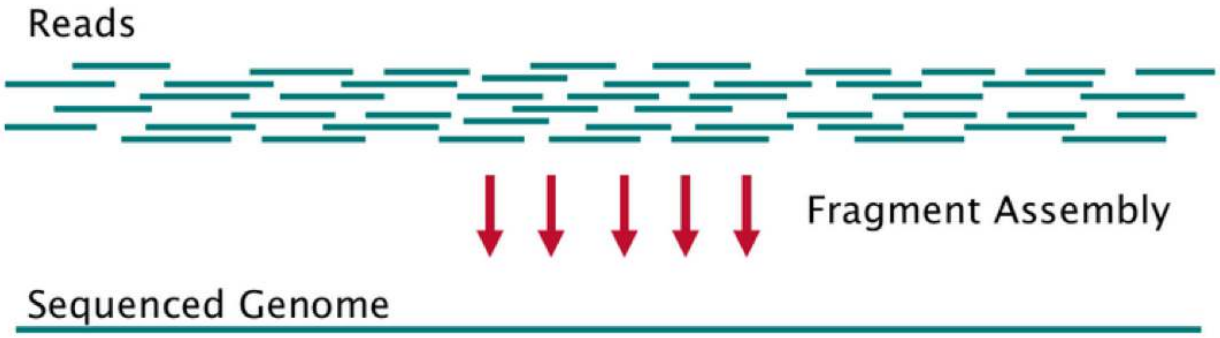
De Novo Assembly

Student:
Aleksey Aleev

Scientific advisors (BioDatomics, Inc):
Maxim Mikheev
Gennady Ganebnii

NGS data assembly pipeline

- Artifacts & contaminants cleaning
- Draft assembly
  - Error correction
  - Contig assembly
  - Repeat resolution & scaffolding
- Postprocessing
- Finishing
- Annotation



Why do we need new assembler?

- Existing solutions are still not efficient enough in terms of memory consumption and execution time
- Many assembly steps could be done using distributed computing
- We'd like to implement and evaluate de novo assembly algorithm based on distributed computing using Hadoop platform

Project goals

1. Learn existing approaches for De Bruijn Graph based assembly
2. Design and implement De Bruijn Graph based prototype for NGS data contig assembly step using Hadoop
3. Construct new or use existing assembly pipeline and integrate implemented assembly step to this pipeline
4. Evaluate quality and performance of implemented solution
5. Make further improvements to implemented solution or implement another step of assembly pipeline also using Hadoop

What was already done

- Existing approaches for De Bruijn Graph based assembly are learnt; particularly, approaches from:
  - Velvet
  - ABySS
  - PASHA
  - SOAPdenovo
- Apache Giraph is chosen as a graph processing system
- DBG vertices and edges representations were designed
  - binary representation of k-mers
  - 4 bits to identify if there are edges which extend vertex via A,C,G and T

Thank you!