Metafast

High-throughput tool for metagenome comparison

V. Ulyantsev, S. Kazakov

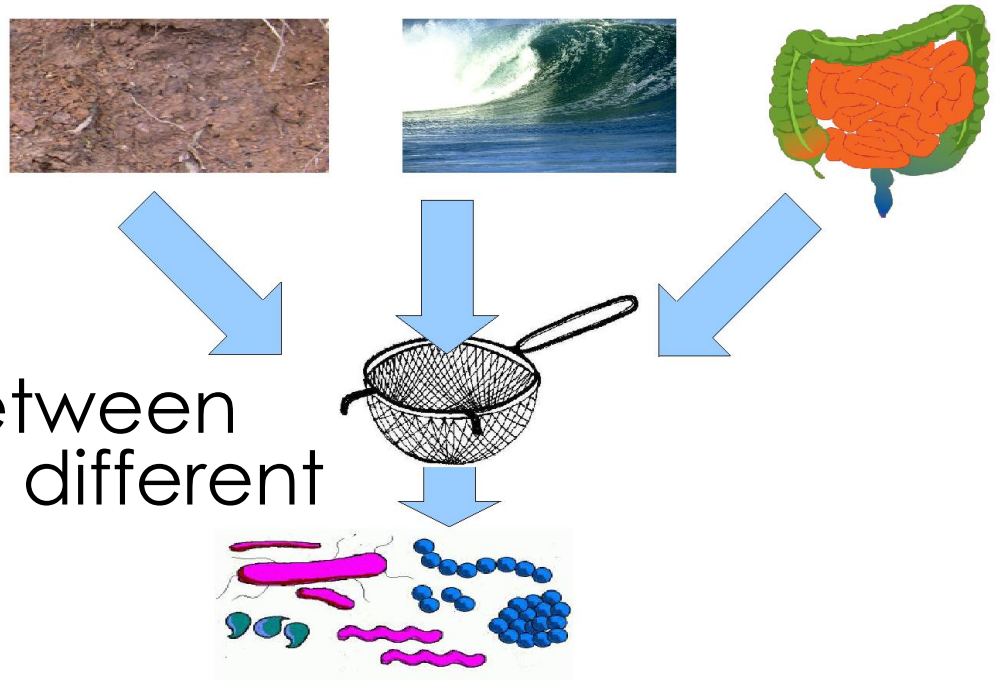
V. Dubinkina, Tyakht A., Alexeev D.

ITMO UNIVERSITY

POST GENOME

ФЕДЕРАЛЬНОЕ МЕДИКО-БИОЛОГИЧЕСКОЕ АГЕНТСТВО
НАУЧНО-ИССЛЕДОВАТЕЛЬСКИЙ ИНСТИТУТ ФИЗИКО-ХИМИЧЕСКОЙ МЕДИЦИНЫ

Comparative metagenomics

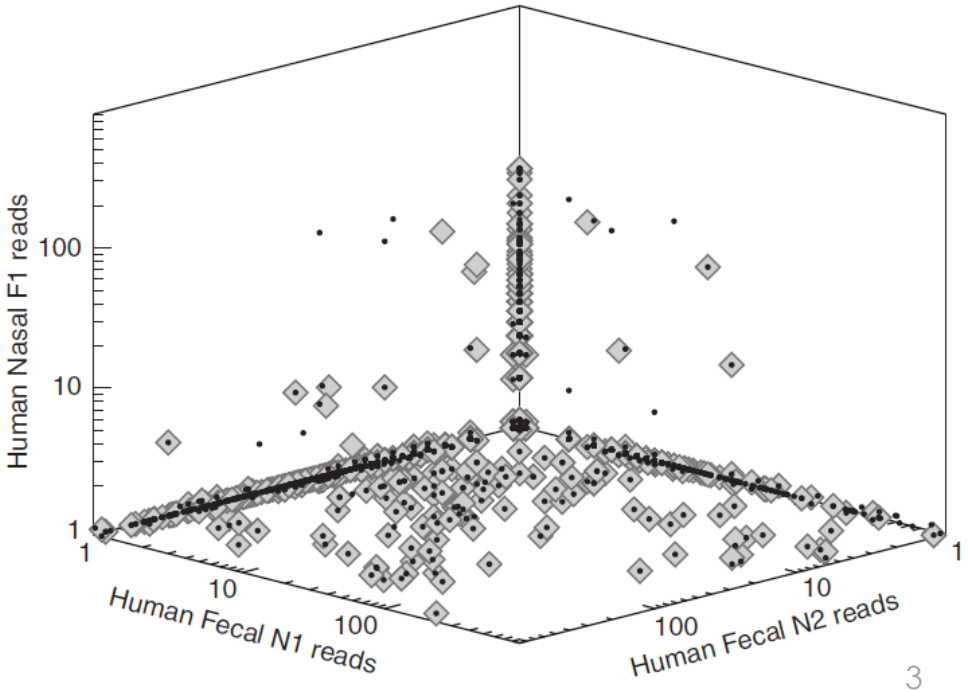


- **Interrelationships** between metagenomes from different samples
  - different biomes
  - different time points
- High level of **unknown** sequences
- Mapping can **limit** the amount of data that can be analyzed

crAss

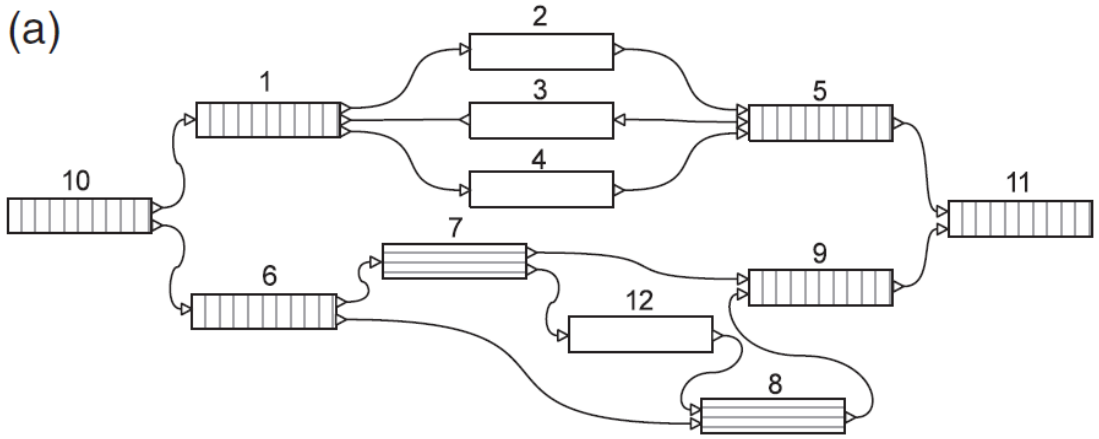
- De novo cross-assembly of all sequence reads
- Number of cross-contigs with reads from both metagenomes

$$d_{i,j} = 1 - \frac{c_{i,j}}{min\{c_i, c_j\}}$$



MaryGold

- Detect and explore genomic variation between metagenomic sequencing samples
- Detect bubble structures in contig graphs using graph decomposition
- 454 and Illumina data



Challenges

- Reference-based (mapping sequences)
  - High level of unknown sequences
- Assembly-based
  - Slow on large datasets

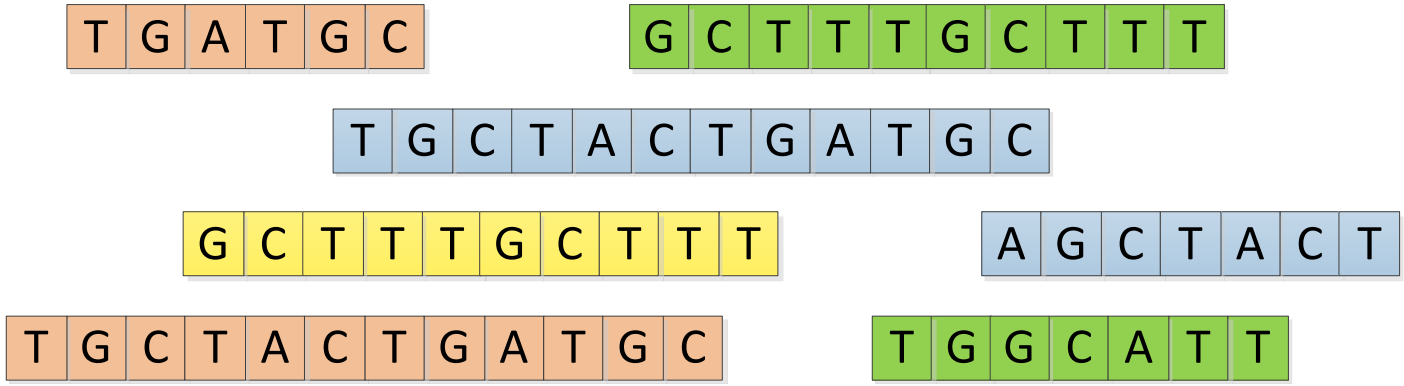
Solution: fast but low quality "semi-assembly" for every library

New algorithm – MetaFast

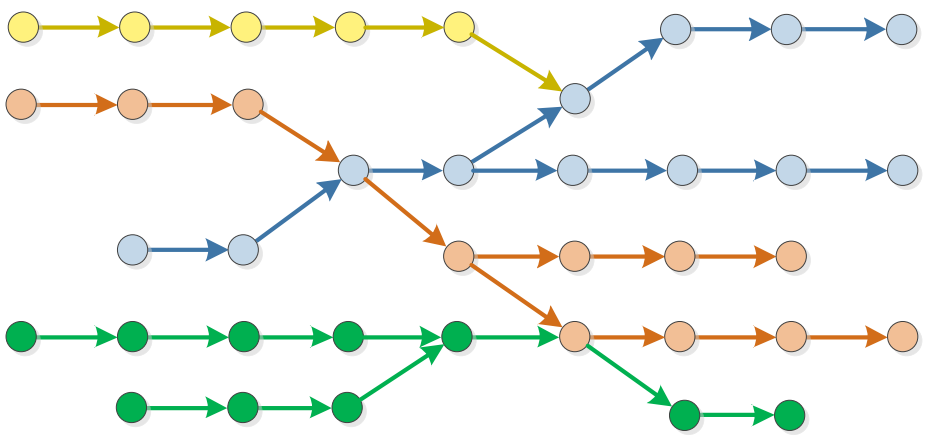
Method, based on simple assembly:

A. For each library:
   - Construct de Bruijn graph
   - Extract simple paths, not contigs
B. For all libraries:
   - Construct de Bruijn graph from found paths
   - Extract components
C. Calculate characteristic vectors for libraries.

1. Construct de Bruijn graph

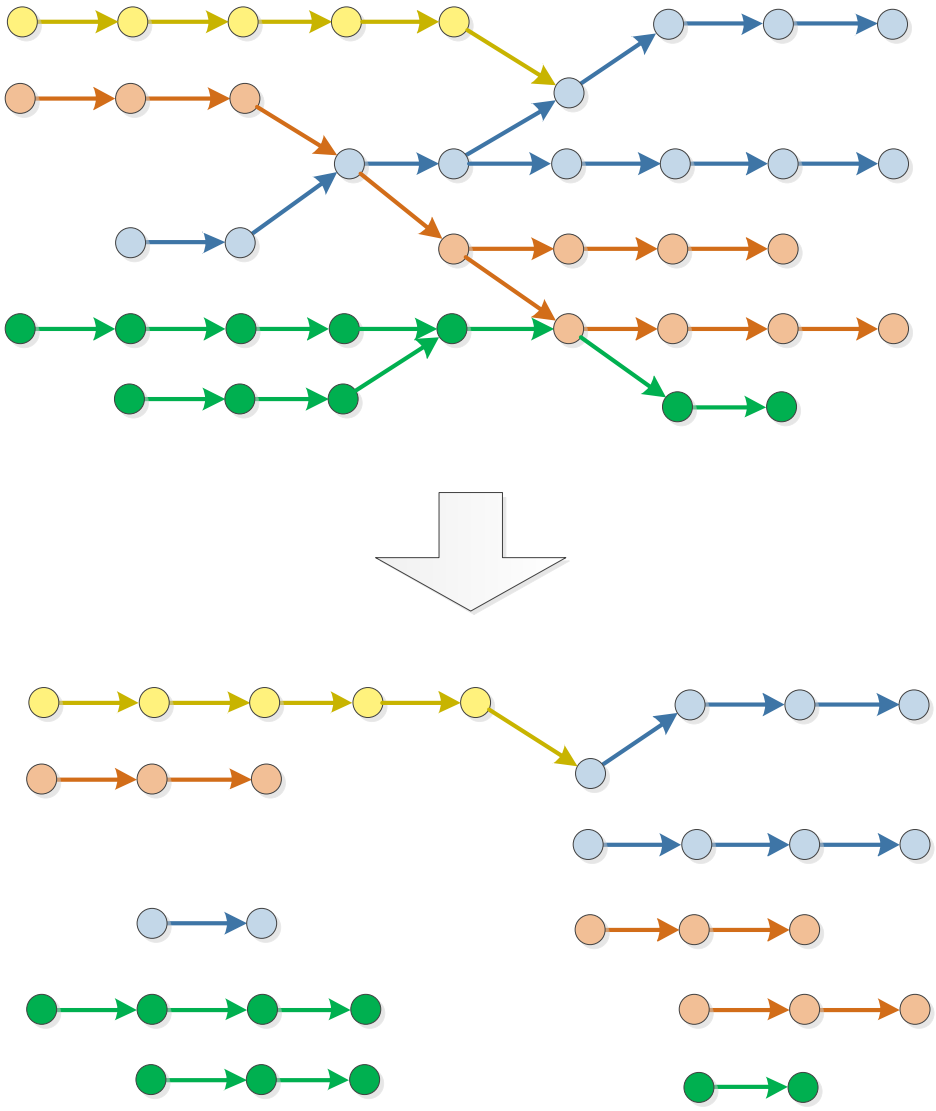


Library



de Bruijn graph

2. Extract simple paths

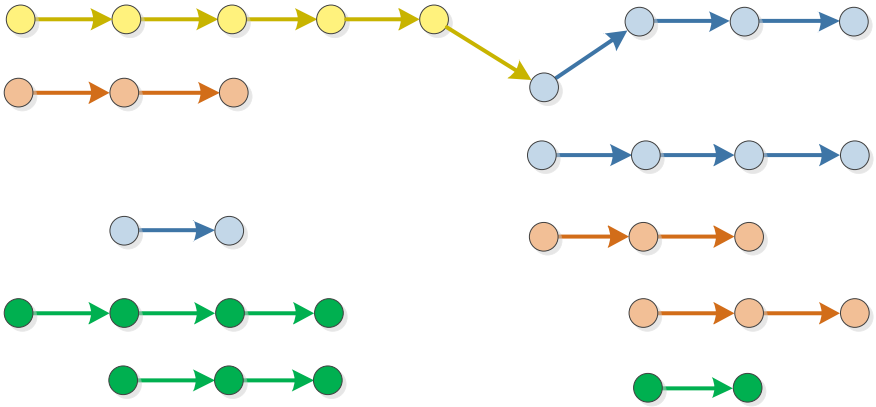


de Bruijn graph

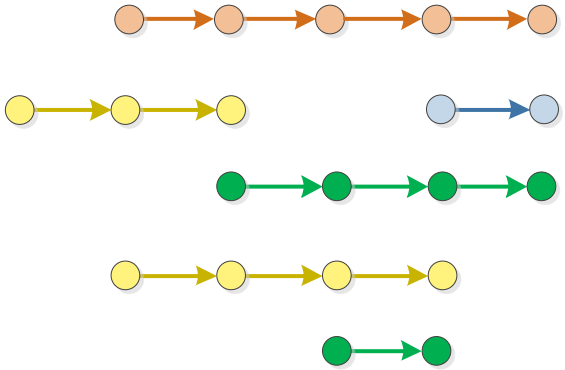
Simple paths

3. Merge paths

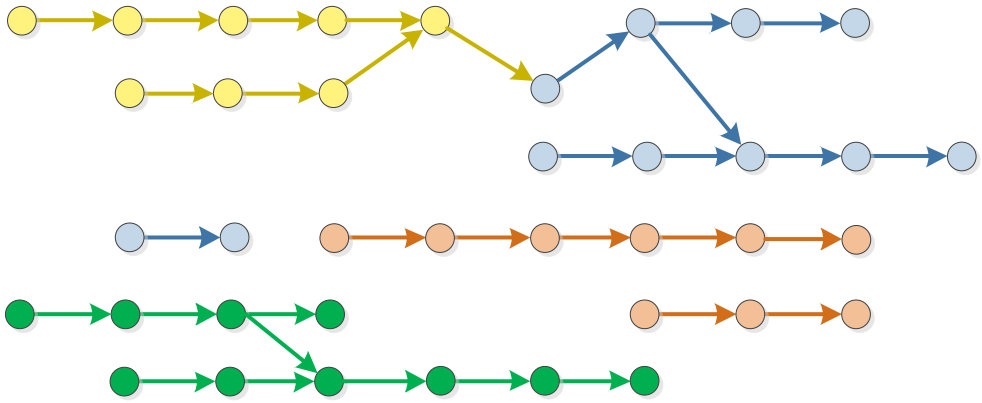
Paths 1



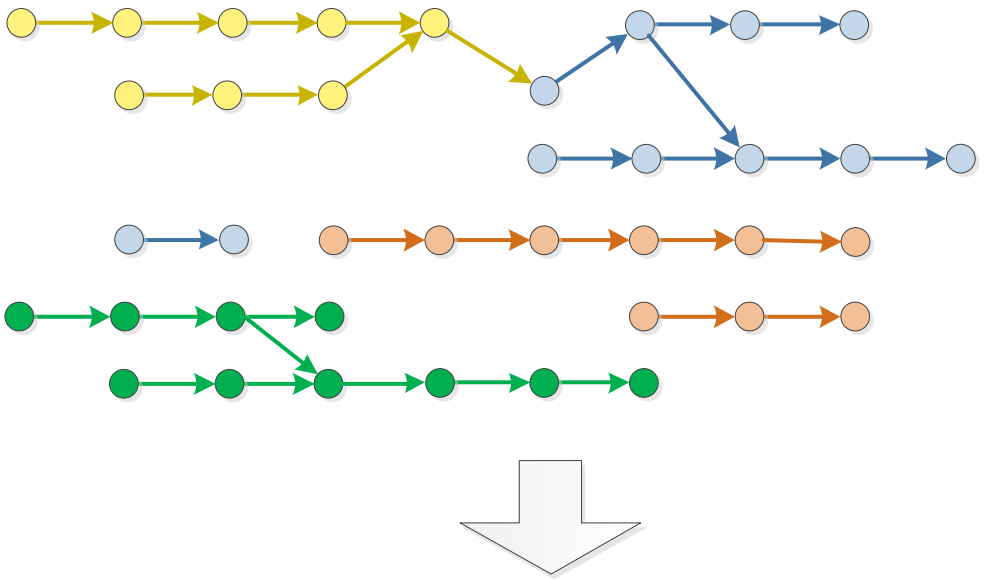
Paths 2



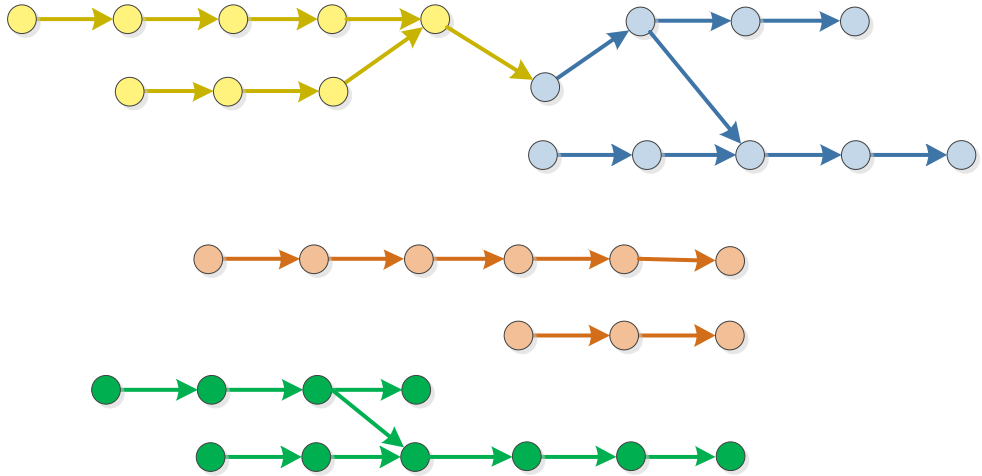
Paths combined in single de Bruijn graph



4. Extract components



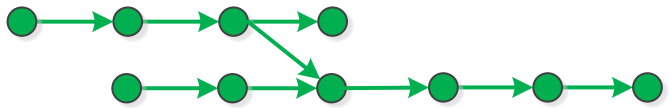
Paths
de Bruijn graph



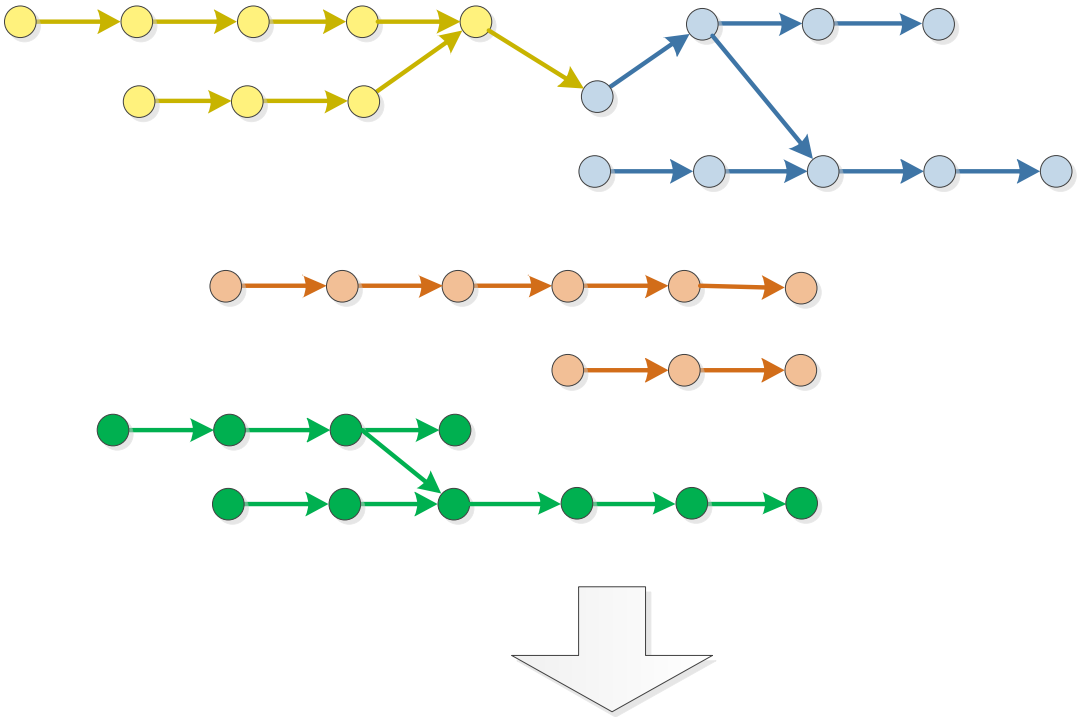
Found
components

Component

- K-mers set
- B_1 <= size <= B_2
- Big component?
  - Iterative algorithm for decomposition



5. Construct characteristic vectors



Components

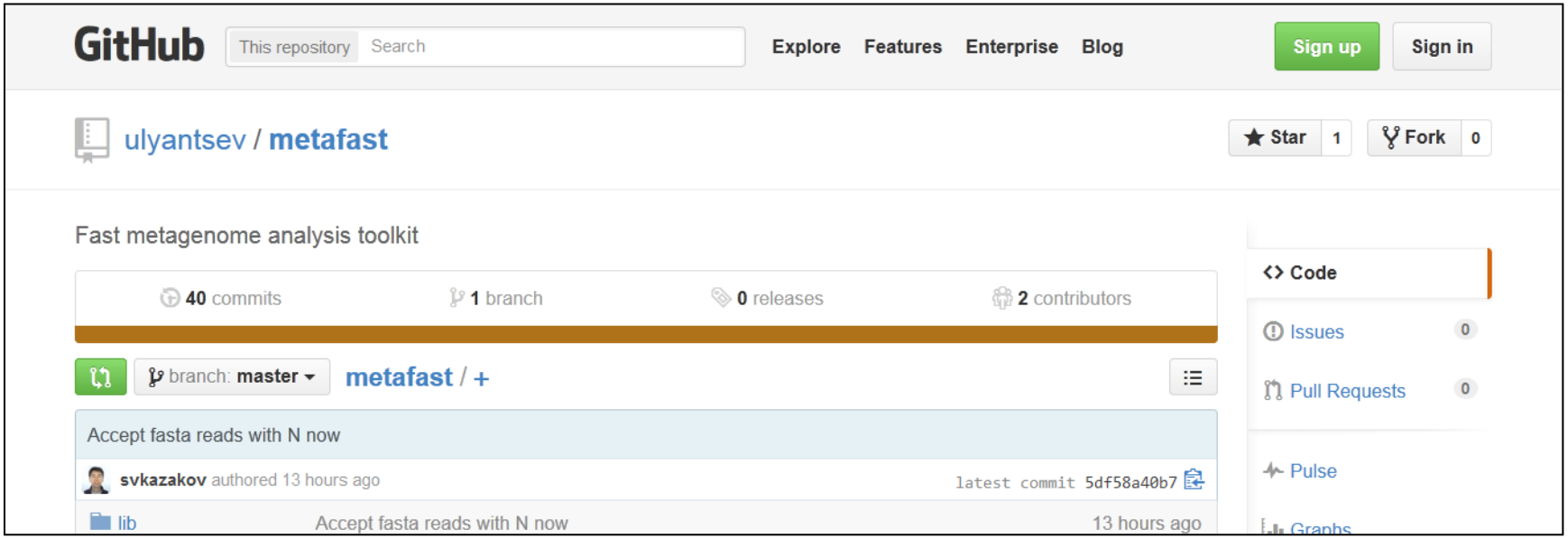
Library 1 **(15, 0, 6)**

Library 2 **(0, 7, 8)**

Characteristic vectors

Implementation

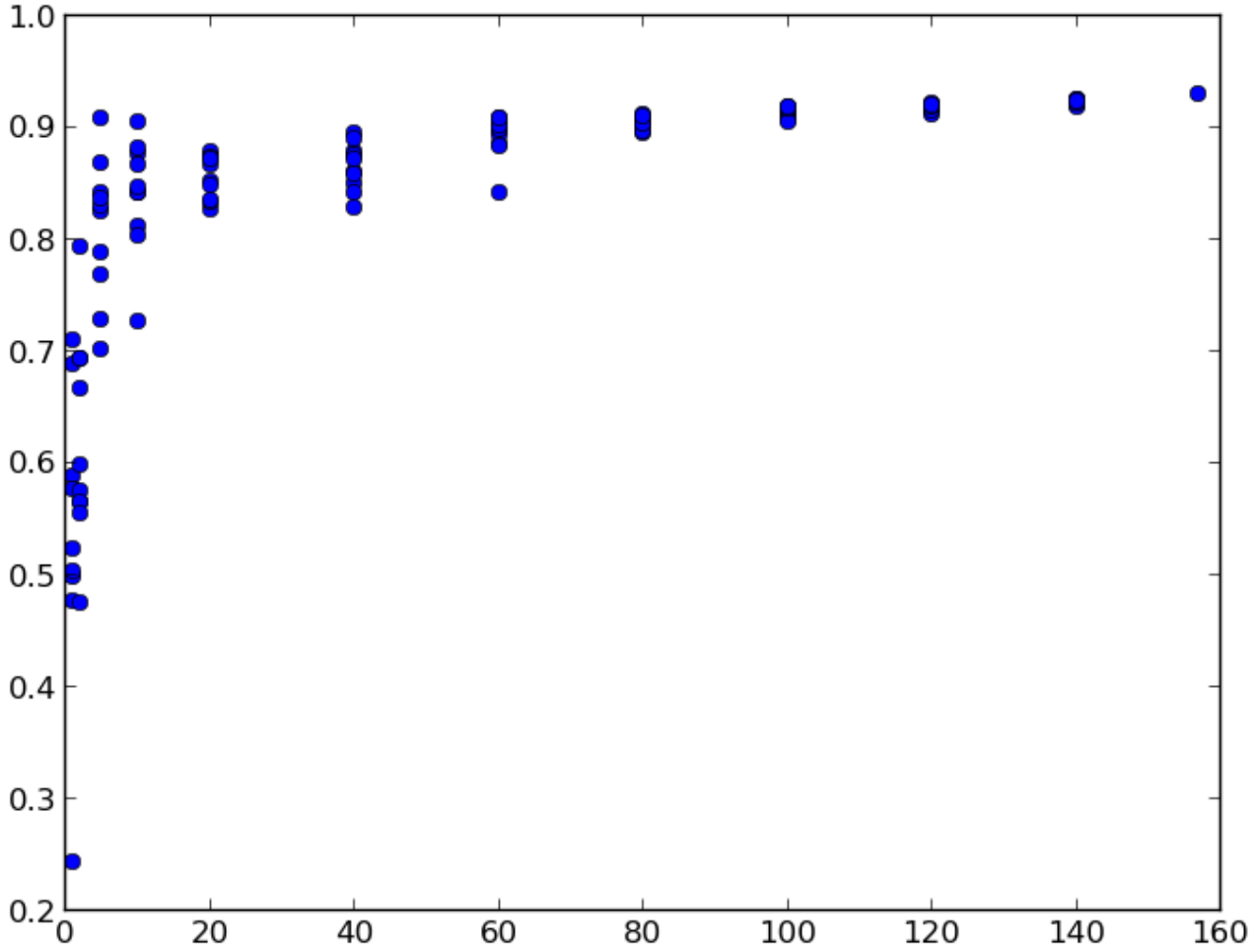
- On *Java*
- Open source project
- http://github.com/ulyantsev/metafast



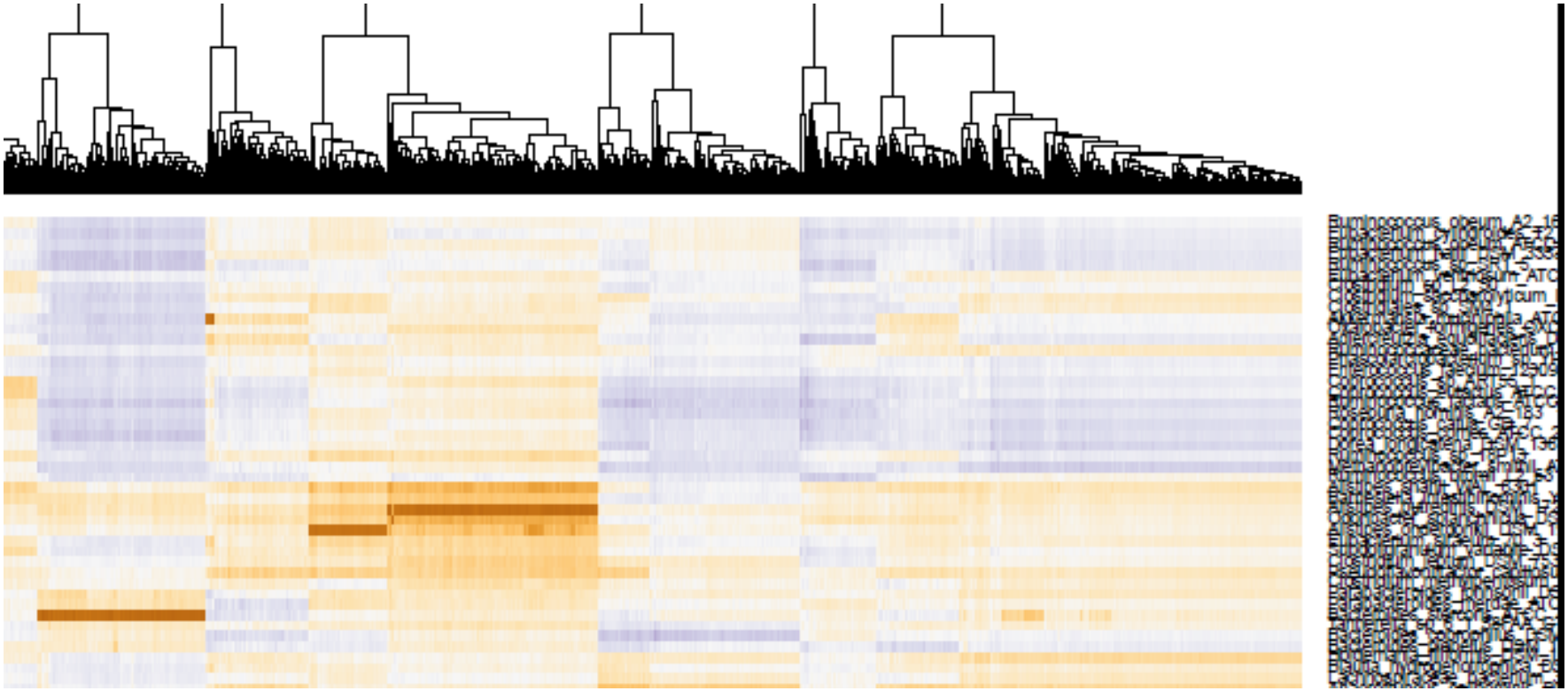
Experiments

- **157** Chinese gut metagenomes (600 Gb)
- **93 %** – correlation between distance matrices based on mapping to knows references and our vectors
- About 10 hours cluster time (not full-loaded)

80 libraries is enough



Components-genes correlation



Results & future work

- MetaFast – new approach and cross-platform tool for comparative metagenomics
- Promising initial experiments

- Experiments with simulated data
- New information about existing metagenomes
- Algorithm modifications

Thank you for attention!

http://github.com/ulyantsev/metafast

V. Ulyantsev

S. Kazakov

V. Dubinkina

A. Tyakht

D. Alexeev

ITMO UNIVERSITY