Data processing after immunoglobuline libraries sequencing

Kolmogorov Mikhail

Supervisor: Karabelskij Alexandr (BioCad)

Task

- Llama`s hypervariable immunoglobuline regions are being sequnced using 454 pyrosequencing (RNA-seq)
- Need to obtain real sequences and correct errors
- Need to classify immunoglobulines by homology, expression rate

Challenges

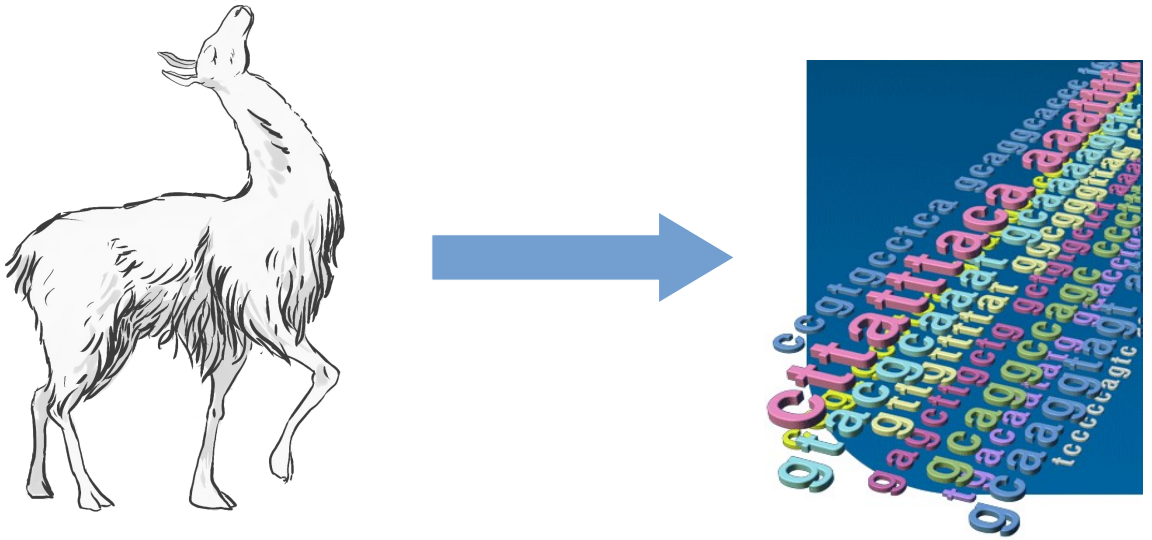
- There is no suitable llama reference (even germline)
- Non-uniform coverage (due to different expression rate)
- Hypervariable regions are hypervariable
- Indels as main error types

Technology

- 454 amplicon seqencing technology is used
- Read length is enough to fully cover antibody variable domain
- Forward and backward reading

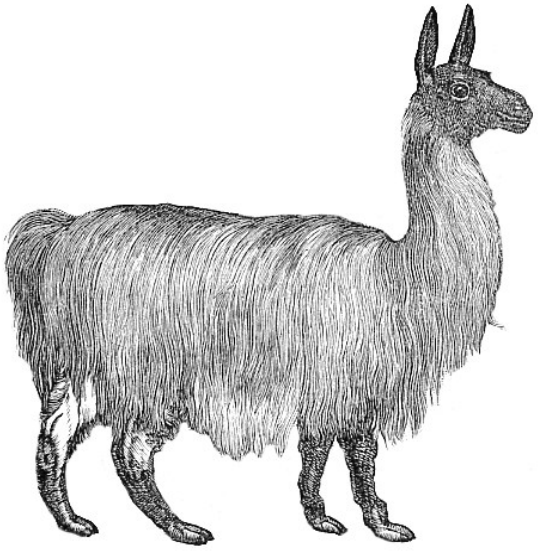
llama-fixer: 3 simple steps to clean sequences

- Step 1: Separating and clipping
- Step 2: Clustering by CDR3
- Step 3: Clustering by nucleotide sequence



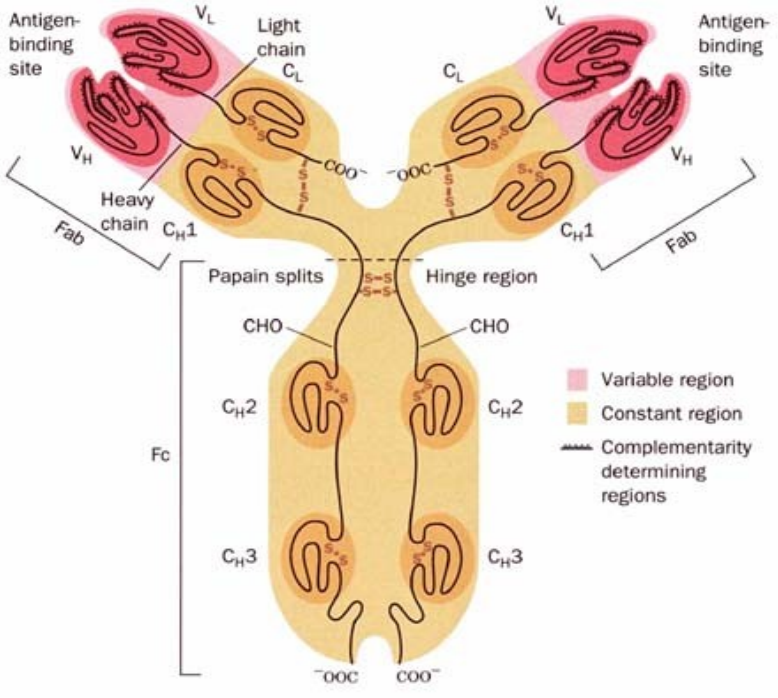
Step 1: Separating and clipping

- Reads separation by MID primer sequence (perfect match)
- Alignment on other primer parts
- Same for reverse direction
- Cut sequence between forward and reverse primers
- Throw away incomplete reads
- Remove duplicates



Hypervariable region structure

- Consists of 3 CDR and 4 FR regions
- CDR`s are highly variable
- Hypothesis: CDR3 is unique for every different variable region

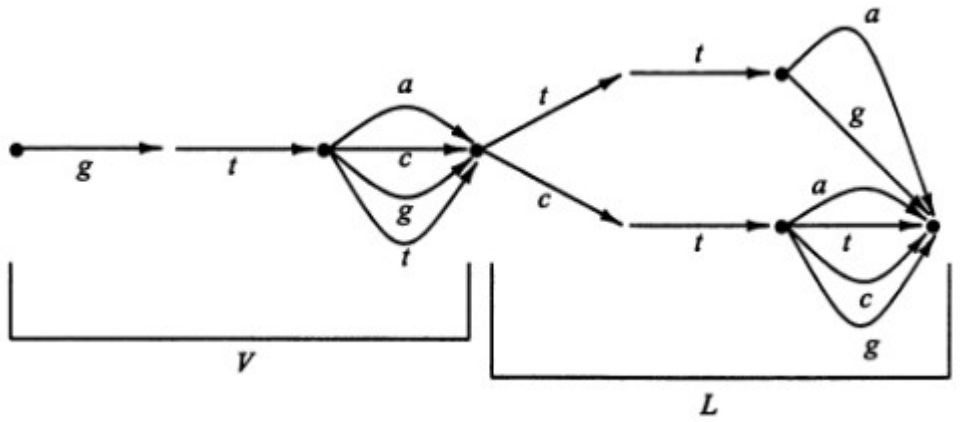


CDR3 extraction

- We don`t have germlines — can`t run alignment-based methods
- We know conservative sequences of end FR3 and begin FR4
- But they are presented as amino acids

Network alignment

- Alignment protein sequence on nucleotide sequence
- Possible nucleotide sequence for corresponding amino acids are represented as graph
- Generalization of ordinary local alignment

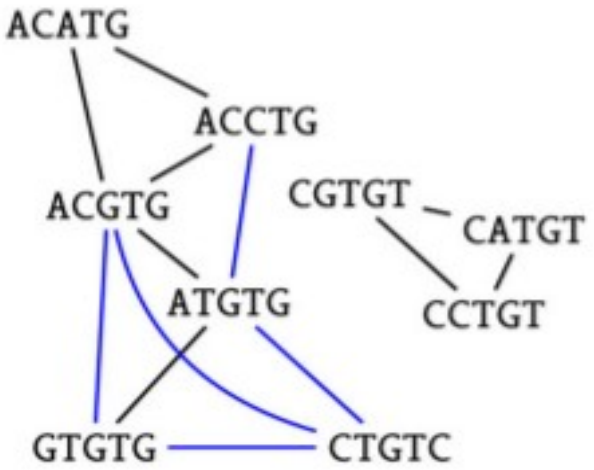


Step 2: Clustering by CDR3

- Extracting CDR3 from each read
- Graph clustering of CDR3 sequences (d = 2)
- Merging CDR3 that differ by 1-2 indels
- Merging CDR3 with 1-2 missmatches, if they correspond to clusters with similar consensus

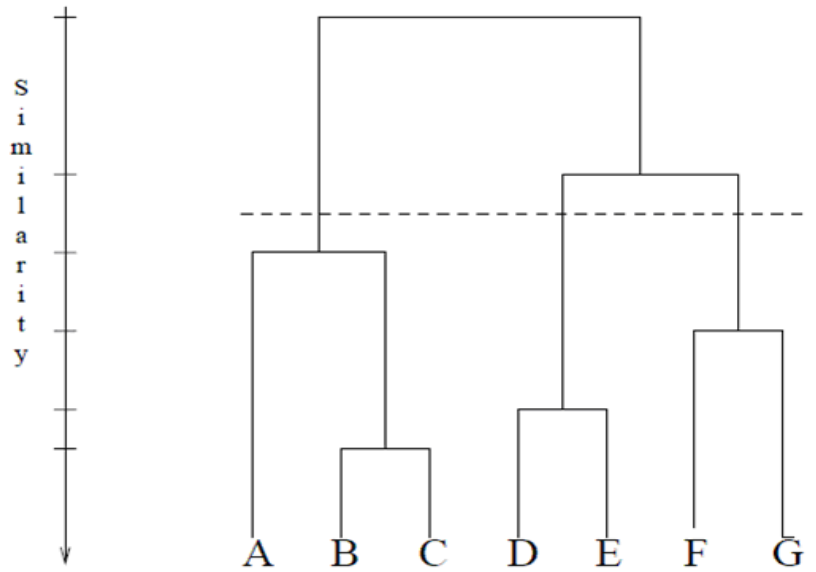
Graph clustering

- Vertexes are adjacent, if levenshtein distance between them is less than threshold
- Each connected component represents similar reads



Hierarchial clustering

- At first, define separate cluster for each read
- On each step two clusters with minimal distance are merged
- Repeat until minimal distance is less than threshold



Step 3: Sequence clusterization

- Graph clusterization (d = 4)
- Hierarchial clusterization (cutoff = 4)
- MSA of clusters
- Consensus through MSA

Results

Chain	VH	VK	VL
Initial diversity	105 502		
After step 1	26 495	20 899	21 994
After step 2	5 306	4 114	6 426
After step 3	15 733	13 399	13 889

Llama says:

Thanks for attention!