Searching for latent viruses in human whole-genome sequencing data

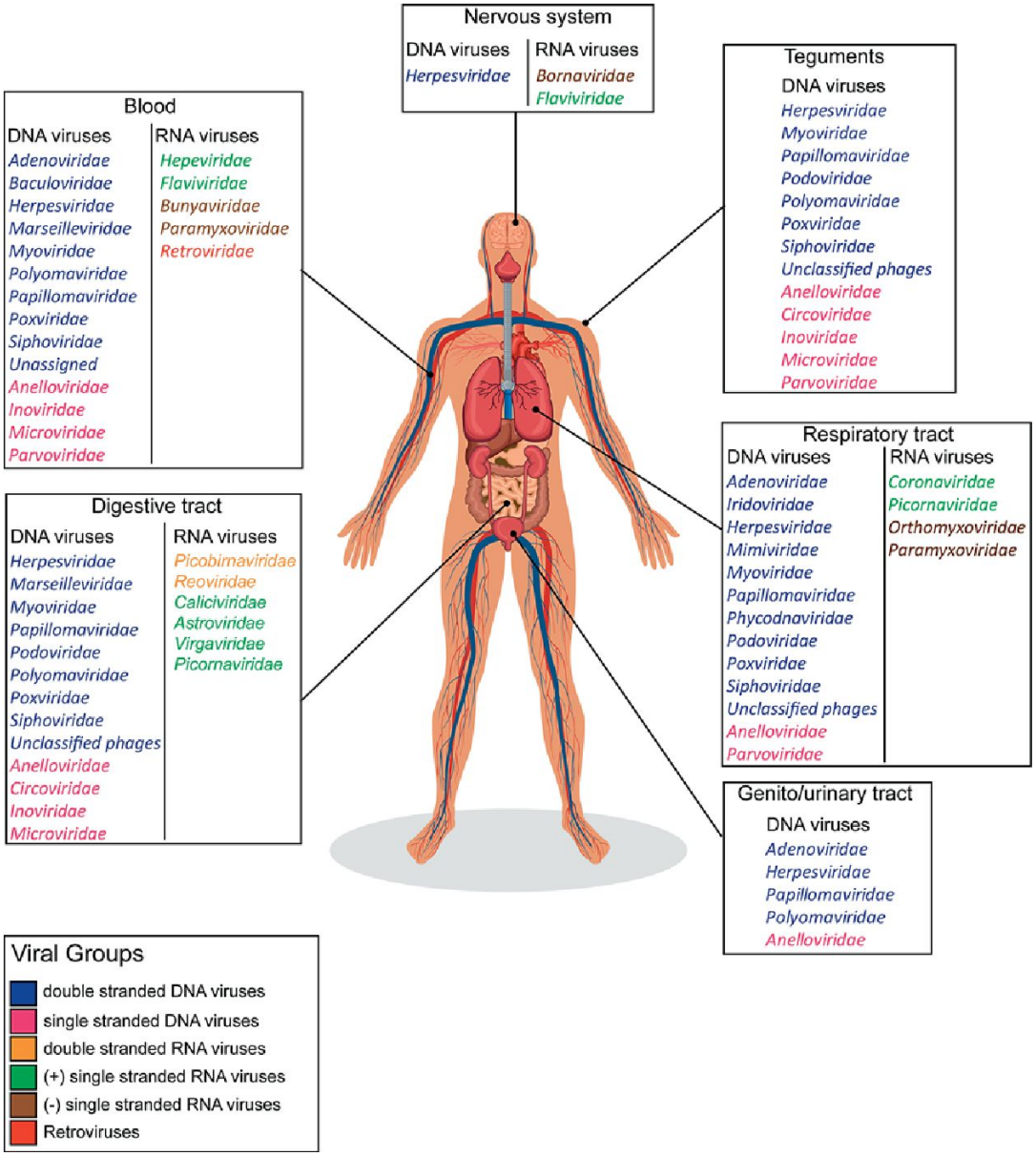
Supervisors (Genotek):
Valery Ilinsky
Alexander Rakitko

Students:
Yura Orlov
Alisa Morshneva
Nadezhda Pogodina

Blood virome

DNA viruses

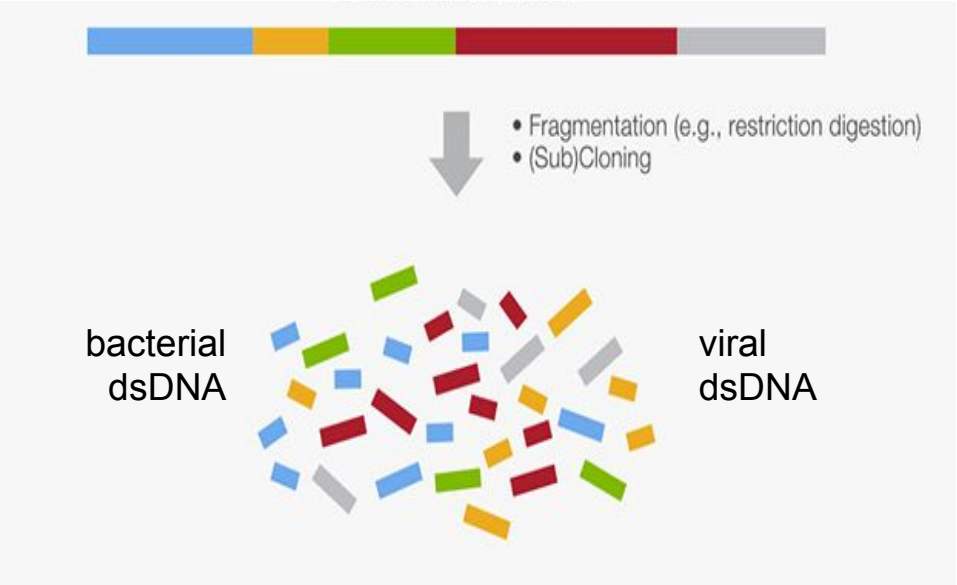
- *Adenoviridae*
- *Baculoviridae*
- *Herpesviridae*
- *Marseilleviridae*
- *Myoviridae*
- *Polyomaviridae*
- *Papillomaviridae*
- *Poxviridae*
- *Siphoviridae*
- *Anelloviridae*
- *Inoviridae*
- *Micoviridae*
- *Parvoviridae*



Popgeorgiev et al. 2013

Shotgun sequencing

Sampling → DNA extraction



thermofisher.com

Goals

Characterize viral representation in human WGS data and search for possible association between viral load and genetic variations

Tasks

1. Explore literature data
2. Create a pipeline for WGS data analysis
3. Find open WGS databases
4. Test pipeline on different samples
5. Count viral load in testing data
6. Create a table of viral load for samples from 1000 Genomes
7. Perform GWAS

Data samples

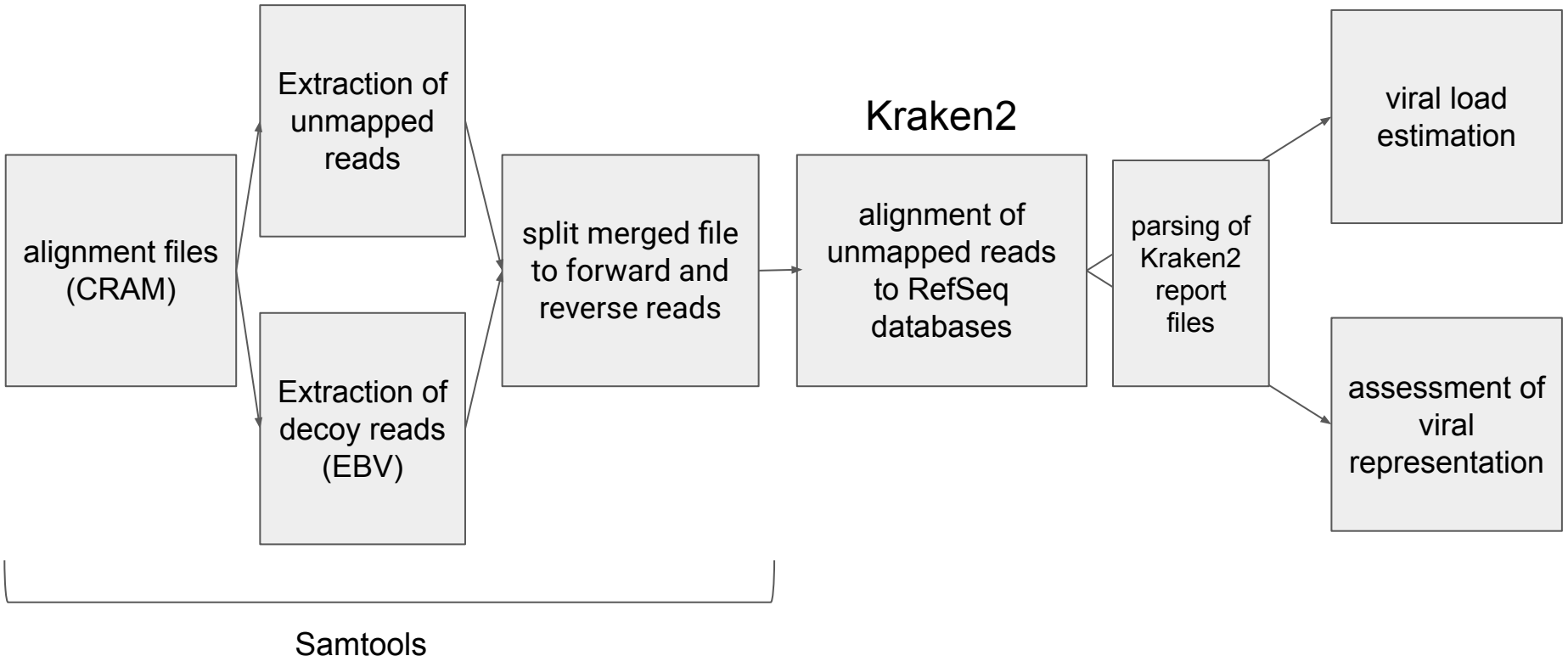
1. One woman's uterine tissue sample from Genotek for creating and testing pipeline
2. WES and WGS samples from Genotek
3. Samples from 1000 Genomes: cram files

Population	Number of samples
CEU	101
TSI	104
IBS	108
GBR	106
FIN	105

5 populations from 1000 Genomes have been analyzed

Pipeline

for the analysis of 1000 Genomes samples (bash script)



Patient with HPV infection

Sample of a woman's uterine tissue that is rich in papillomavirus 16, caused cervical cancer

Virus-only subset:

```
1090 Human papillomavirus type 16 (taxid 333760)
 426 Human mastadenovirus C (taxid 129951)
   9 dsDNA viruses, no RNA stage (taxid 35237)
   6 Pa6virus (taxid 1982251)
   3 Human papillomavirus 4 (taxid 10617)
```

Total diversity:

```
6181 Paracoccus yeei (taxid 147645)
3430 Cutibacterium acnes (taxid 1747)
3424 Paracoccus (taxid 265)
2274 Bacteria (taxid 2)
1616 Proteobacteria (taxid 1224)
1401 Comamonadaceae (taxid 80864)
1151 root (taxid 1)
1090 Human papillomavirus type 16 (taxid 333760)
 928 Micrococcus luteus (taxid 1270)
 643 Acidovorax sp. JS42 (taxid 232721)
 557 Paracoccus sp. Arc7-R13 (taxid 2500532)
 543 Sphingobium yanoikuyae (taxid 13690)
 432 Acidovorax (taxid 12916)
 426 Human mastadenovirus C (taxid 129951)
 387 Pseudomonas koreensis (taxid 198620)
 365 Alphaproteobacteria (taxid 28211)
 364 Acidovorax ebreus TPSY (taxid 535289)
 327 Diaphorobacter polyhydroxybutyrativorans (taxid 1546149)
 324 Pseudomonas (taxid 286)
 286 Homo sapiens (taxid 9606)
 251 Paracoccus mutanolyticus (taxid 1499308)
 234 Enterobacteriaceae (taxid 543)
```

Viral load estimation

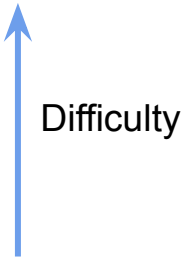
$$C = \frac{2 \times \frac{\text{number of reads mapped to virus genome}}{\text{virus genom size}}}{\frac{\text{number of reads mapped to human genome}}{\text{human genome size}}}$$

$$C = 2 \times \frac{\text{virus genome coverage}}{\text{human genome coverage}}$$

$$C = \frac{2 \times \frac{\text{number of reads unmapped to human genome classified as viral by Kraken}}{\text{virus genom size}}}{\frac{\text{number of raw reads mapped to human genome}}{\text{human genome size}}}$$

Approaches:

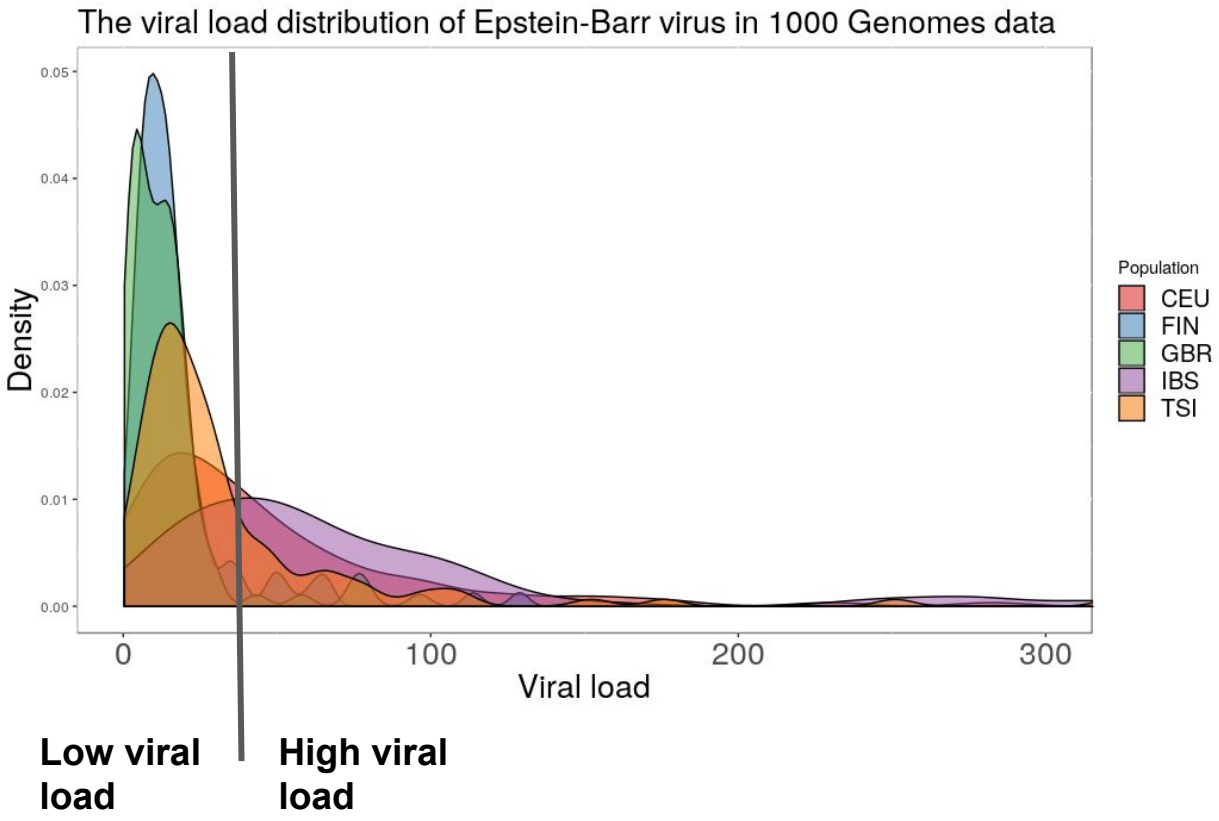
1. Using coverage values (median or average)
2. Using number of mapped reads
3. Using number of reads classified by Kraken



Viral representation in 1000 Genomes data

	Number of samples
Total	524
Human gammaherpesvirus positive	524
Human mastadenovirus positive	123

EBV load in 1000 Genomes data

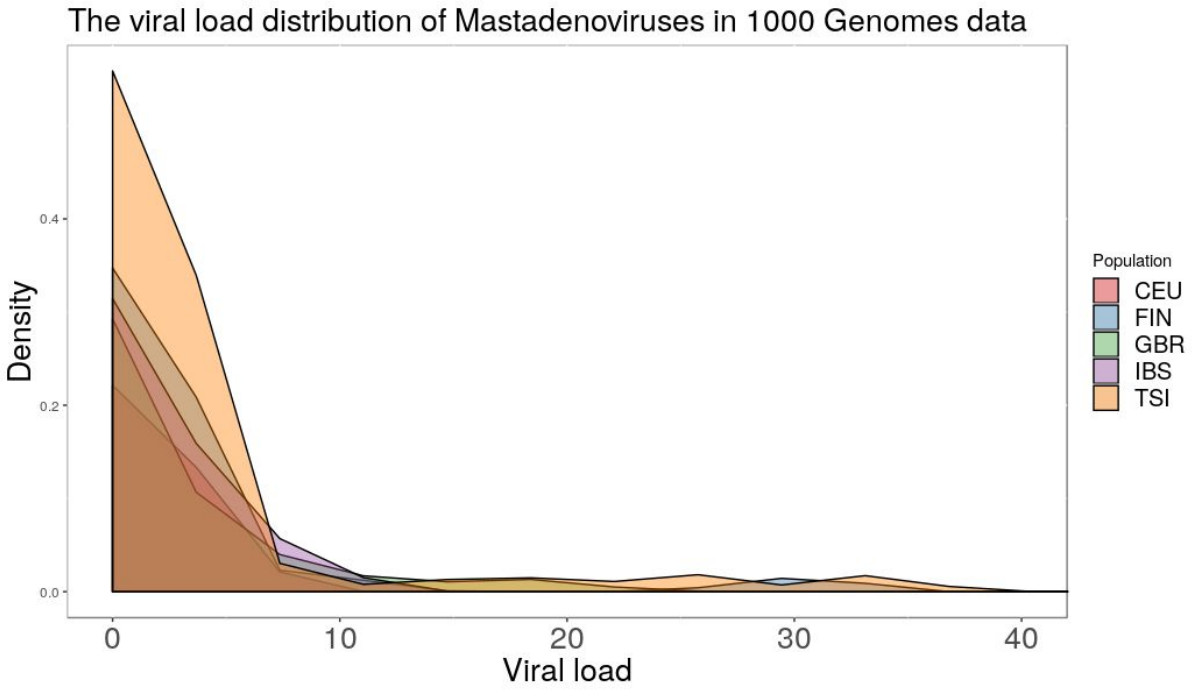


Low viral load

High viral load

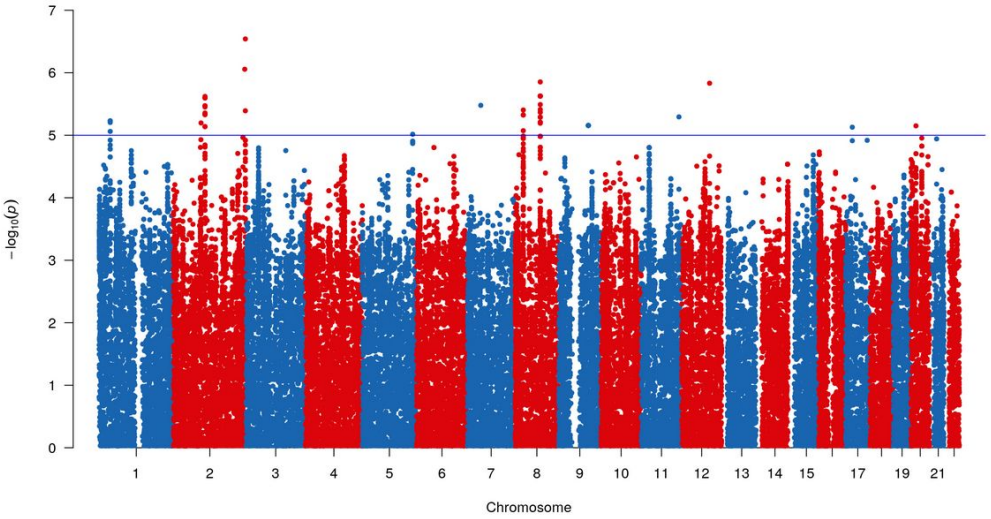
Group	Number of samples
Low EBV	305
High EBV	219

Mastadenovirus load in 1000 Genomes data



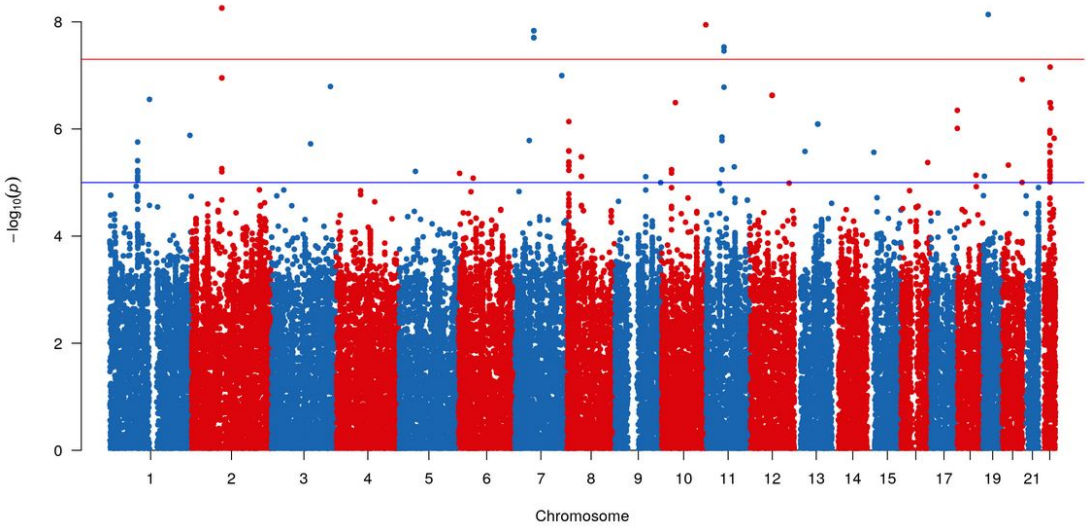
Group	Number of samples
Data with adenoviruses	123
Control	401

GWAS analysis (EBV)



SNP	Gene	P-value
rs6710977	LOC105373958 : Intron Variant	8.795e-7
rs11780074	None	0.000001404
rs10532235	LOC643339 : Intron Variant	0.000001472
rs10808854	None	0.000002363
rs2033302	LOC102724691 : Intron Variant	0.000002396

GWAS analysis (Mastadenoviruses)



SNP	Gene	P-value
rs28529415	None	6.983e-8
rs138432335	CDH4 : Intron Variant	1.187e-7
rs112308716	None	3.254e-7
rs529717656	ROCK1P1 : Non Coding Transcript Variant	4.495e-7
rs12681923	LOC105377793 : Intron Variant	7.282e-7

Project overall results

1. Pipeline for working with WGS data was created
2. Viral load of EBV and mastadenoviruses was counted
3. GWAS analyses was performed for 5 populations from 1000 Genomes

Thank you for your attention!

Project GitHub link:
https://github.com/Alisa1195/Searching-for-latent-viruses-in-human-whole-genome-sequencing-data