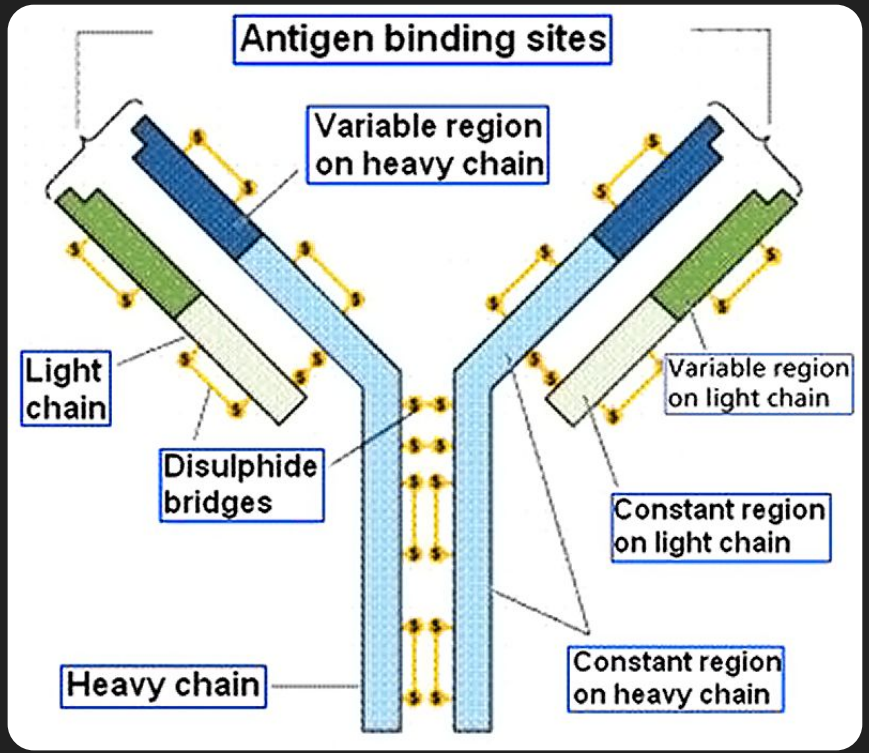
Antibody repertoire construction from Ion Torrent reads

Supervisor:	**Student:**
Bankevich Sergey	Chukharev Konstantin
Center for Algorithmic Biotechnology	ITMO University

Antibody structure

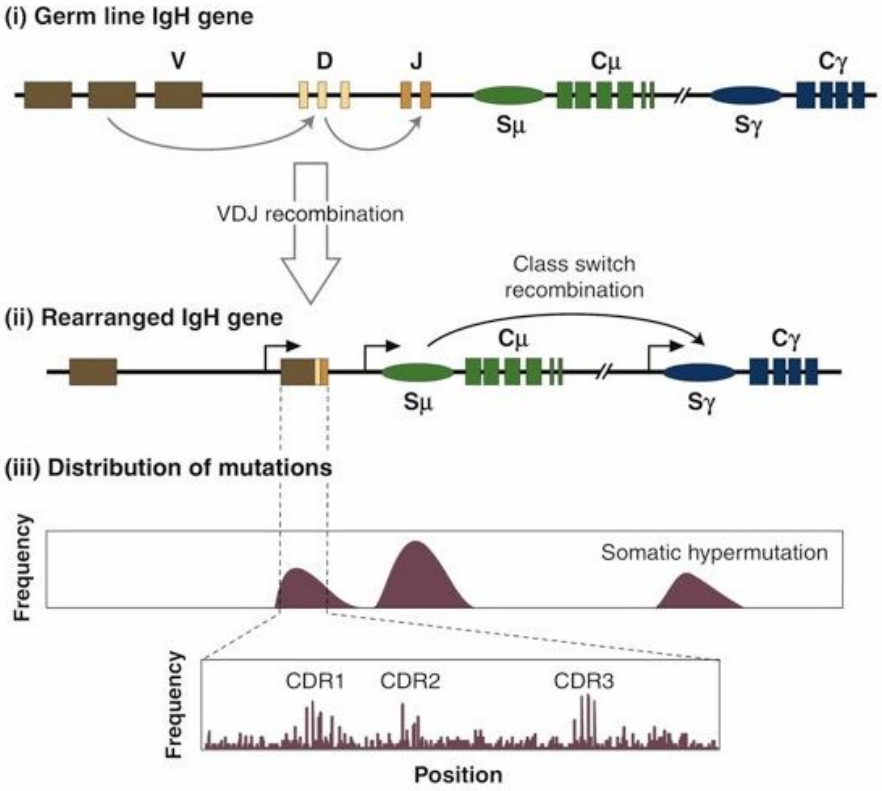
- ***Constant*** region
- ***Variable*** region
  - 110-130 aa
  - *hypervariable* region
  - *framework* region
- Bound by disulphide bridges



VDJ recombination and somatic hypermutation

From \ To		Pyr T	Pyr C	Pur G	Pur A	
Pyr	T		7	4	2	13
	C	16		2	4	22
Pur	G	7	6		15	28
	A	3	12	22		37

Di Noia JM and Neuberger MS.
Annu Rev Biochem. 76:1 (2007)



Problems

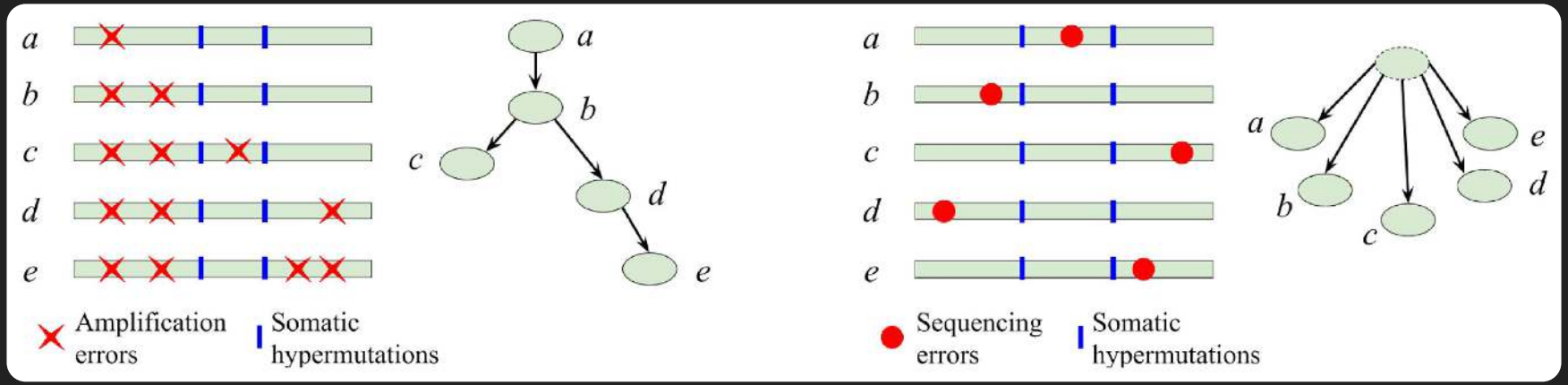
3 different *clustering problems* with increasing granularity:

- ***VDJ*** classification
  - 255x30x15 clusters possible
- ***CDR3*** classification
  - *CDR3 region* is the most biologically important segment
- ***Full length*** antibody repertoire classification
  - *Somatic hypermutations* (SHMs) accounted

What can go wrong?

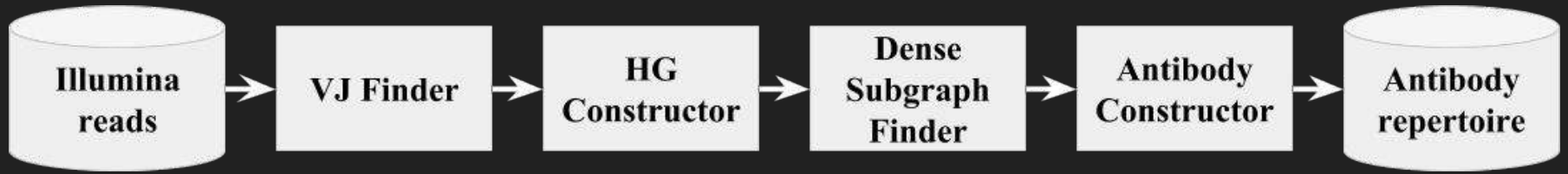
- *Amplification* errors

  ⇒ pseudo-diversity of an antibody repertoire
- *Sequencing* errors



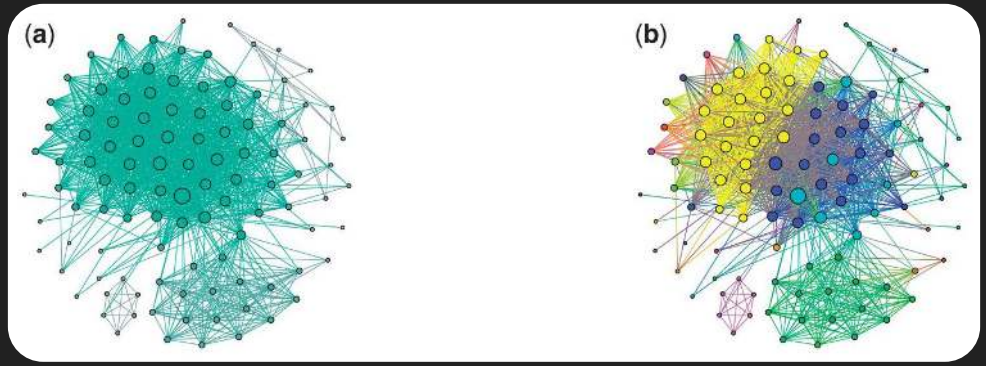
IgReC pipeline

- Align reads against germline genes (*V(D)J labeling problem*)
  - Fast *VJ Finder* tool
- Construct *Hamming graph*
- Analyze *connected components* (*almost-cliques*)
- Each CC — single or some similar antibodies



Hamming graph

- `HG(Reads,`τ`)` is a subgraph of Hamming graph, where edge (s_1, s_2) exists iff `d(s1,s2)`$\leq\tau$
  - Illumina: `d` = Hamming distance (almost)
  - Ion Torrent: `d` = Edit distance
- Ideally, proper τ makes `HG(Reads,`τ`)` — *clique graph*
- Reality:



Filtration

To build HG we need to calc distances between all reads

Not good.

To reduce number of comparisons use *filtration techniques*

Fact: “two strings of length L differing in at most τ positions must share at least one l-mer of length $L/(\tau+1)$” (Knuth, 1998)

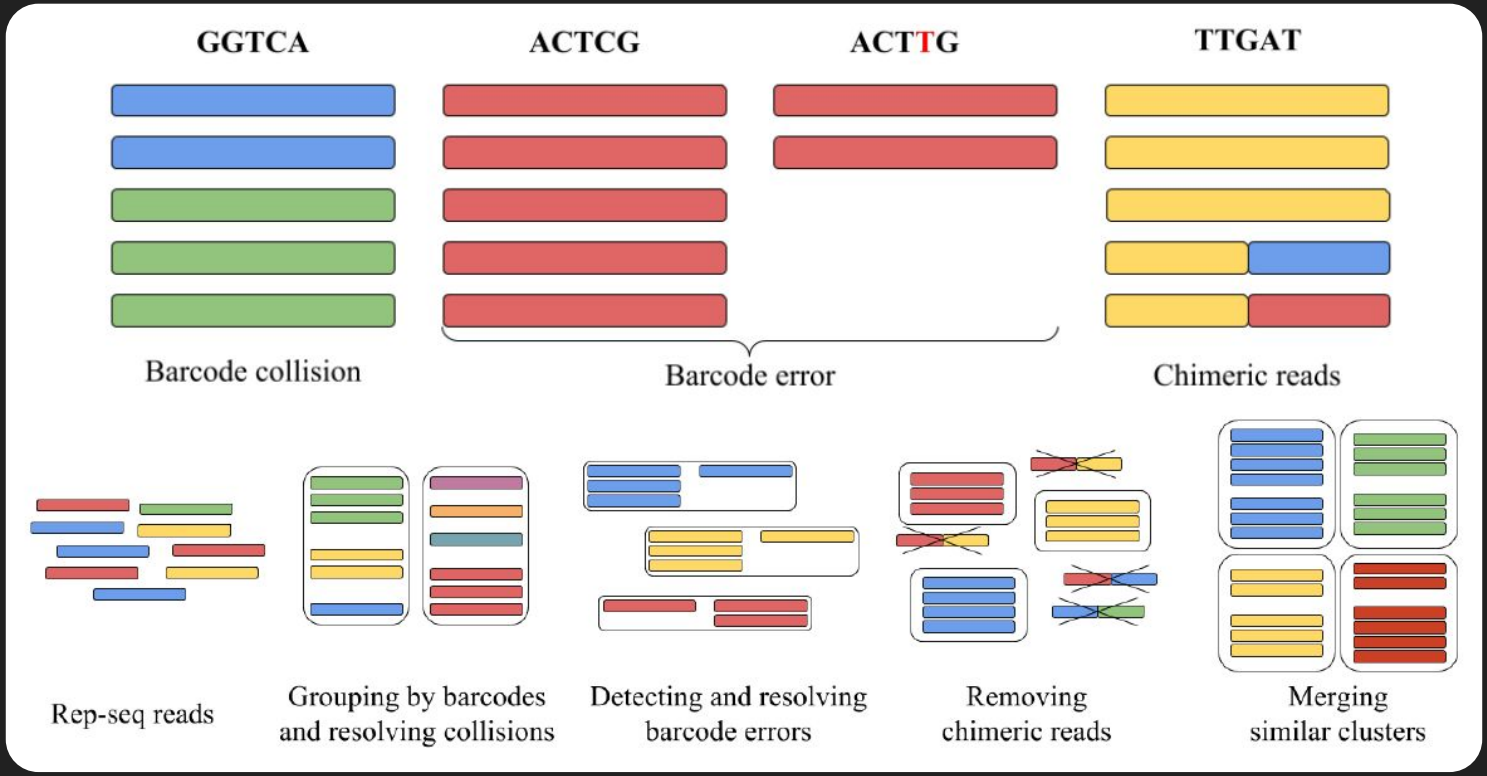
Filtration ⟶ verification ⟶ distances!

Minimizers

Observation about two strings of length L differing in at most τ positions: “given a set of τ+1 non-overlapping k-mers in the first string, at least one of them appears in the second string”

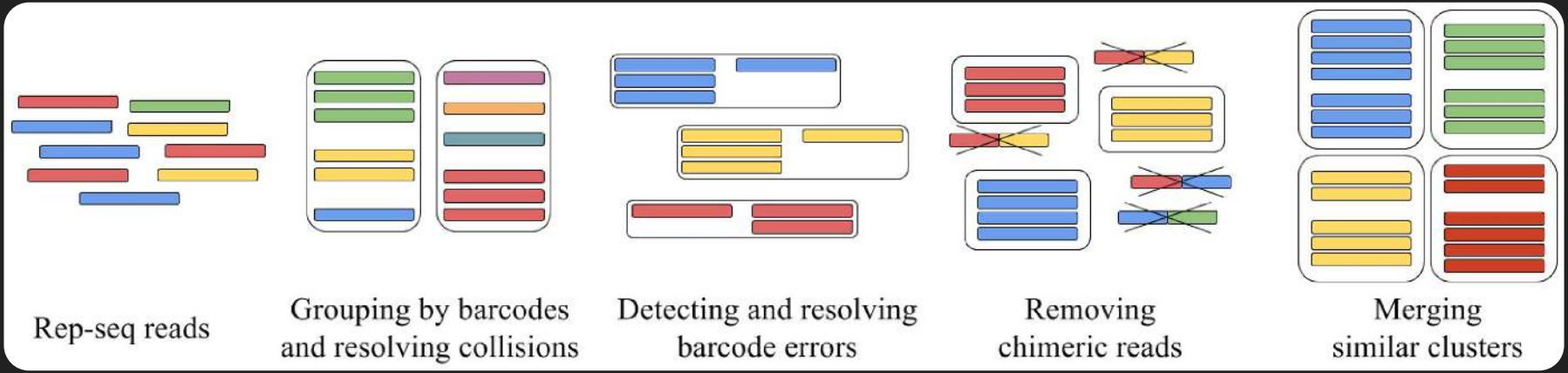
Minimizers — τ+1 rarest and non-overlapping k-mers

Filtration ⟶ *verification* ⟶ *distances!*



(top) barcoding/amplification errors

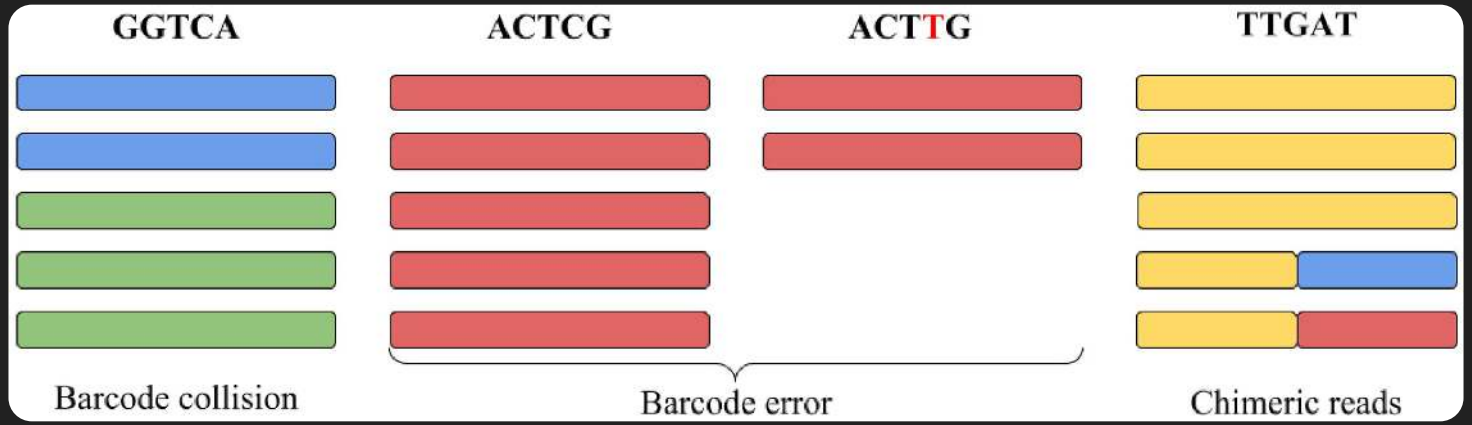
(bot) `barcodedIgReC` pipeline



barcodedIgReC pipeline

Barcoding and amplification artifacts

- Amplification and sequencing errors
- Barcodes collision
- Chimeric reads

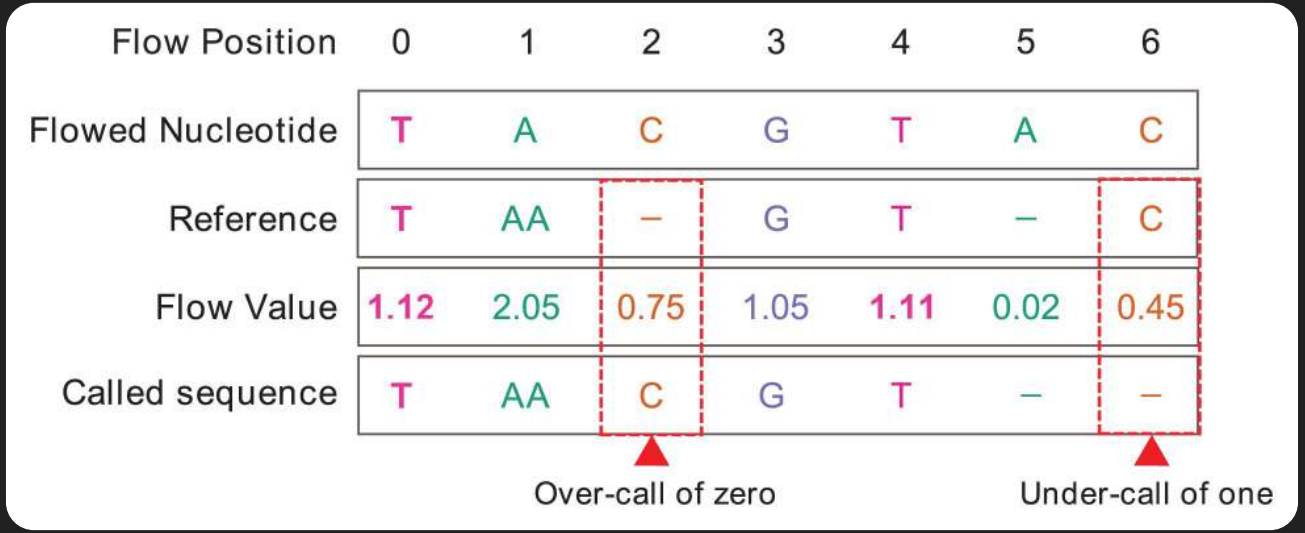


Characterising Ion Torrent errors

- Indels
- Homopolymer errors
- Substitution errors

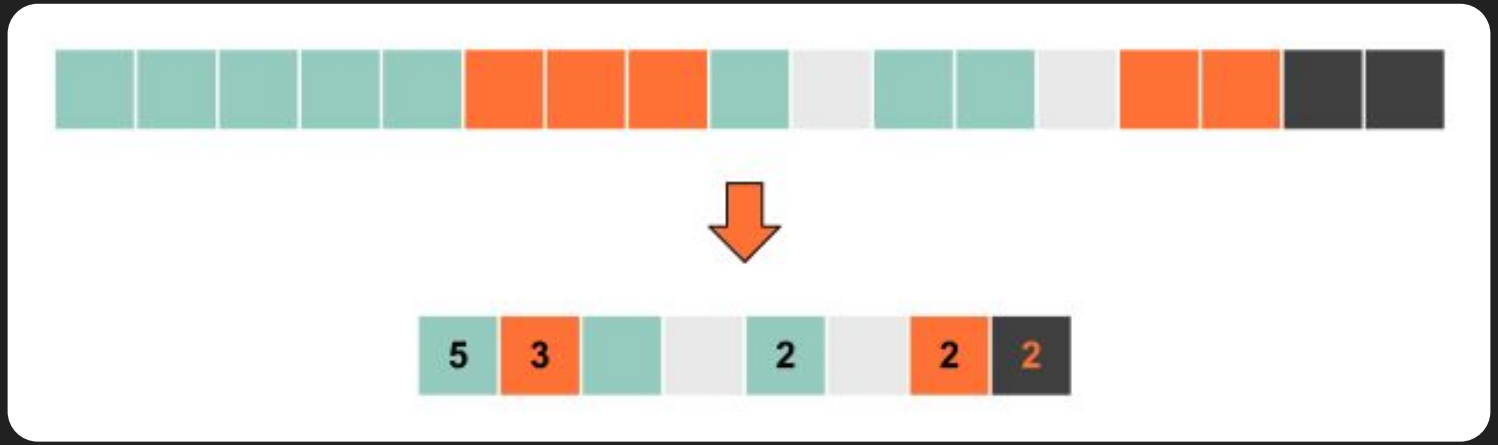
Homopolymer errors

- The term came from Roche 454 pyrosequencing
- Indels are subclass of homopolymer errors



RLE (Run-Length Encoding)

- To somehow analyze data with homopolymer errors, just shrink consecutive nucleotides into one



Run-Length Decoding

- We only know how to decompress initial reads and germline
- Possible decoding techniques:
  - search *k*-mers from consensus in reads/germline
  - align reads to consensus
  - ...

Results

- Initial data:
  - 4 datasets with Ion Torrent reads
- IgReC output on encoded reads with encoded germline:
  - “encoded” antibody repertoire — compressed consensuses
- Postprocessing:
  - 4 “decoded” antibody repertoires

Thanks for attention!

BIOINFORMATICS
INSTITUTE