Data processing after immunoglobuline libraries sequencing

Kolmogorov Mikhail

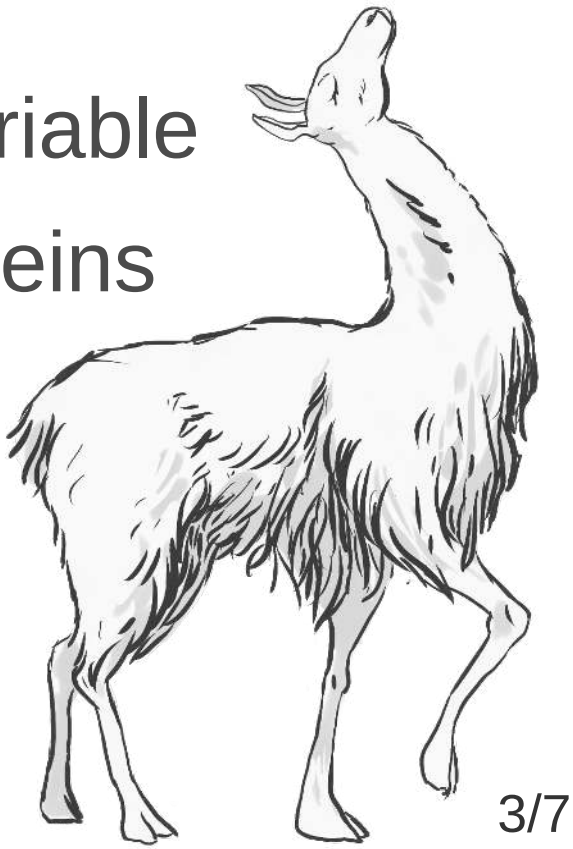
Supervisor: Karabelskij Alexandr (BioCad)

Task

- Lama`s hypervariable immunoglobuline regions are being sequnced using 454 pyrosequencing
- Need to obtain real sequences and discard/correct errors
- Need to classify immunoglobulines by homology, expression rate
- Analysis of typical substitutions

Challenges

- There is no suitable lama reference (even camel reference is not available)
- Big difference in immunoglobuline expressinon in immunized lama
- Hypervariable regions are hypervariable
- Indels makes comparison with proteins databases impossible



Technology

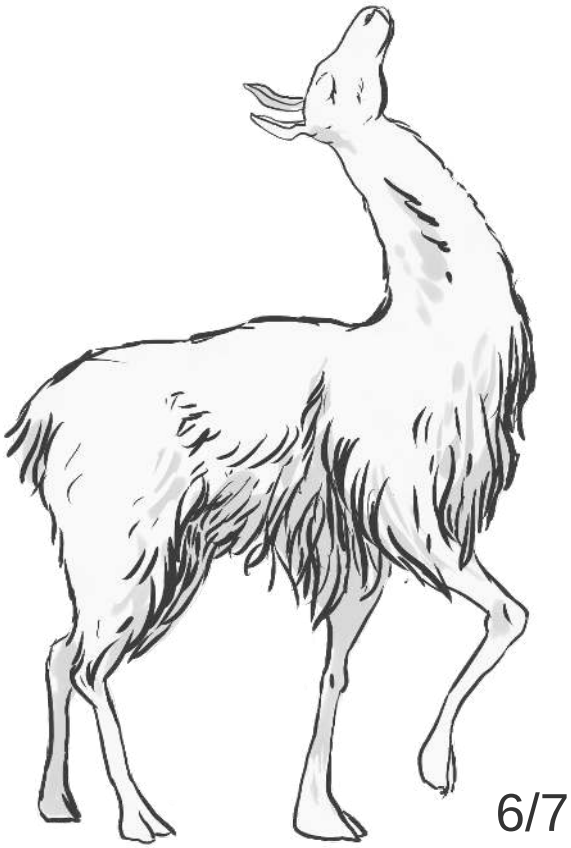
- 454 amplicon seqencing technology is used
- Read length is enough to fully cover antibody variable domain
- Forward and backward reads can be distinguished by barcodes

Thoughts

- Some general error correction (AmpliconNoise)
- Extract trusted reads (that have perfect matches between each other)
- Align rest of reads on trusted clusters
- Align reads on each other and define clusters as their consensus seuqence
- When ORF is restored, we can find regions and continue error correction according to them

Goal

- Ready-to-go pipeline for error correction and clusterisation of NGS data
- Creation of «Lama immunoglobuline database»
- Tables with typical substitutions



Lama says:

Thanks for attention!