref_small.bed \
         --str-vcf       AFR_calls.vcf.gz

# Index the output VCF with tabix
tabix -p vcf AFR_calls.vcf.gz
```

These steps will write per sample genotype information for the target repeat in `AFR_calls.vcf.gz`. Repeat the same procedure for European samples.

1.4 4. Running dumpSTR

Before moving forward we will want to filter low quality genotypes based on quality score using dumpSTR as below:

```
dumpSTR  \
    --vcf AFR_calls.vcf.gz \
    --out AFR_calls_filtered \
    --vcftype hipstr \
    --hipstr-min-call-Q 0.9 \
    --hipstr-min-call-DP 10

bgzip AFR_calls_filtered.vcf
tabix -p vcf AFR_calls_filtered.vcf.gz
```

This will output a VCF file similar to HipSTR output, with low quality calls filtered and replace with no call. Repeat the same procedure for European samples.

1.5 5. Inspect repeat length distributions by population

You can use statSTR to make plots of the length distribution of alleles in each population:

```
statSTR --vcf AFR_calls_filtered.vcf.gz --plot-afreq --out AFR
statSTR --vcf EUR_calls_filtered.vcf.gz --plot-afreq --out EUR
```

This will output pngs with histograms of repeat lengths at each STR (e.g. `AFR-chr11-57528484.pdf`, which you can view in datahub). Compare the distributions for each STR across populations to identify any cases where the length distributions look substantially different.

1.6 6. Peform the association analysis

Summarize the genotype information and sample names in the filtered VCF file with bcftools:

```
bcftools query -f "%CHROM\t%POS\t%PERIOD\t%END\t[%GB\t]\n" AFR_calls_filtered.vcf.gz > AFR_GB.t
bcftools query -l AFR_calls_filtered.vcf.gz > AFR_names.txt
```

Run the python script provided to find if expression of any gene is associated with the copy number of the target repeat. This script performs a linear regression for each STR-gene pair.

```
python3 ~/public/project2-association/project2_association_script.py AFR 11 AFR_names.txt AFR_(
```

This will output a file `AFR_expr_results.tab` giving association statistics for each STR-gene pair.

Repeat the above procedure for European samples.

1.7 7. Write a report summarizing your results

1. Did you identify any STRs that show different length distributions across different populations?
2. Did you find any STR significantly associated with expression of a gene in this region? Was it associated in one or both populations?

3. As a bonus, you can make a plot of STR genotype vs. expression for any significant associations identified.

Each group should submit a written report and presentation to: https://bit.ly/3WvnUqW