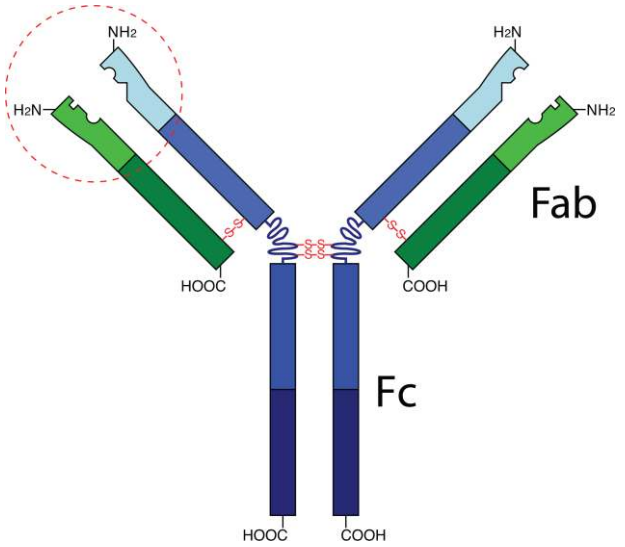
HIERARCHICAL CLUSTERING OF NGS DATA

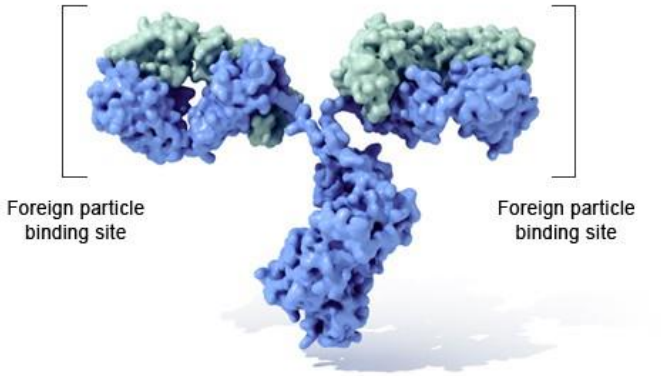
Oleg Yasnev

Advisor: Pavel Yakovlev

Subject



Immunoglobulin G (IgG)



U.S. National Library of Medicine

- Immunoglobulin – part of our immune system
- Fab fragments are hypervariable
- Hard to distinguish sequencing errors from a real variability

Task

Input

- Many (~30k) short and very resembling reads
- Sequencing hypervariable regions with Roche-454

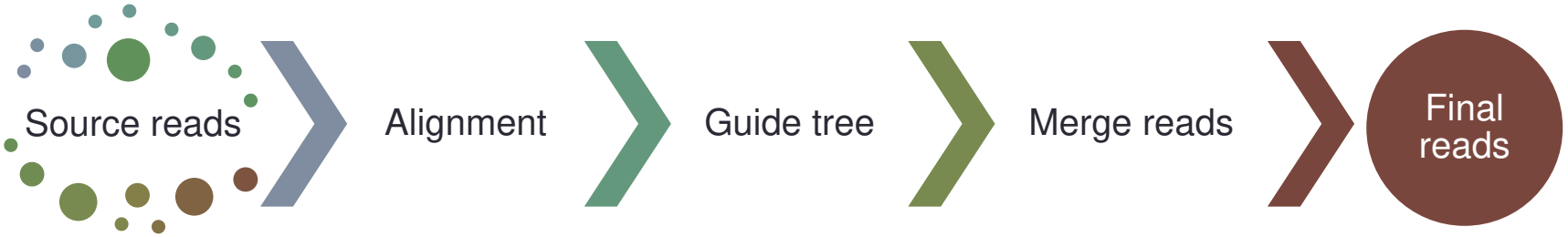
Output

- Error correction
- Hierarchical clustering, such that one leaf – one unique immunoglobulin

How to solve

1. Multiple pairwise alignment with errors in homopolymers
2. Hierarchical clustering
3. Merge resembling reads
4. Reclustering

Pipeline


Source reads
Alignment
Guide tree
Merge reads
Final reads
Homopolymer Alignment tool
Clustal Omega
• Merge Tree tool
• Clustal Omega

Homopolymer Alignment

- Difference between homopolymers in 1—2 nucleotides is not such bad
- But the more, the worse

$$Hscore = score \cdot max - \frac{|score|}{(max - min)^2},$$

score – usual score from NUC4.4

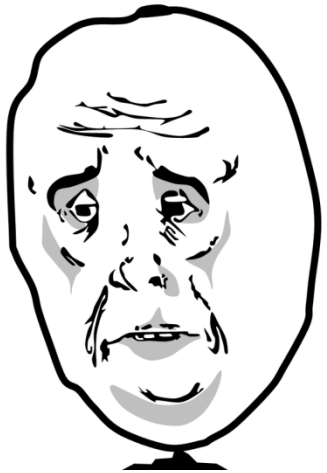
max, ***min*** – maximal and minimal number of nucleotides in homopolymers

ClustalO's little surprise

```
>clustalo.exe -i VL.fasta --distmat-in
score_matrix.dat
```

FATAL: FIXME: reading of distance matrix from file not implemented

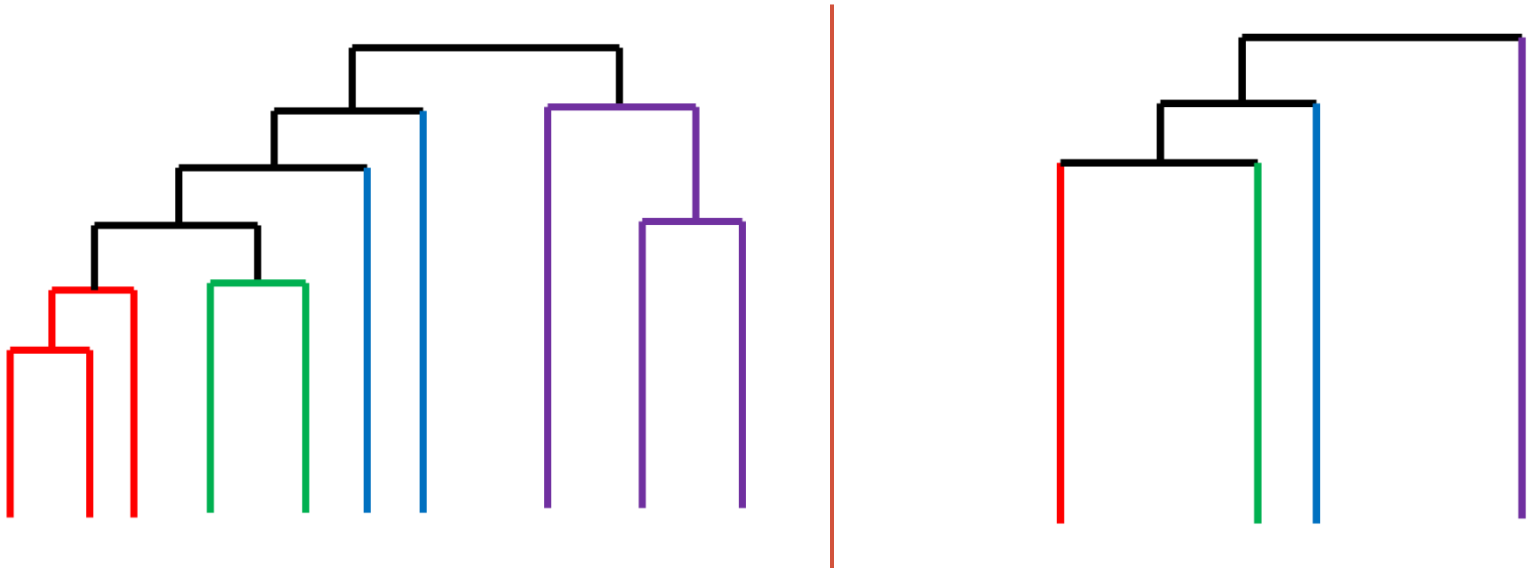
Okay, I'll fix…



Done!

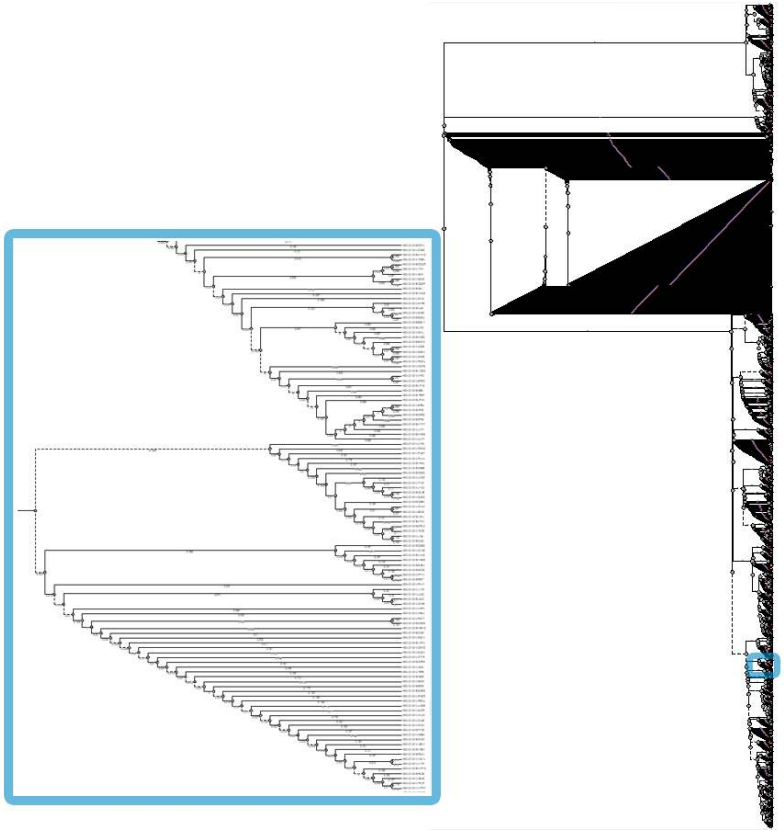
Merge tree

- Tree can be **merged** if:
  - it is a leaf
  - one subtree is a leaf and second can be merged
- Tree can be **partially merged** if:
  - one subtree can be merged



Results

Before



After