Enlarged clinical Belarusians' exomes: opportunities and restrictions of additional analysis

Danat Yermakovich, Aleh Liaudanski, IGC NAS, Minsk, Belarus bioinfgroup@igc.by

The new era of large NGS comes to Belarus. With the quite fast production of big data in wet labs, the problem of processing and analysis it raises up. We as a young bioinformatics group faced the necessarily recycling data for system analysis versus routine clinical investigations on the presence of pathogenic variants. According to the type of data, we chose the population genetics field. After the common variant calling from enlarged clinical exomes NGS Illumina data, we expect to get plausibly infers for the Belarusian population using generally known types of analysis. The obtained already PCA plot shows the distinction of Belarussians from other 1000G populations.

Motivation

Currently 200 enlarged clinical exomes are being sequenced within the institute, and we, a recently assembled bioinformatics group, would like to utilize the data for our own project. We see the restrictions of the data: the sample count, small number of variants, their distribution across the genome, the limitation to exonic variations, and the sample consists only of people with Mendelian diseases. Nevertheless, this will be the first systematically obtained genomic big data for Belarusians.

Aim

We plan to: a) Use this cohort for phasing and to impute from this cohort to other smaller data; b) Visualize the distinction between Belarusian from other 1000G populations via PCA plot; c) Conduct a GWAS. Half of the cohort have CAKUT, the other cardiomyopathies. The diseases seem to be Mendelian, but the rate of pathogenic's mutation finding is not high; d) Infer a demographic history; e) Make admixture with ancient populations. We are considering obtaining microarray data for the cohort and mixing it with clinical exome data in way of enlarging it.

METHODS

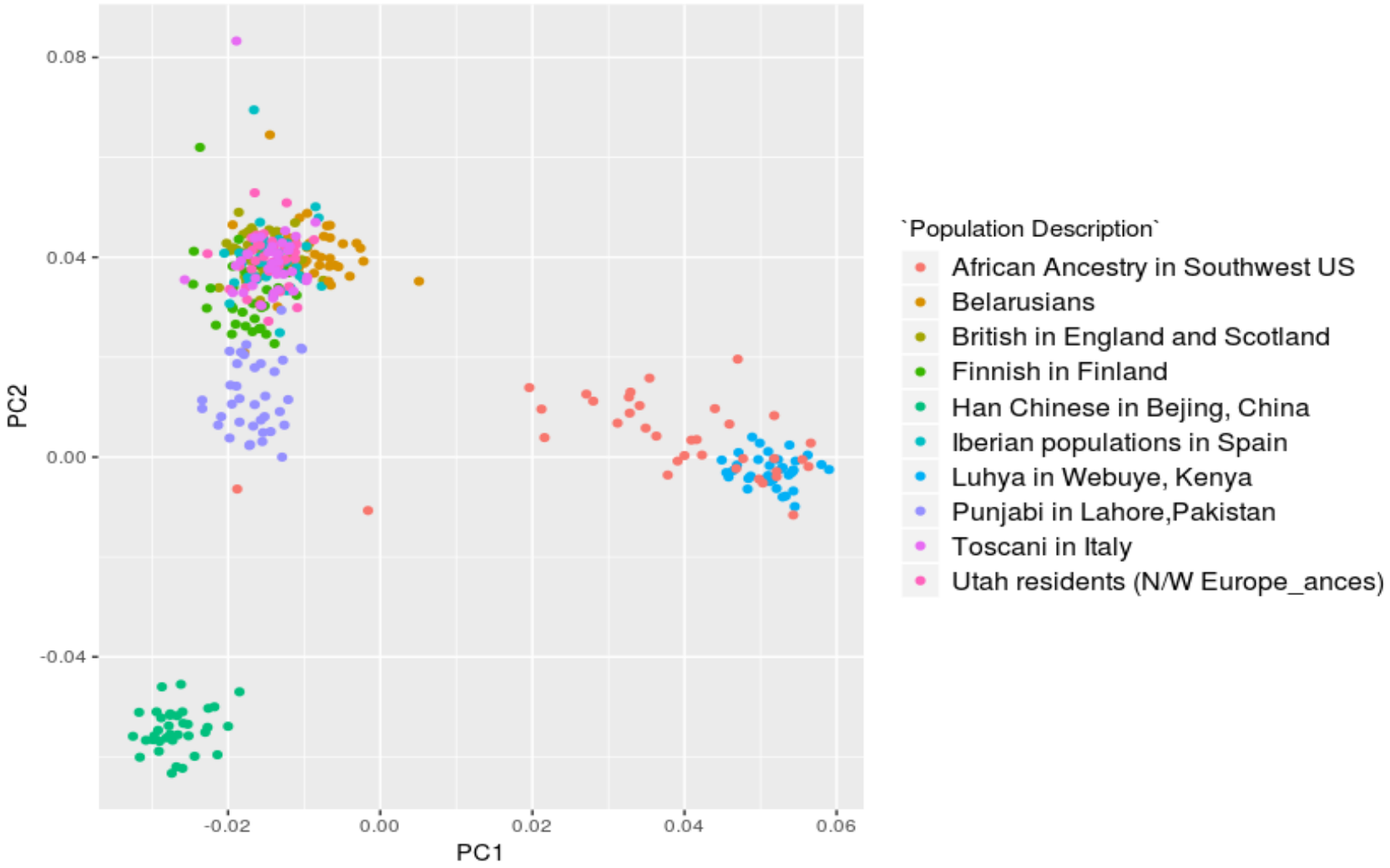
The targeted Illumina NGS is being carried out by usingx Gen®ExomeResearch Panel v1.0 (around 20000 genes). The analysis was made on received at the moment samples. The last pipeline includes the further steps: a) preprocessing: trimmomatic (cutting), bwa (aligning), samtools/picard (converting, marking, sorting), b) variant calling for clinical purposes: samtools/varscan2, freebayes, gatk 4 haplotype caller (calling), bcftools – (filtering, merging callers' results), annovar (annotating). c) variant calling for popgen purposes: freebayes, bcftools. Obtained cohort vcf-file was imputed and phased by shapeit, beagle, eagle in order to compare these tools. Also, the Michigan Imputation Server was tested. Finally, the pca plot was done by using smartpca from eigensoft. The input file for it was converted by plink. Currently, all our pipelines have been written on bash or Snakemake.

RESULTS

We tested our pipelines on 36 obtained yet samples and GIAB sample. By launching phasing/imputation tools, we found out the complexity of direct results' comparison of such tools due to the process of haplotype's assignment during their work. We're waiting for other samples to explore more aspects of the Belarussian population genetics. The first pca plot shows the relation of Belarusians to other nations.

ACKNOWLEDGMENT

The work is being done as part of the Belarus state research program "Biotechnology", 2019 - 2020; subprogram 2 "Structural and functional genomics", task 2.45



The PCA plot was made from a subset of 1000G (36 samples in each population; chr 12 13 18 19)