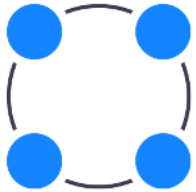


ИНСТИТУТ
БИОИНФОРМАТИКИ

Finding new V genes in genomic data

Ivan Sosin

Supervisors: Alexander Shlemov,
Timofey Prodanov,
Center for Algorithmic Biotechnology
St. Petersburg State University

Ultimate goal – V-genes validation

Sub goals:

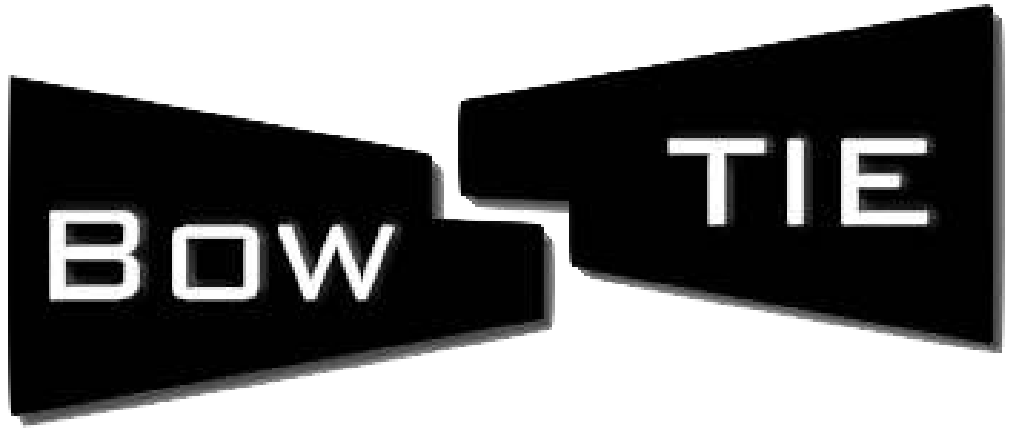
- Data correction
- Finding V-genes in reads from High coverage WGS (1000 genomes project)
- Finding specific sequences before and after genes

First steps

- Let’s check whether we can map reads for V-genes from well-established database.

- What tools do we have?

Toolkit: working horses



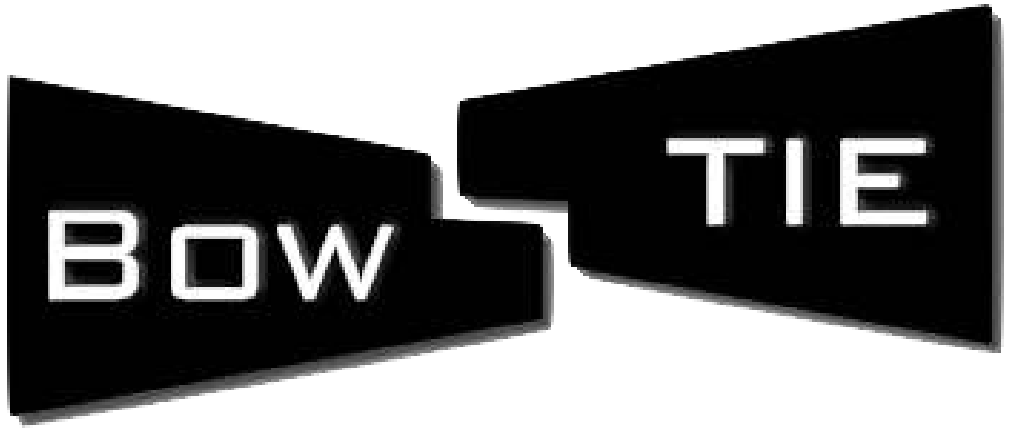
Bowtie 2
Fast and sensitive read alignment

SAMtools

SAM Tools provide various utilities for manipulating alignments in the SAM format

Workflow routine design

- Bowtie2
  - Indexing v-genes
  - Aligning reads to index

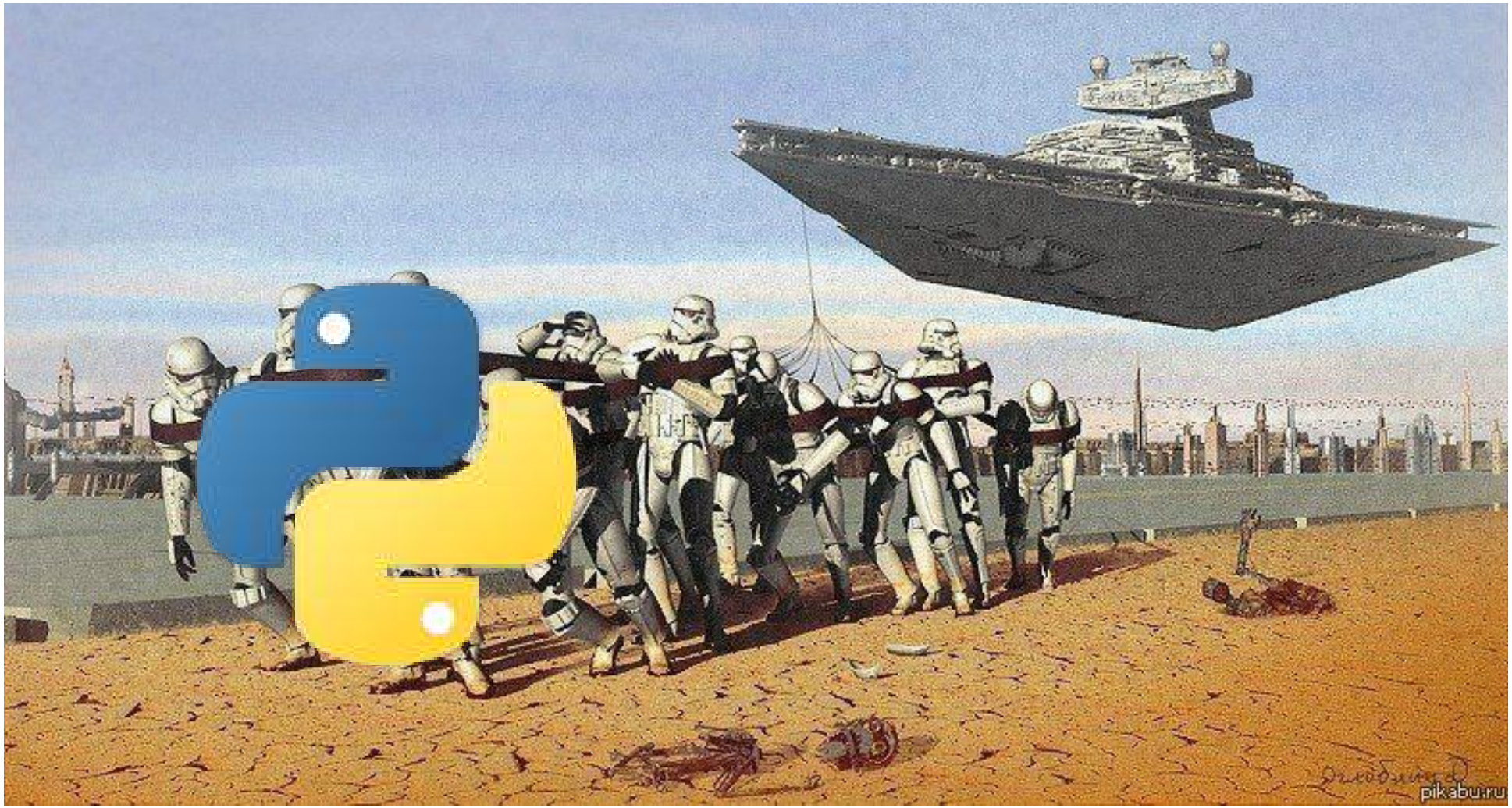


- SAMtools
  - Converting SAM to BAM
  - Sorting BAM
  - Indexing sorted
  - Getting stats

SAMtools

Ok. That’s clear.
What’s next?

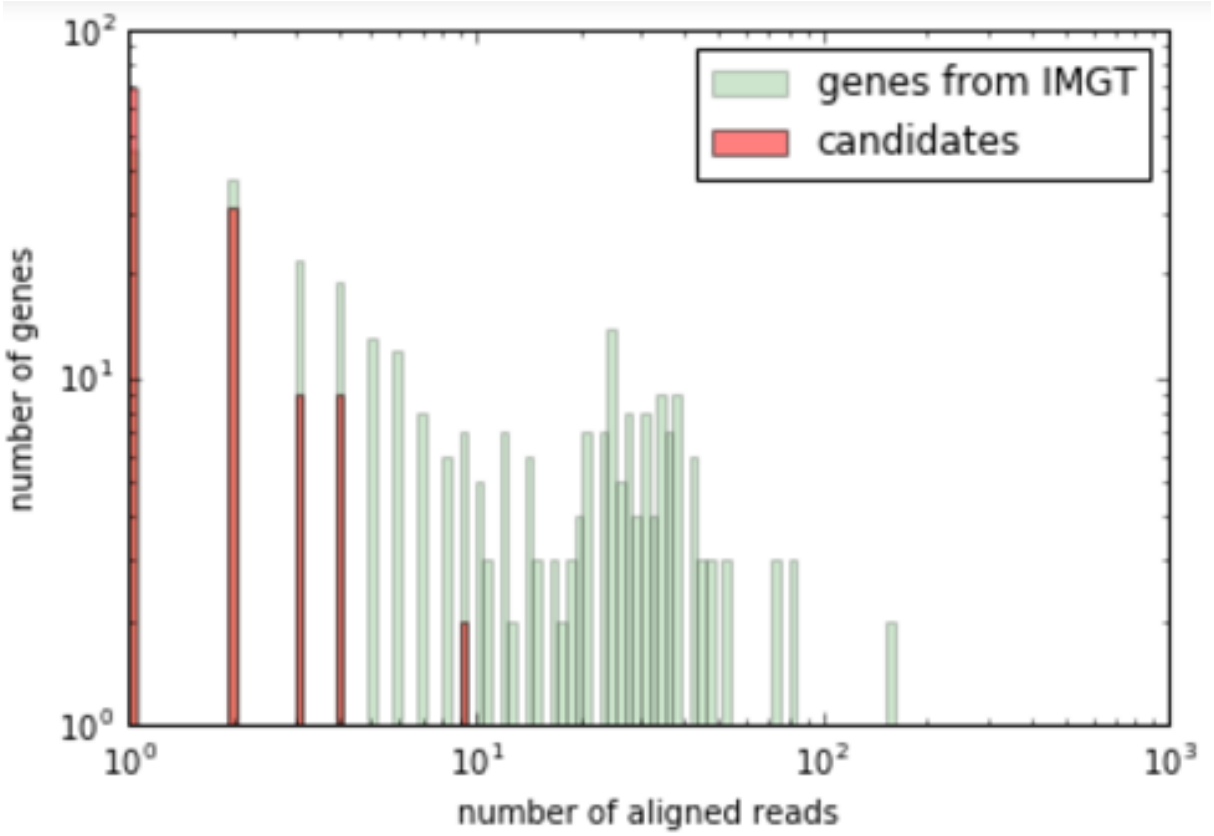
Make Python do it for us!



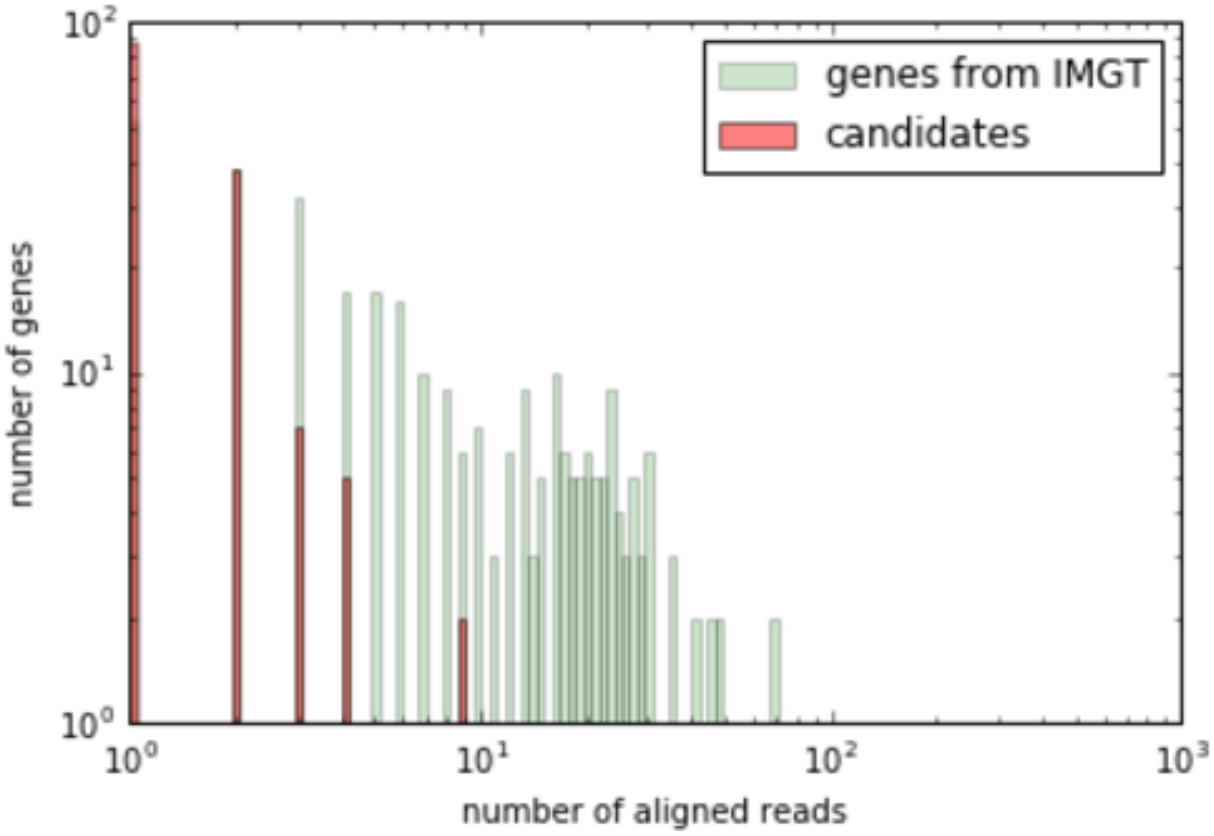

ВЖУХ
И РИДЫ ПРИЛОЖИЛИСЬ
risovach.ru

Paired vs Unpaired alignment

Paired



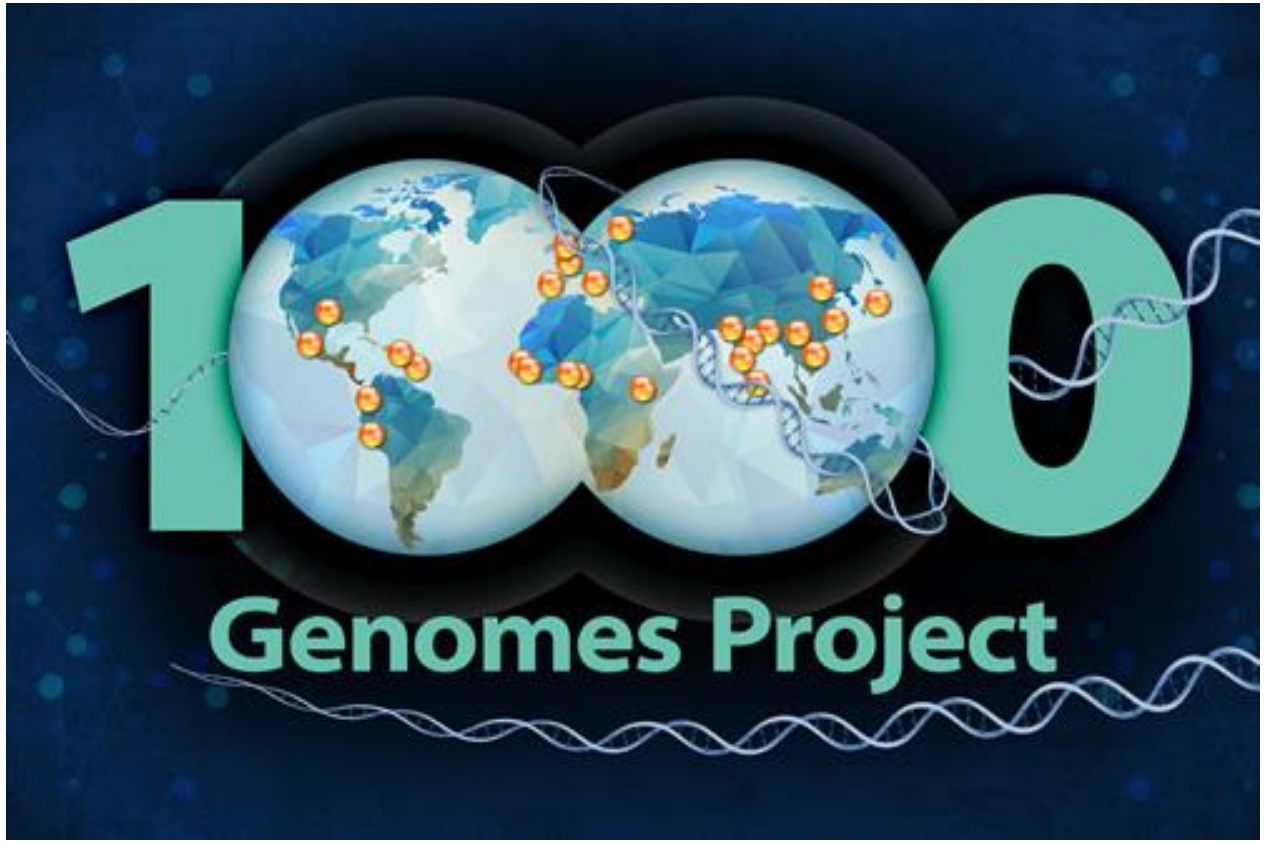
Unpaired



What about different populations?

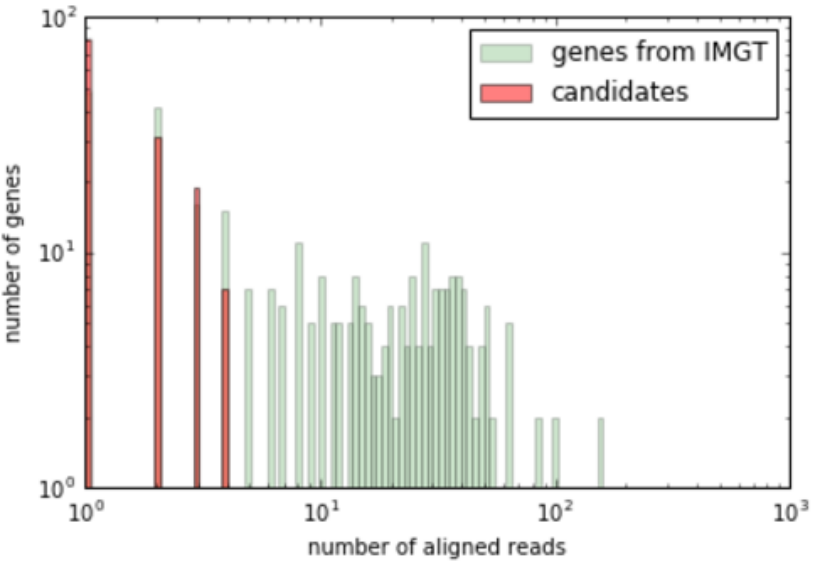
We took WGS to play with from:

- Great Britain,
- Chinese Han,
- Utah residents with Northern and Western European ancestry

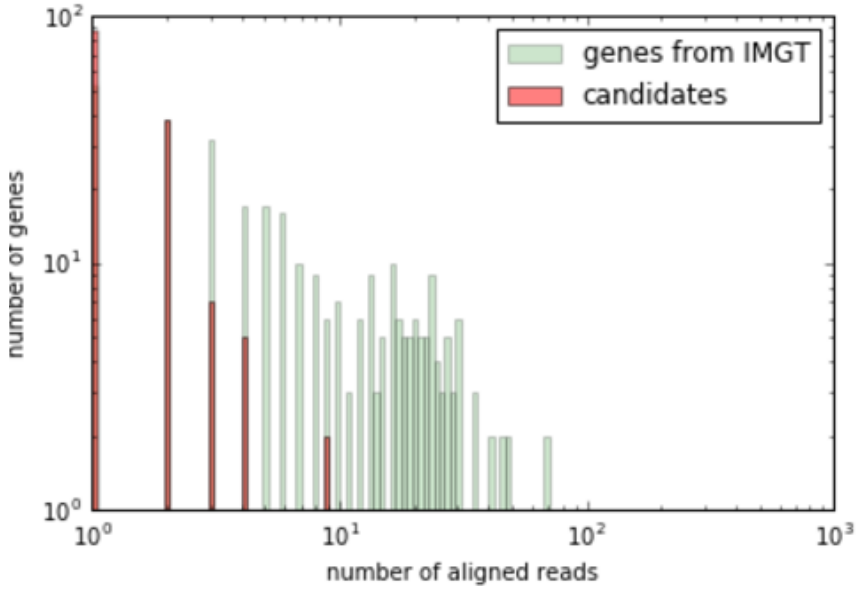


Chinese vs Great Britain vs Utah

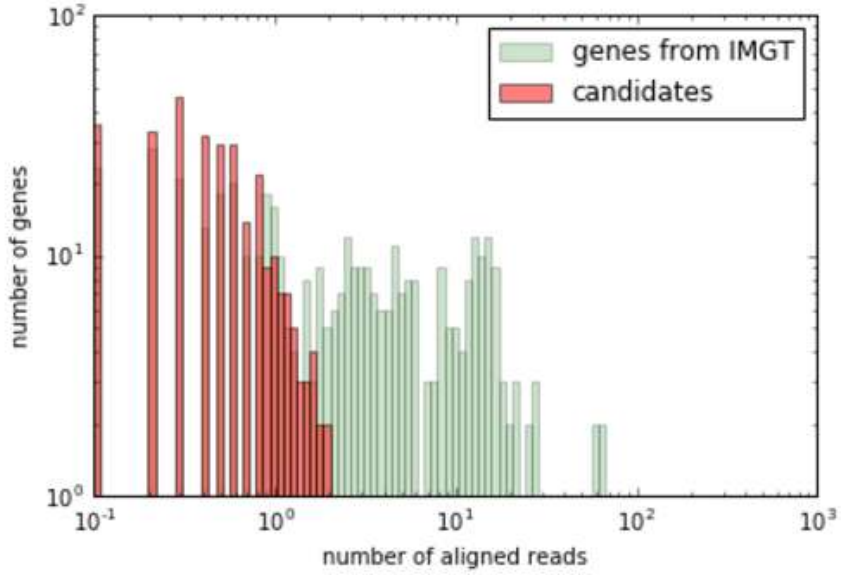
Chinese



Great Britain



Utah

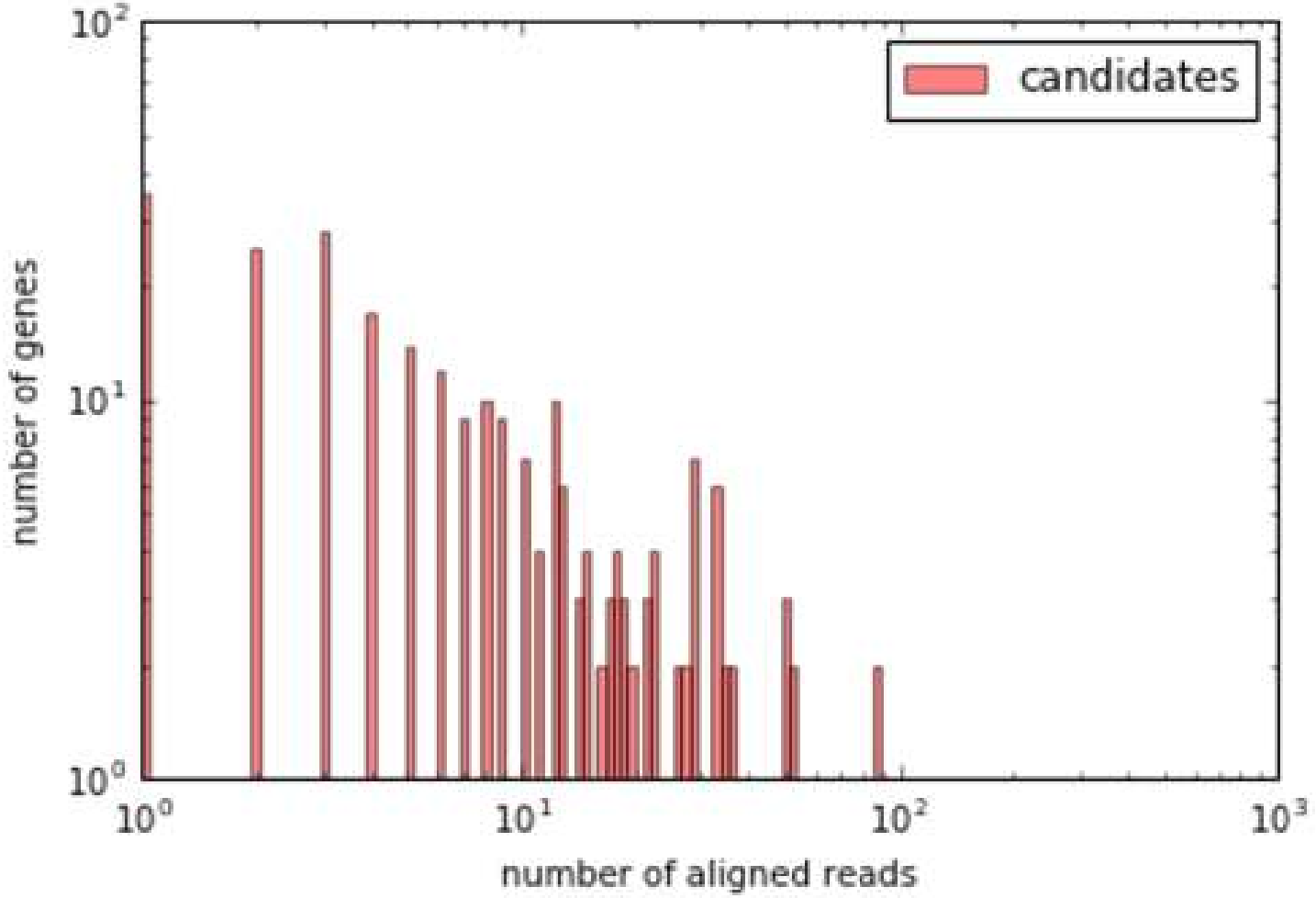


Unplanned discoveries

Several genes from IMGT database aligned to great number of reads > 100000.

Reasons are unknown.

Candidates only



Results

- I got my hands dirty in bioinformatics
- I got experience in working with bioinformatics tools

- We found several super-genes
- Several candidates showed interesting signals
- V-genes from different populations show similarity

What’s next?

- Investigate candidate’s alignment in order to find signal sequences
- Assemble reads mapped to candidates
- Check signal sequences
- Super-genes from database should be examined

Questions?