Tommaso Alfonsi
Master Student

Expected graduation date: April 2020

1000 Genomes Project: data integration and identification of control populations

Supervisor: Prof. Marco Masseroli
Co-supervisors: Anna Bernasconi and Arif Canakoglu
Collaborating with: Laura Riva (Sanger UK)

POLITECNICO
MILANO 1863

What we did:
Integration into a GDM repository

Main issues:

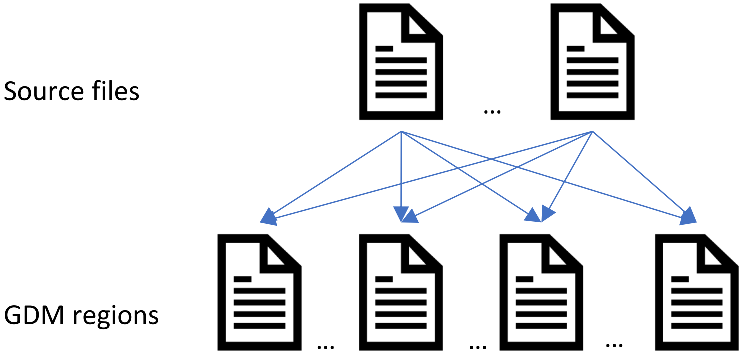
- Volume of the data

  ~ 88 million mutations 2500 individuals 4 TB

Solutions:

To reduce the execution time:
Dynamic source code generation and compilation
Parallel transformation

- Shape of the transformation



Extended capabilities of the framework Metadata-Manager to accommodate this kind of transformations

- Source mutations encoded in VCF
  (1-based coordinates, largely extensible format, mostly single-end)

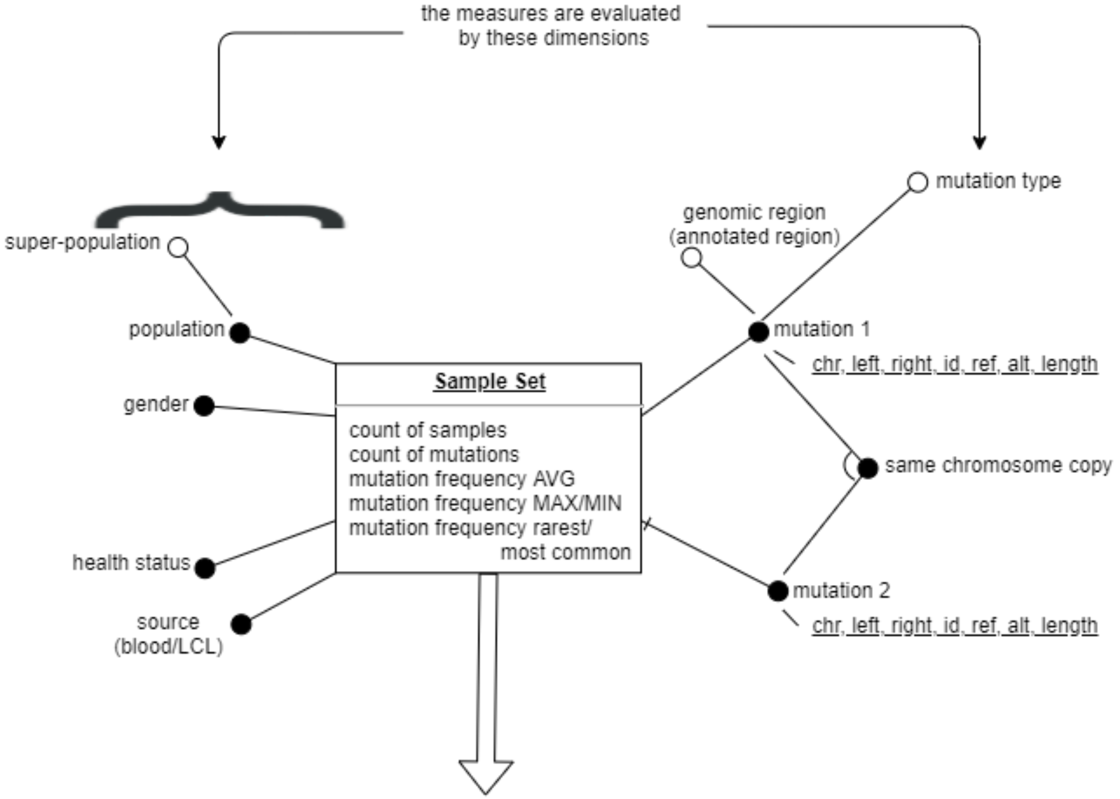
Development of tools for managing VCF and
conversion into GDM regions
(0-based coordinates, paired-end)

What we are doing: Identification of control populations

Use case:

- User can select the desired characteristics:
  - Country of provenance
  - Gender
  - Origin of the sample, i.e. blood / LCL
  - Assembly
  - Having some user-provided mutations

The application returns summary statistics evaluated on the free dimensions, e.g. The most frequent mutations in the South Asian individuals, aggregated by provenance and gender.

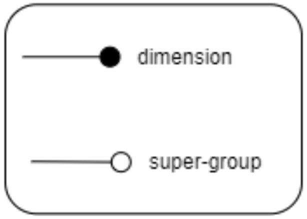


the measures are evaluated by these dimensions
super-population
population
gender
health status
source (blood/LCL)
Sample Set
count of samples
count of mutations
mutation frequency AVG
mutation frequency MAX/MIN
mutation frequency rarest/ most common
mutation type
genomic region (annotated region)
mutation 1
chr, left, right, id, ref, alt, length
same chromosome copy
mutation 2
chr, left, right, id, ref, alt, length

these are the measures of the sample set.
These are all numbers so we will have for example:

pie charts for the cont of samples and
for the count of mutations

box plots for the mutation frequency with the AVG, MAX, MIN as annotations
and rankings for the most common / rarest mutations



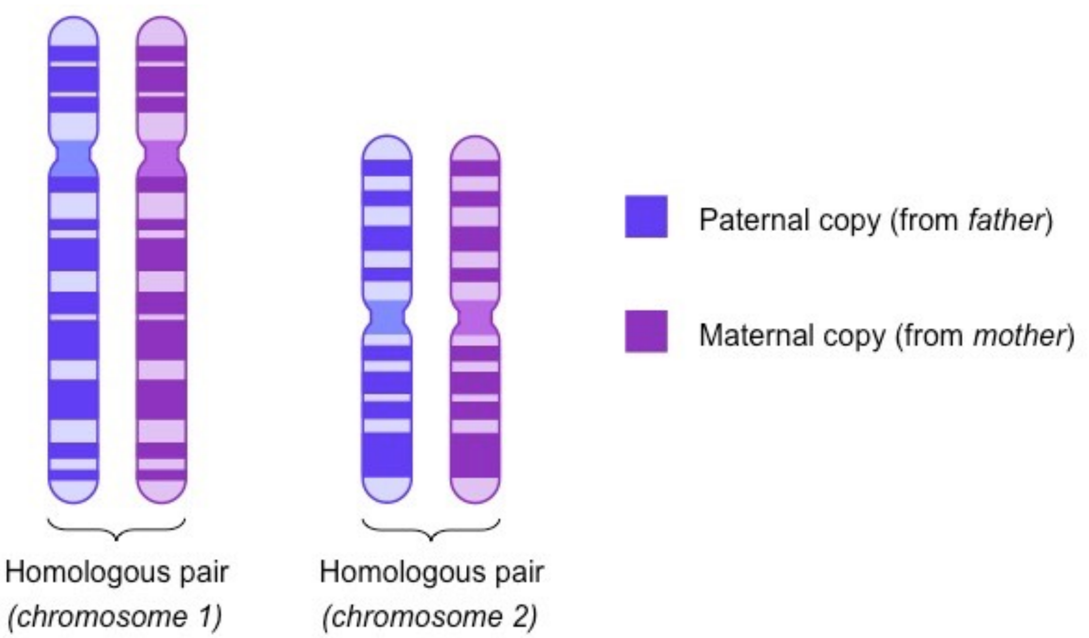
What we are doing: Identification of control populations

Use case:

- User can select the desired characteristics:
  - Country of provenance
  - Gender
  - Origin of the sample,
    i.e. blood / LCL
  - Assembly
  - Having some user-provided mutations

- also it is possible to specify if the nutations must be on the same / different chromosome copy (chromatid)

The 1000 Genomes Project mutations are phased, i.e. we know for each sample which chromosome copy shows the mutation.



e.g. Show the distribution by gender of the individuals from Puerto Rico on assembly hg19 and having mutations with id=rs123 and id=rs456 on the same chromosome copy

What we are doing: Identification of control populations

Issue:

- Response time when working with tables filled with billions of rows (~ 11 250 000 000 rows)

Solutions we thought of:

- Working on the definition of indexes for the region data (under development)
- Study the query plan and look for possible optimizations
- Save the query results into temporary tables and use them as a shared caching mechanism (under development)
- Possibly precompute some summary statistics offline