GenomePrep

Release 0.1

chandicelu

Apr 04, 2022

CONTENTS

GenomePrep is a Python library for genetics enthusiasts that prepare direct-to-consumer DNA data for analysis using popular bioinformatics tools. It performs the following tasks:

1. Parse SNPs from popular DTC providers
2. Check for missing calls and duplicated SNPs
3. Determine assembly, and sanity check for similarity to the reference genome
4. Deduce the genotyping array version
5. Apply a genotyping-array-based SNP sanity filter (optional)
6. Convert to 23andMe-like format and VCF format for downstream analysis

This project was developed on the goodwill of over 7,000 open genome data made public between 2011 and 2020, addressing the problem of processing raw DTC DNA data in the context of the present: genotype arrays.

More information about genotyping array, ancestral relatedness can be found in our paper on CSBJ:

C. Lu, B. Greshake Tzovaras, J. Gough, A survey of direct-to-consumer genotype data,and quality control tool (GenomePrep) for research, Computational and Structural Biotechnology Journal(2021)

You can upload and process your genome on server.

You can also download preprocessed open genomes. Check out the *Usage* section for further information, including how to *Installation* the project.

Note: This project is under active development.

CHAPTER

ONE

CONTENTS

1.1 Usage

1.1.1 Installation

Source code from Github :

```
$ git clone git@github.com:changlubio/GenomePrep.git
```

1.1.2 Setup

To use GenomePrep, first download relevant files:

```
$ make datadir
$ cd datadir
$ wget tp://ftp.ensembl.org/pub/release-75/fasta/homo_sapiens/dna/Homo_sapiens.GRCh37.75.
↪dna.toplevel.fa.gz
$ gunzip Homo_sapiens.GRCh37.75.dna.toplevel.fa.gz
$ wget https://supfam.mrc-lmb.cam.ac.uk/GenomePrep/datadir/api.23andme.com
$ wget https://supfam.mrc-lmb.cam.ac.uk/GenomePrep/datadir/badalleles.dat
$ wget https://supfam.mrc-lmb.cam.ac.uk/GenomePrep/datadir/RS2GRCh37Orien_1.dat
$ wget https://supfam.mrc-lmb.cam.ac.uk/GenomePrep/datadir/THE_LIST.dat
```

Use liftOver for build other than GRCh37

```
$ wget ftp://hgdownload.cse.ucsc.edu/goldenPath/hg38/liftOver/hg38ToHg19.over.chain.gz
$ wget http://hgdownload.soe.ucsc.edu/goldenPath/hg18/liftOver/hg18ToHg19.over.chain.gz
```

1.1.3 Running test DNA

Run GenomePrep on a typical 23andMe file

```
$ cd ..
$ bin/process.py tutorial/testgenome.zip -d ./datadir -o ./outputs -i vcfindex
```