



УНИВЕРСИТЕТ ИТМО

# Разработка методов построения и визуализации геномного контекста в метагеномных данных с использованием Hi-C связей на графах де Брейна

Шостина А.Д.,  
Научный руководитель – Ульяновцев В.И.

## Полногеномное секвенирование

Полногеномное секвенирование (WGS) — это лабораторная процедура, определяющая порядок нуклеотидных оснований (А, Т, С и G) в геноме организма.

Пример: ACGTCCCGTCAA,  $k = 3$

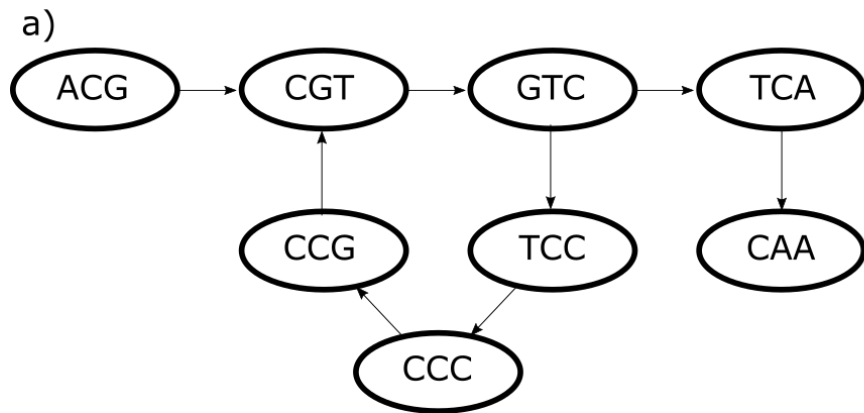


Рисунок 1. Граф де Брейна

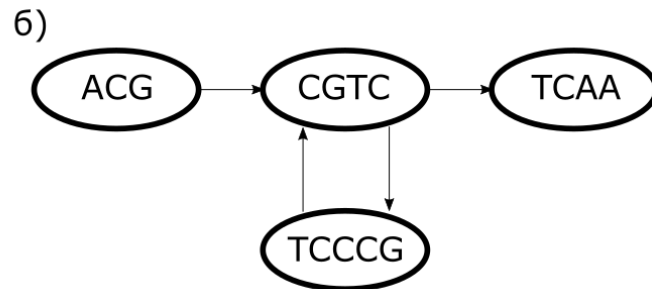
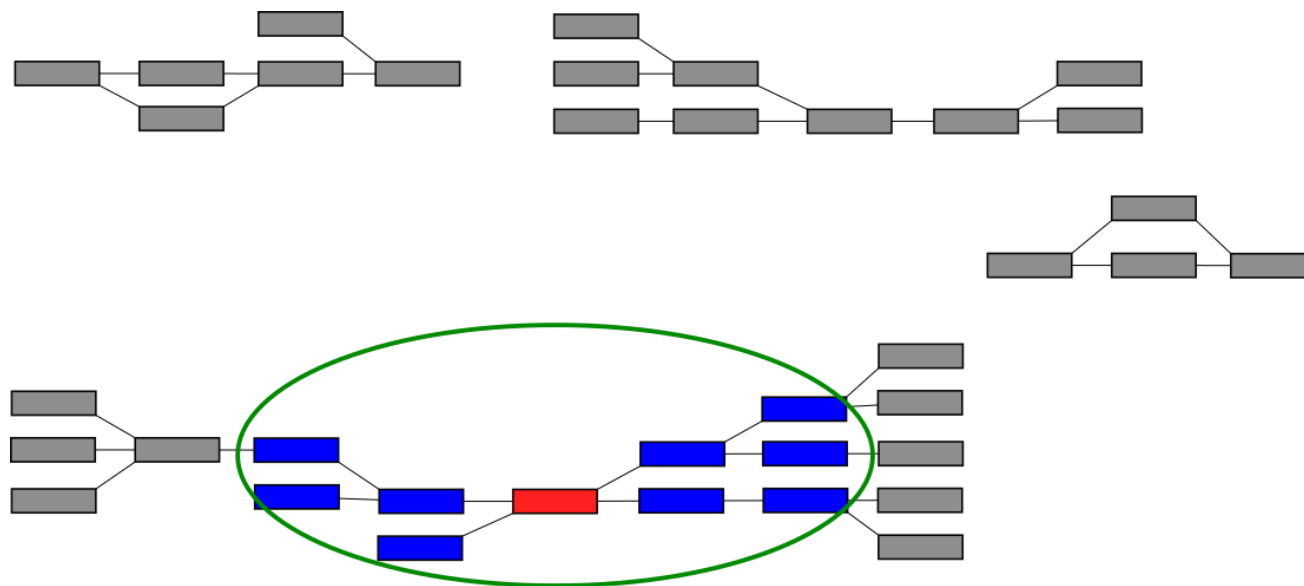


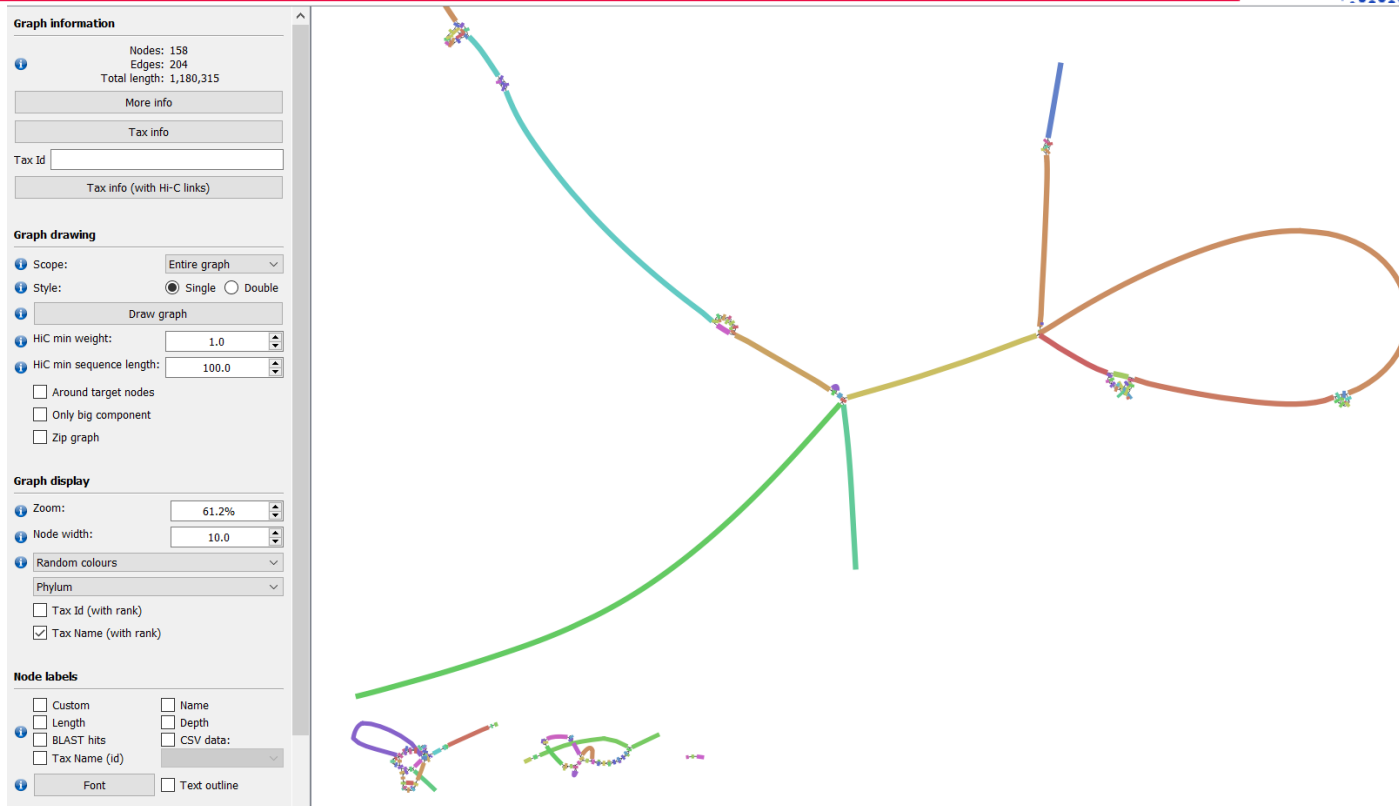
Рисунок 2. Сжатый граф де Брейна

## Геномный контекст

Геномный контекст – это подграф графа де Брейна, построенный вокруг анализируемого гена. Построить геномный контекст можно при помощи приложения **MetaCherchant**.

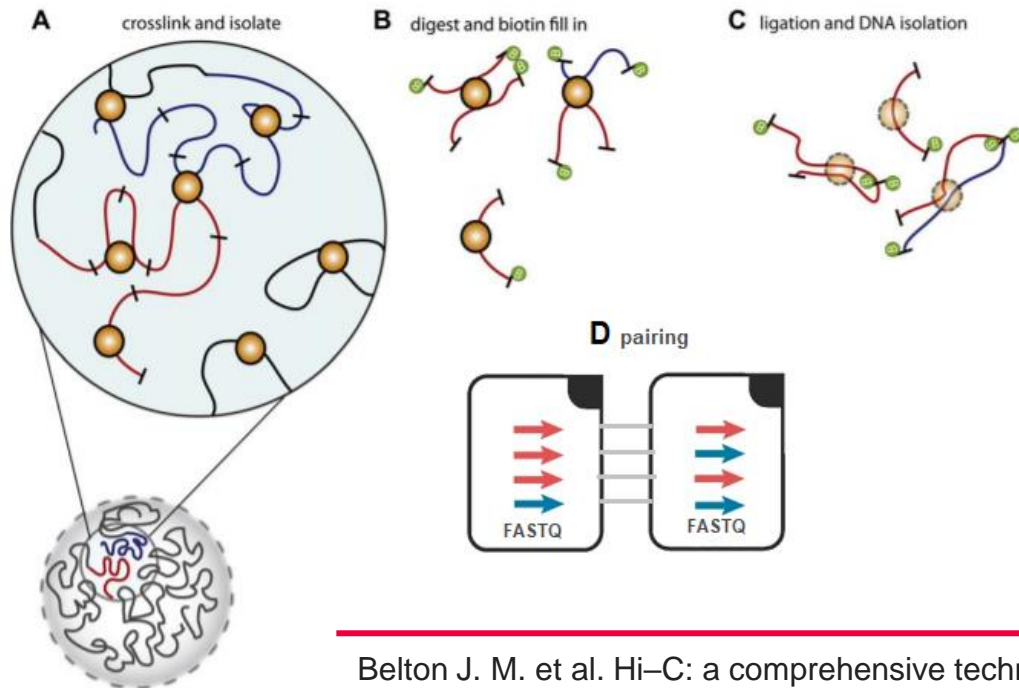


## Визуализация графа де Брейна в приложении Bandage



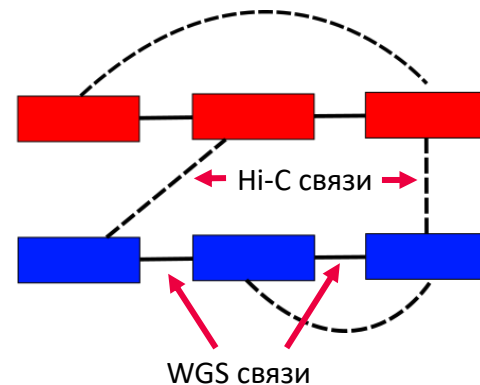
## Определение конформации хромосом (Hi-C)

**Hi-C секвенирование** - один из методов изучения пространственной организации генома.



Хромосомная  
ДНК Бактерии

ДНК Плазмиды



Belton J. M. et al. Hi-C: a comprehensive technique to capture the conformation of genomes //Methods. – 2012. – Т. 58. – №. 3. – С. 268-276.

## Биологический вызов

- Антибиотико-резистентные гены могут располагаться в плазмидах.
- Бактерия приобретает свойство устойчивости к антибиотикам, если в ее клетках есть плазмида с АРГ.
- Хромосомная ДНК бактерии и геном плазмиды имеют Hi-C связи друг с другом, если лежат в одной клетке.

Биологическая задача:

**Получение информации о взаимосвязях мобильных элементов (например, плазмид) и их носителей.**



## Цели и задачи

---

### Цель:

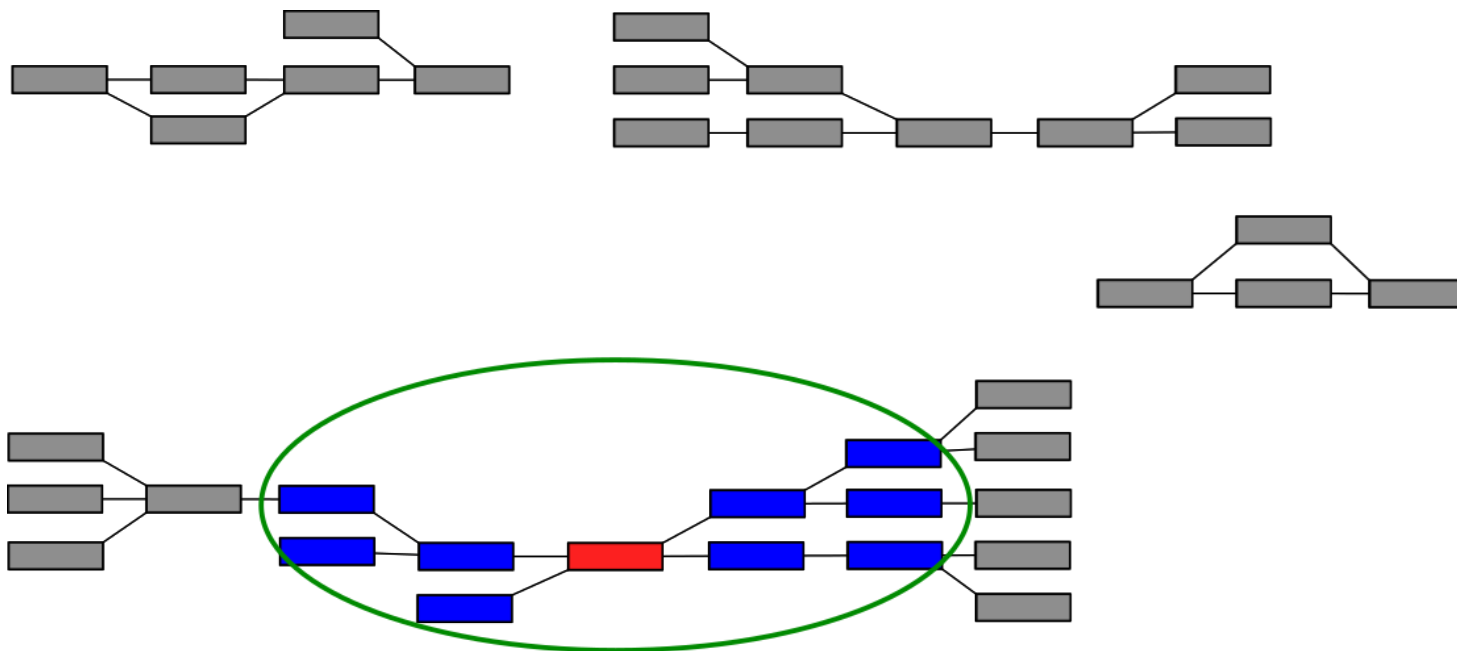
Совершенствование методов построения и визуализации геномного контекста.

### Задача:

- 1) Разработка метода построения геномного контекста вокруг анализируемых генов с учетом Hi-C связей на основе приложения MetaCherchant.
- 2) Использование данных Hi-C секвенирования и результатов таксономического анализа при визуализации графа де Брейна в приложении Bandage.

## Этап 1: Построение исходного геномного контекста

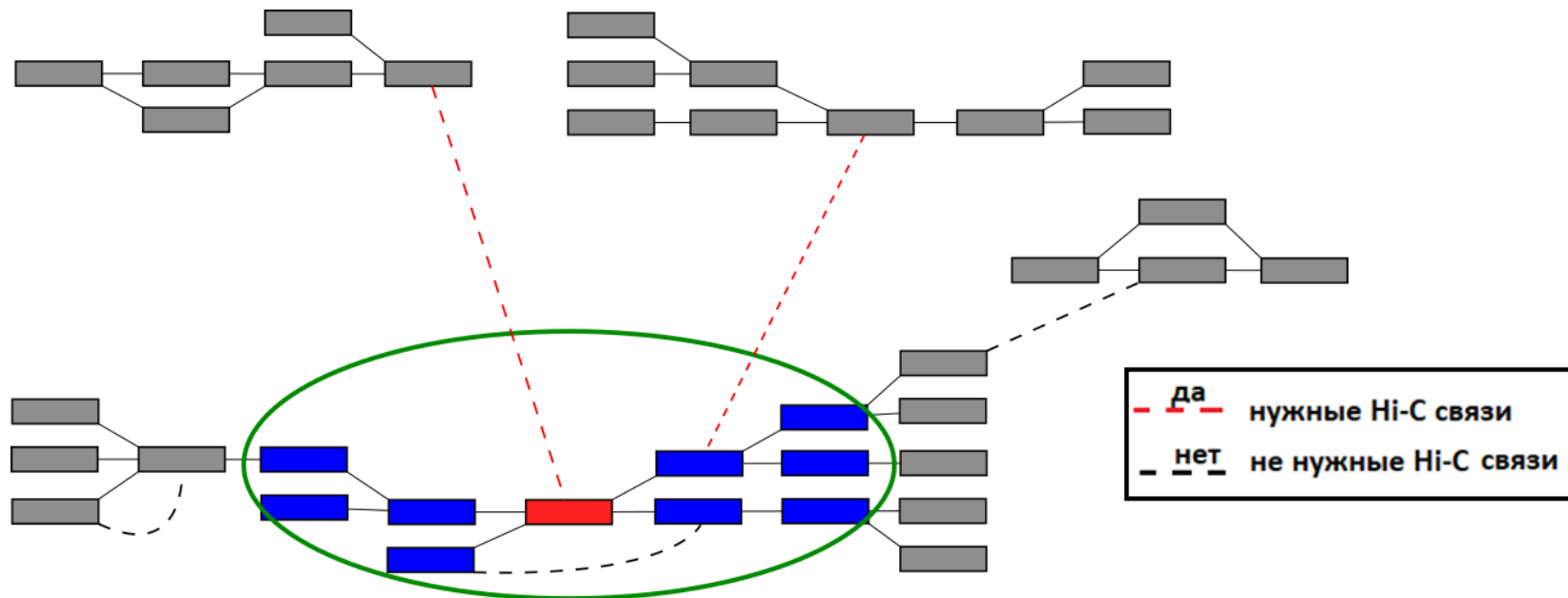
Построение подграфа графа де Брейна при помощи приложения **MetaCherchant**.





## Поиск Hi-C связей, расширяющих геномный контекст

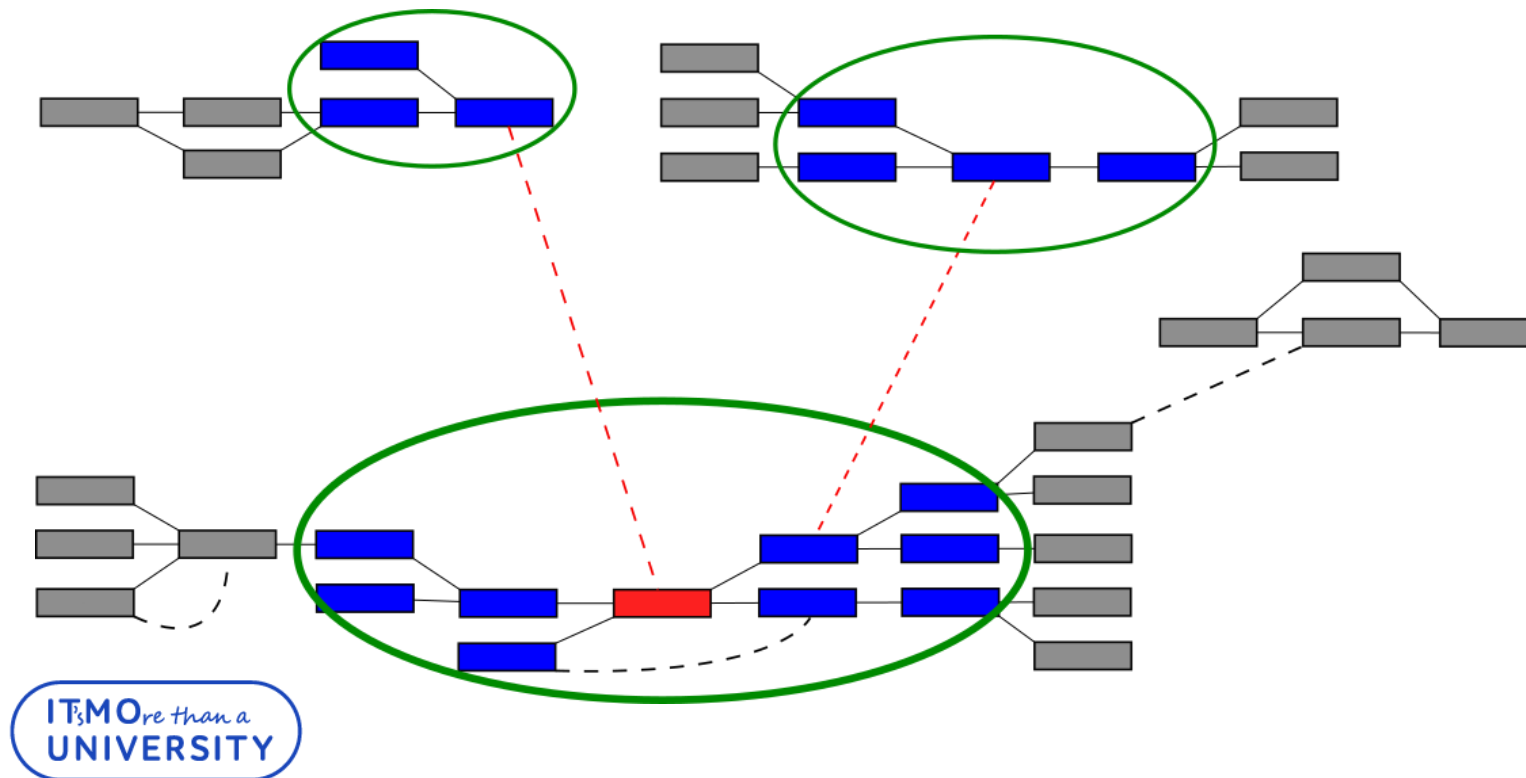
Поиск Hi-C ридов при помощи утилит  
**BWA** [1] и **SAMTools** [2].



1. Li H., Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform //bioinformatics. – 2009. – Т. 25. – №. 14. – С. 1754-1760.
2. Danecek P. et al. Twelve years of SAMtools and BCFtools //Gigascience. – 2021. – Т. 10. – №. 2. – С. giab008.

## Построение геномного контекста с учетом Hi-C связей

Построение расширенного геномного контекста при помощи приложения **MetaCherchant**.



# Визуализация графа в приложении Bandage до его модификации

**Graph information**

Nodes: 158  
Edges: 204  
Total length: 1,180,315

More info

Tax info

Tax Id

Tax info (with Hi-C links)

**Graph drawing**

Scope: Entire graph

Style: ☒ Single ☐ Double

Draw graph

HiC min weight: 1.0

HiC min sequence length: 100.0

☐ Around target nodes

☐ Only big component

☐ Zip graph

**Graph display**

Zoom: 61.2%

Node width: 10.0

Random colours

Phylum

☐ Tax Id (with rank)

☒ Tax Name (with rank)

**Node labels**

☐ Custom

☐ Name

☐ Length

☐ Depth

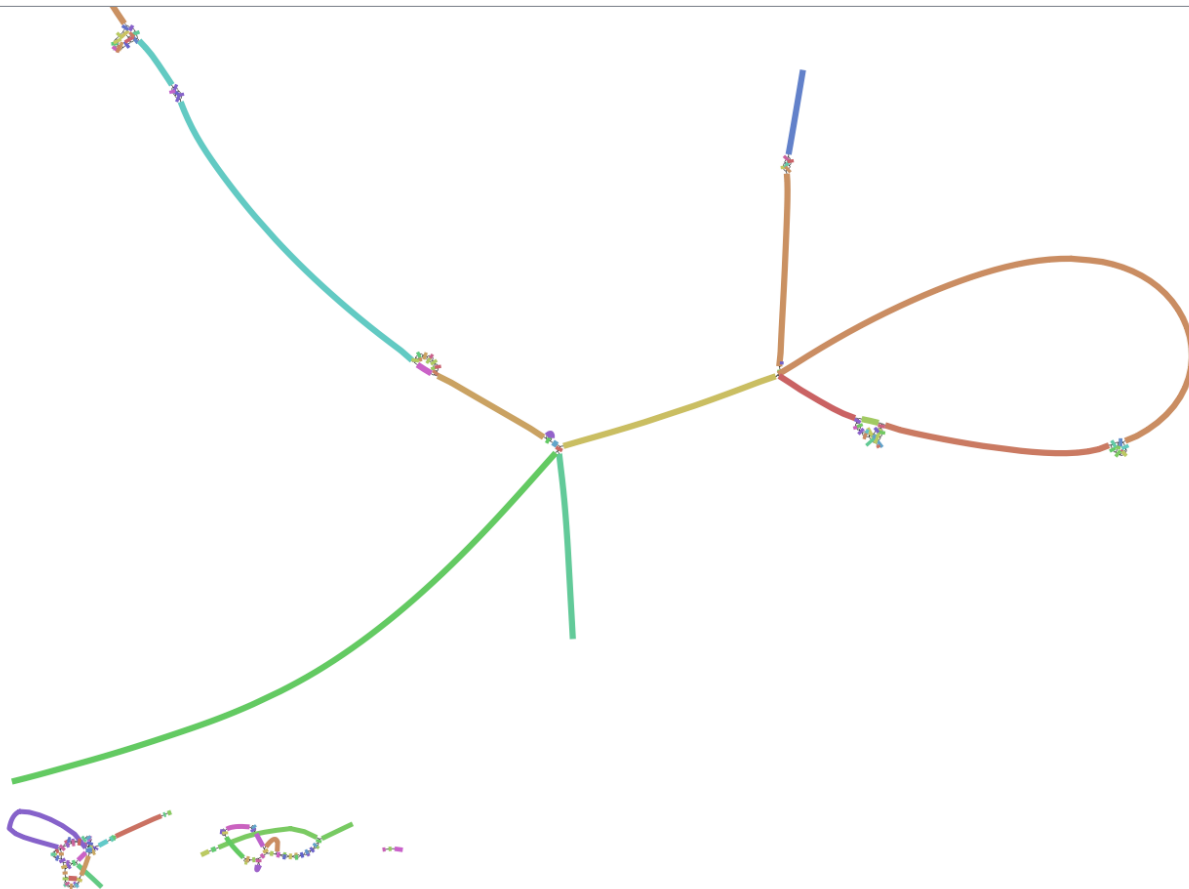
☐ BLAST hits

☐ CSV data:

☐ Tax Name (id)

Font

☐ Text outline



## В приложении Bandage добавлен следующий функционал

### Таксономический анализ

- Информация про таксоны вершины
- Раскраска по таксонам
- Продление раскраски
- Подписи таксонов
- Фильтрация по таксонам
- Фильтрация по таксонам с учетом Hi-C
- Отчет с общей информацией о таксонах
- Отчет с информацией о Hi-C связях таксона

### Сжатие графа

- Раскрытие сжатой вершины

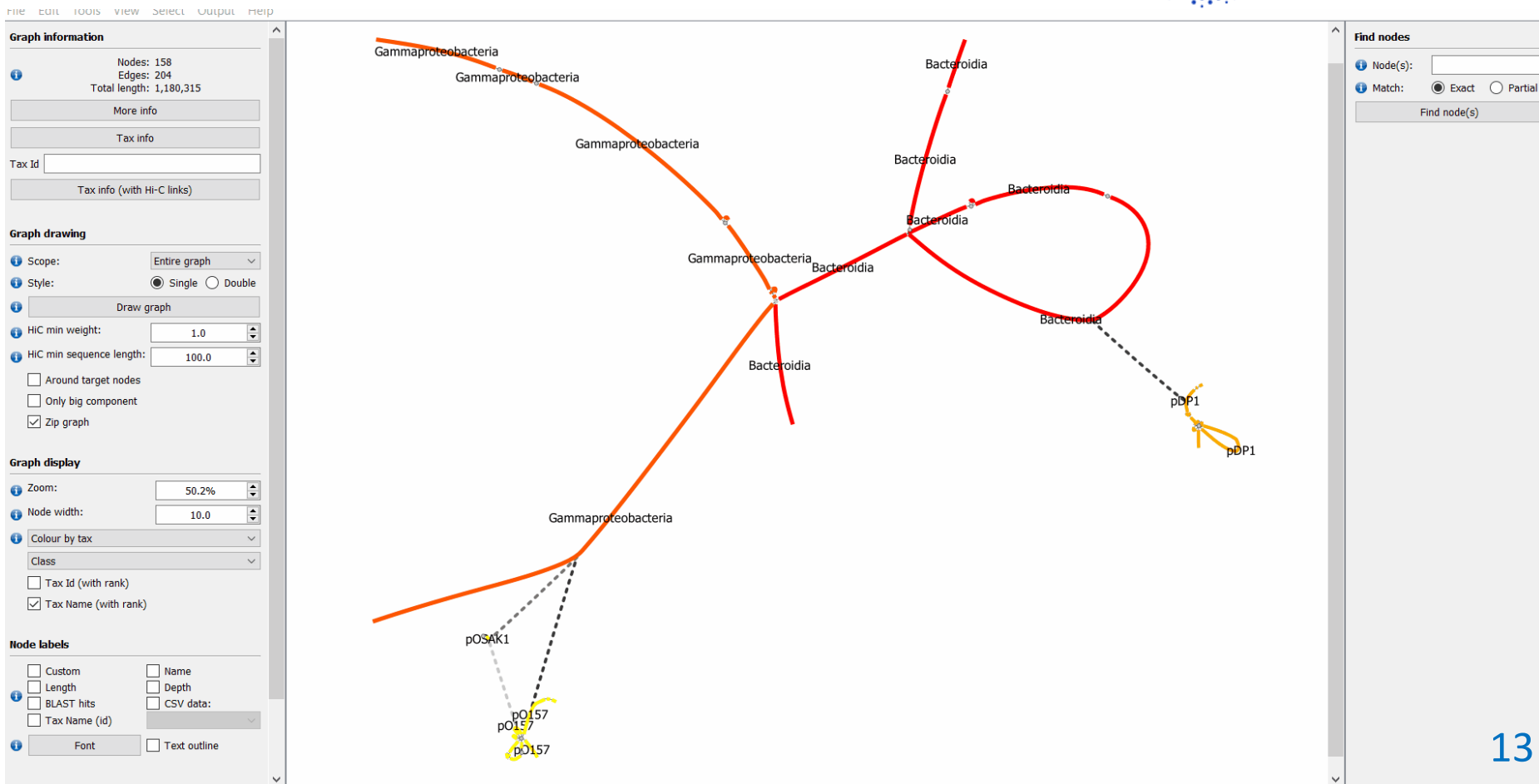
### Визуализация Hi-C связей

- Фильтры Hi-C ребер
- Минимальный вес Hi-C связи
- Минимальная длина контигов
- Hi-C связи между различными компонентами связности
- Ровно одна Hi-C связь между различными компонентами
- Раскраска по компонентам связности

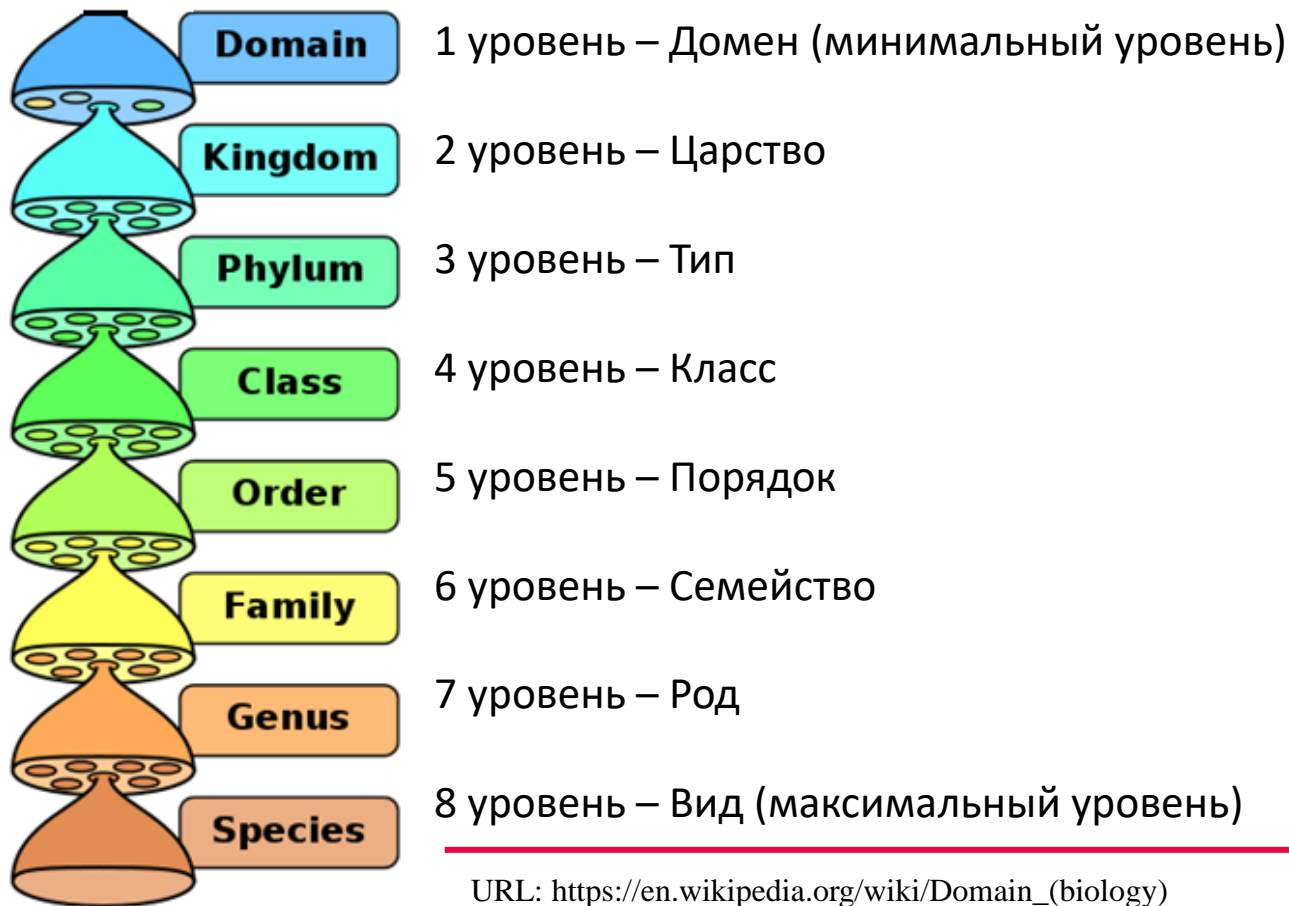
### Фильтры компонент связности

- Только большие компоненты связности
- Только связанные Hi-C ребрами с целевой компонентой

## Визуализация результатов таксономического анализа



## Хранение данных таксономического анализа



URL: [https://en.wikipedia.org/wiki/Domain\\_\(biology\)](https://en.wikipedia.org/wiki/Domain_(biology))

## Хранение данных таксономического анализа

### Список вершин:

Name: Node\_1

....

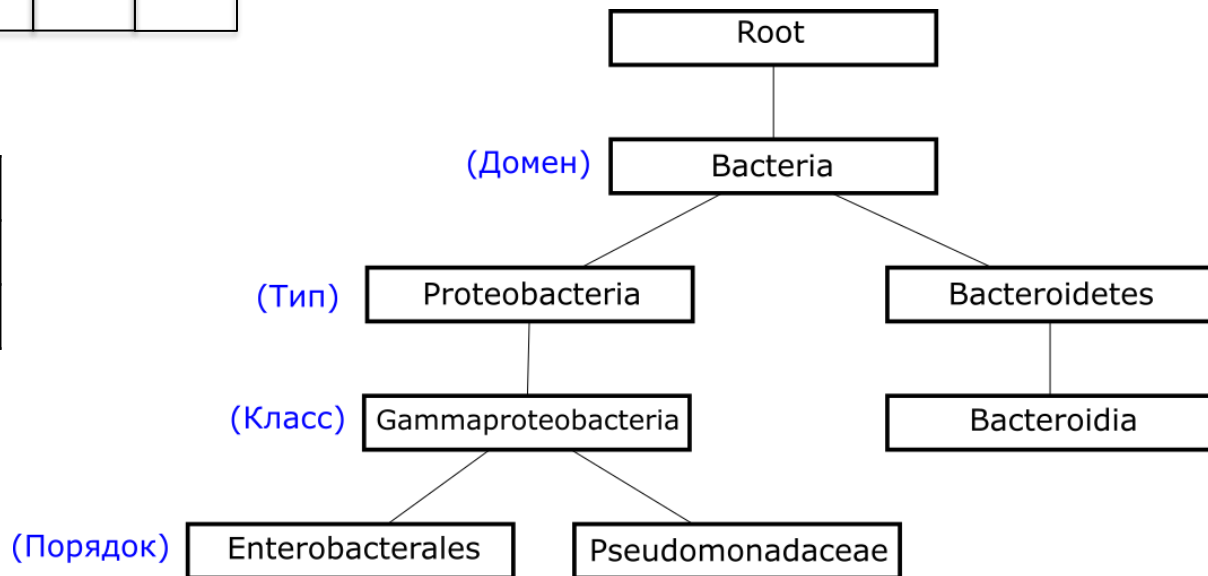
Тах\*: Gammaproteobacteria

Класс	Gammaproteobacteria
-------	---------------------

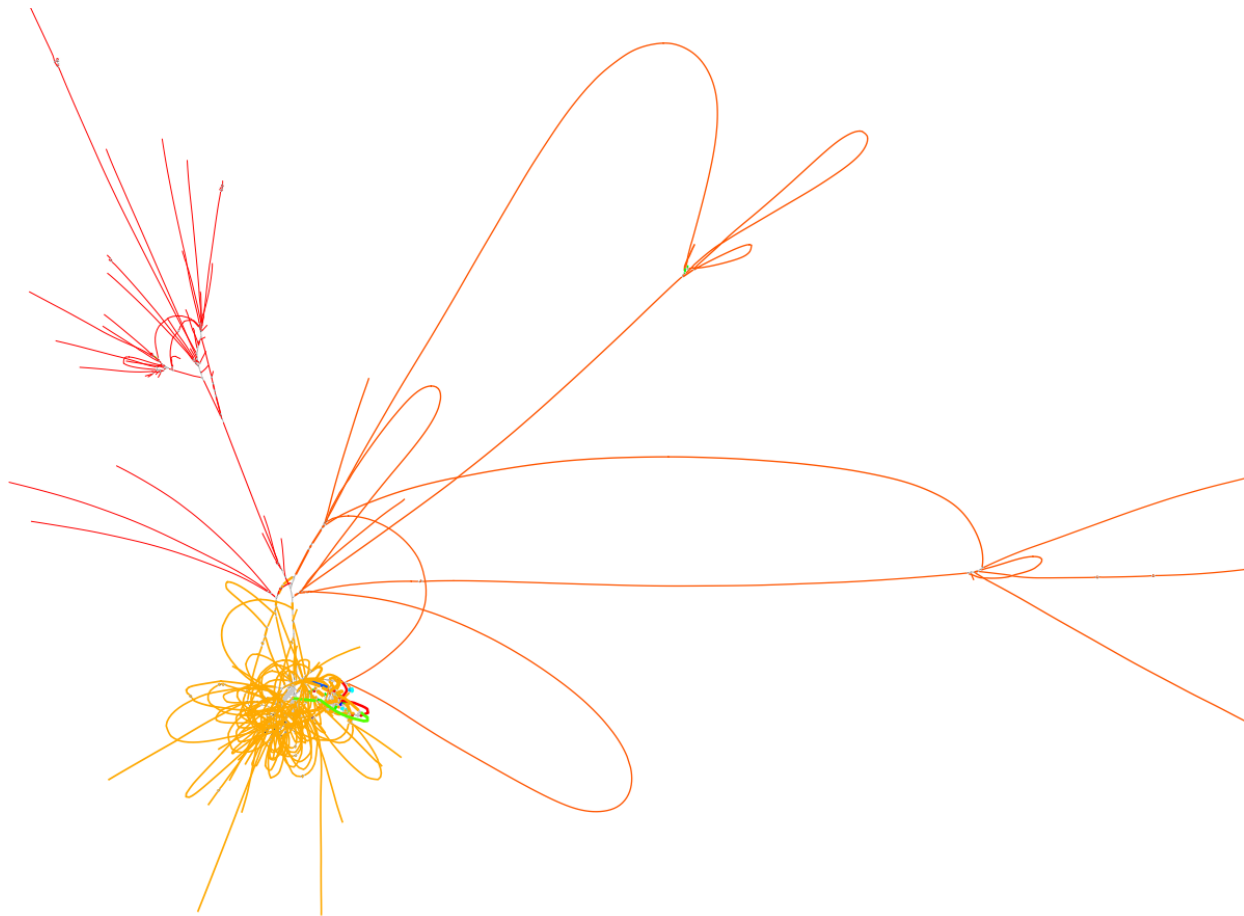
Тип	Proteobacteria
-----	----------------

Домен	Bacteria
-------	----------

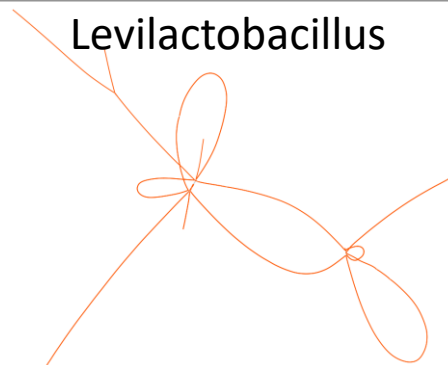
### Дерево таксонов



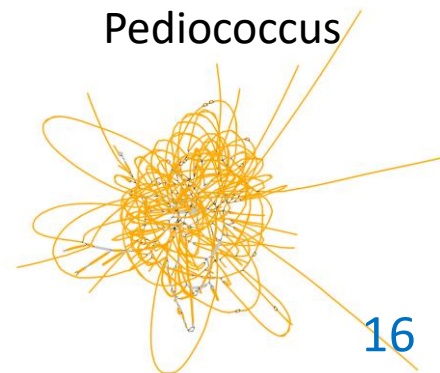
## Фильтрация вершин по таксонам



Levilactobacillus



Pediococcus





# Фильтрация вершин по таксонам с учетом Hi-C связей

**Graph information**

Nodes: 158  
Edges: 204  
Total length: 1,180,315

More info

Tax info

Tax Id

Tax info (with Hi-C links)

**Graph drawing**

Scope: Around tax

Style: ☒ Single ☐ Double

Draw graph

Tax id: 365491

☒ With Hi-C

☒ With Distance 5

HiC min weight: 1.0

HiC min sequence length: 100.0

☐ Around target nodes

☐ Only big component

☒ Zip graph

**Graph display**

Zoom: 135.4%

Node width: 10.0

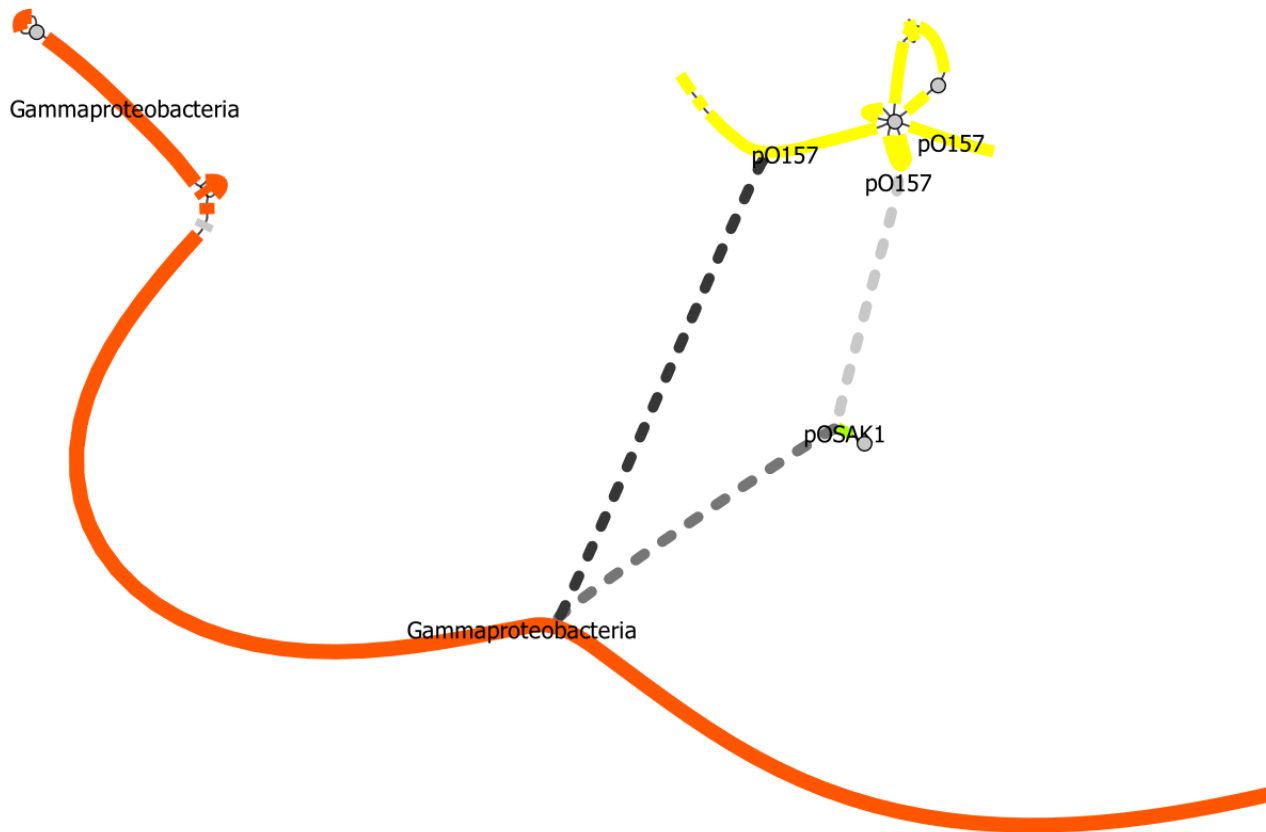
Colour by tax

Class

☐ Tax Id (with rank)

☒ Tax Name (with rank)

**Node labels**



## Сжатие графа

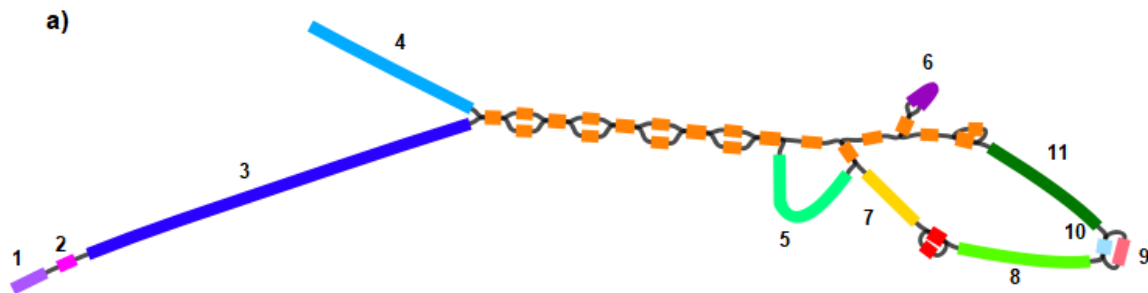


Рисунок 1. Несжатый граф

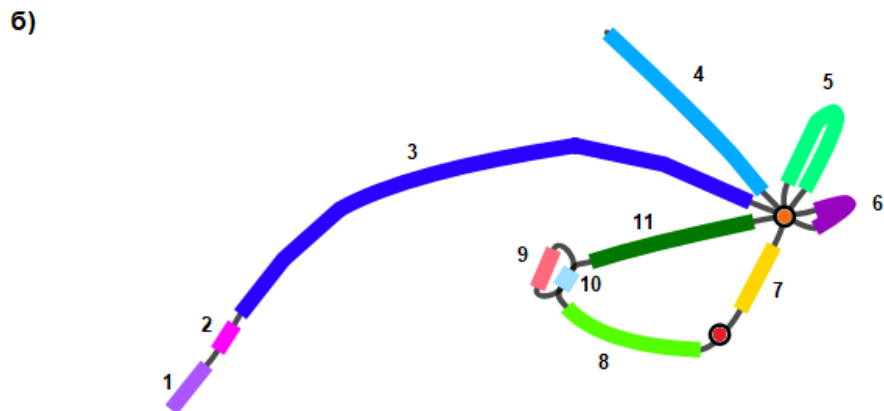


Рисунок 2. Сжатый граф

# Раскрытие одной сжатой вершины графа

**Graph information**

Nodes: 142  
Edges: 190  
Total length: 999,255

More info

Tax info

Tax Id

Tax info (with Hi-C links)

**Graph drawing**

Scope: Entire graph

Style: ☒ Single ☐ Double

Draw graph

HiC min weight: 1.0

HiC min sequence length: 100.0

☐ Around target nodes  
☐ Only big component  
☒ Zip graph

**Graph display**

Zoom: 73.5%

Node width: 10.0

Random colours

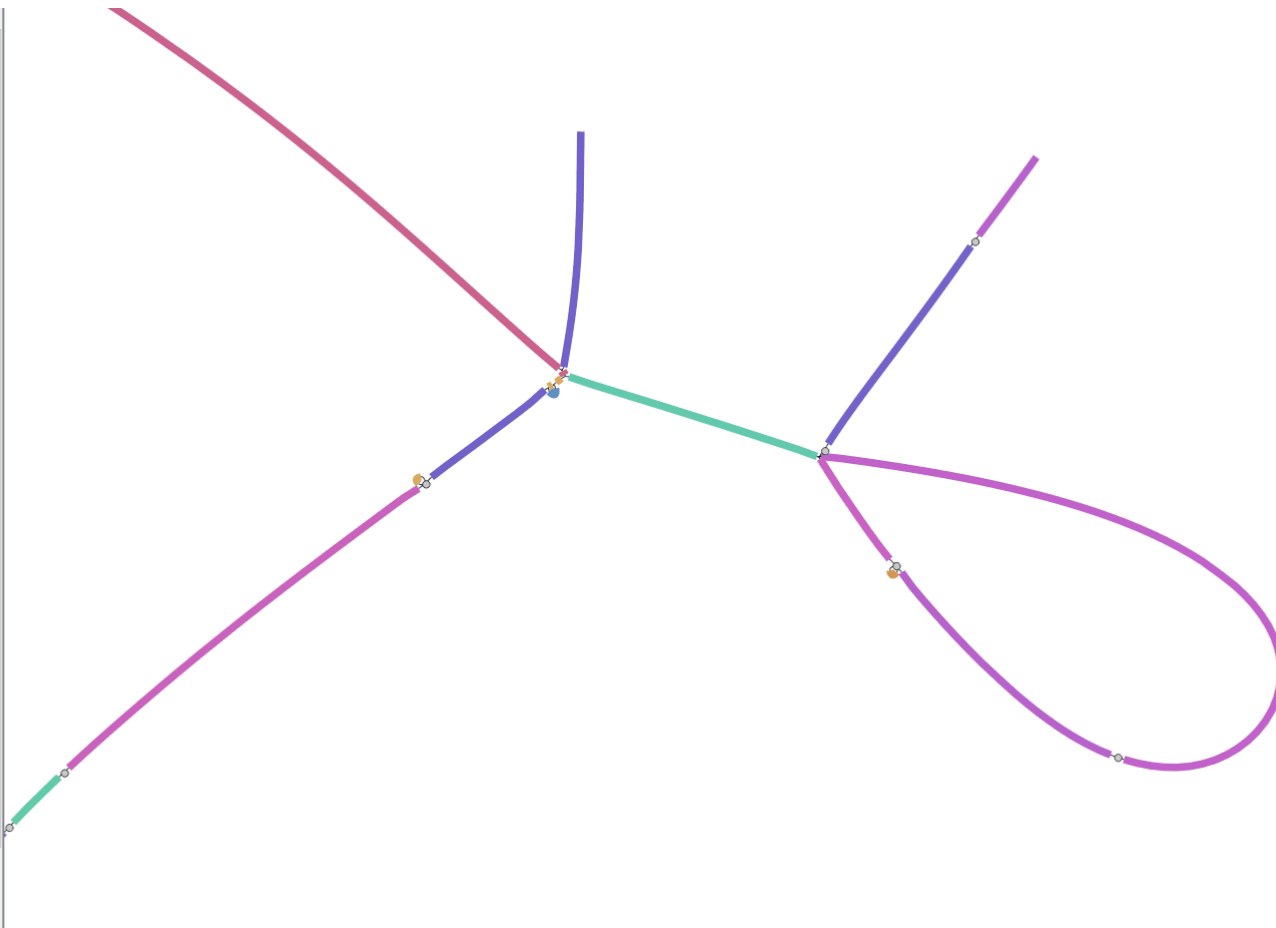
Domain

☐ Tax Id (with rank)  
☐ Tax Name (with rank)

**Node labels**

☐ Custom ☐ Name  
☐ Length ☐ Depth  
☐ BLAST hits ☐ CSV data:  
☐ Tax Name (id)

Font ☐ Text outline



**Find nodes**

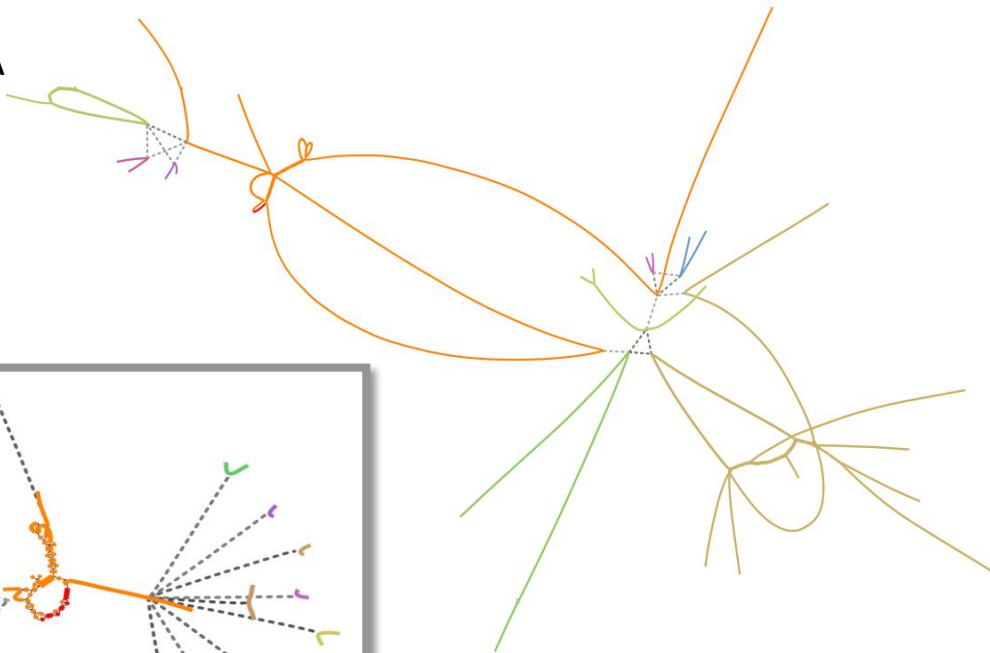
Node(s): union9

Match: ☒ Exact ☐ Partial

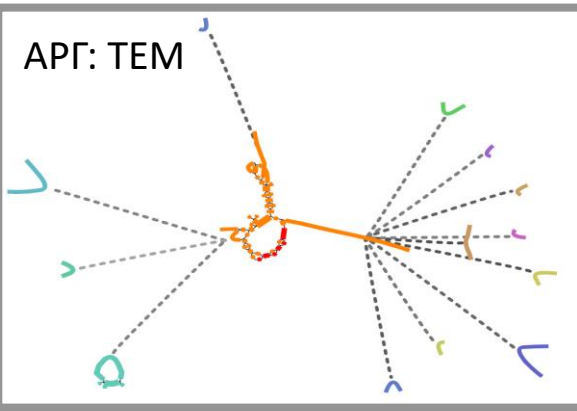
Find node(s)

## Тестирование на реальных данных

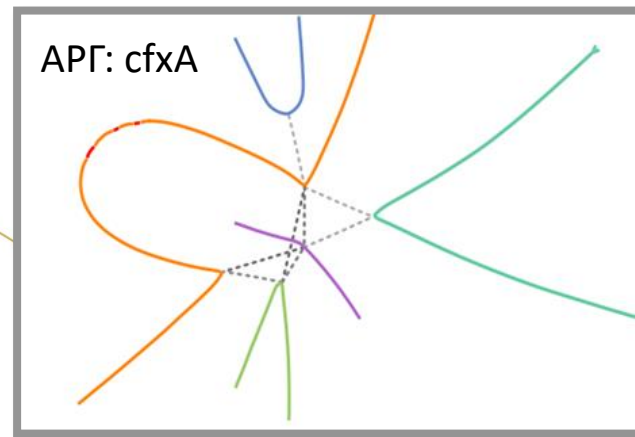
АРГ: ОХА



АРГ: TEM



АРГ: cfxA



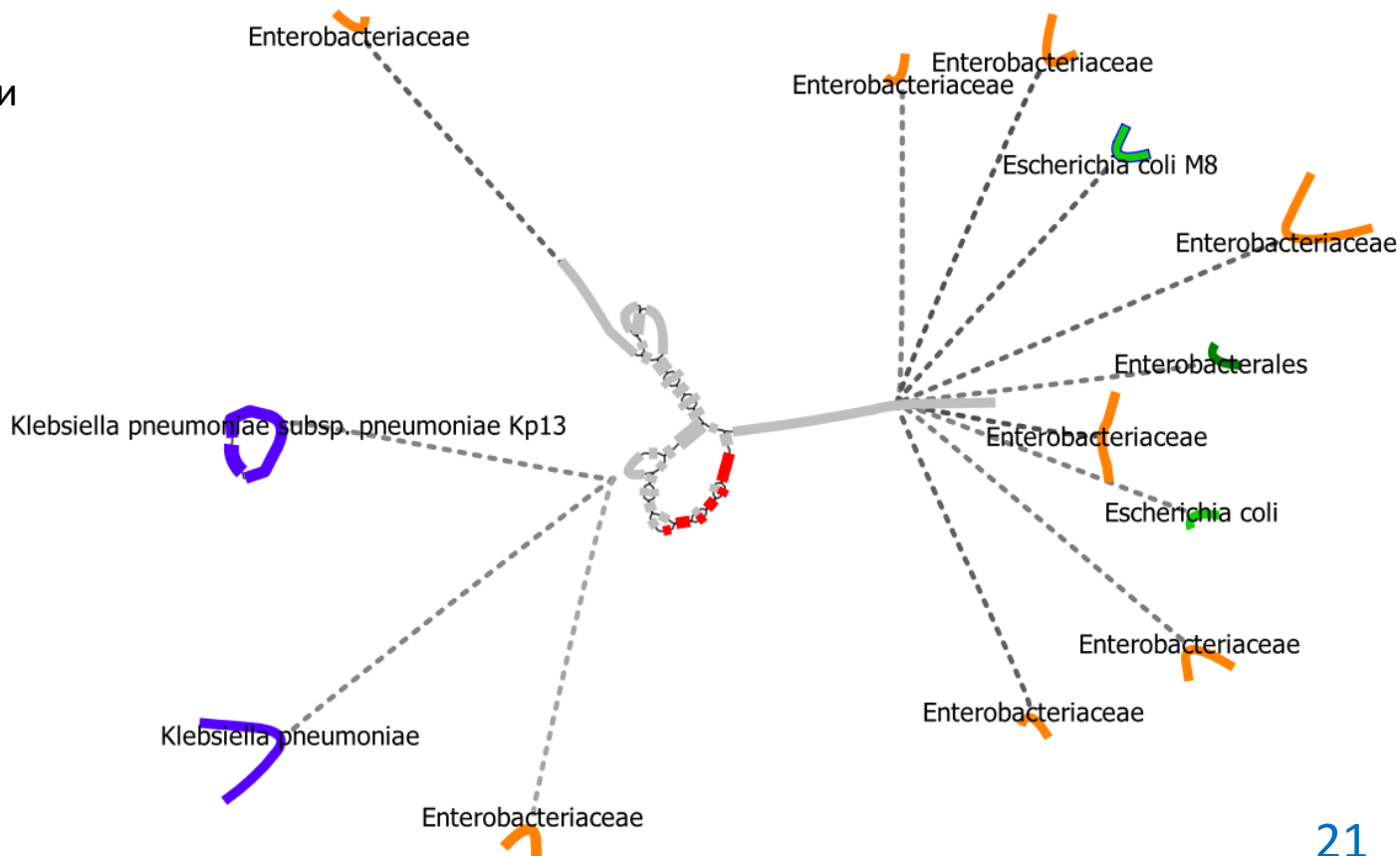
Ivanova V., Chernevskaya E., Vasiluev P., Ivanov A. et al. Hi-C Metagenomics in the ICU: Exploring Clinically Relevant Features of Gut Microbiome in Chronically Critically Ill Patients // Front Microbiol. 2022. Vol 12. URL: <https://doi.org/10.3389/fmicb.2021.770323> (дата обращения: 03.03.2022)

## Тестирование с данными таксономического анализа

АРГ: TEM

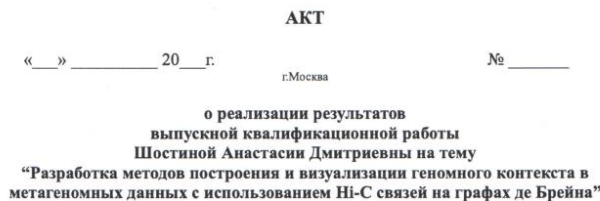
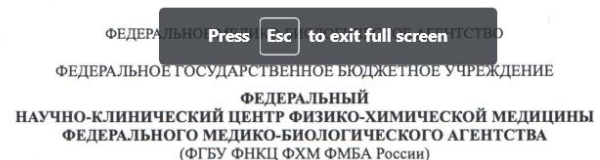
Hi-C связи с семействами  
бактерий:

- 1) Escherichia
- 2) Klebsiella



# ЗАКЛЮЧЕНИЕ

- 1) Разработан метод построения геномного контекста с учетом Hi-C связей.
- 2) Реализован новый функционал в приложении Bandage:
  - 2.1) Отображение Hi-C связей
  - 2.2) Отображение таксонов
  - 2.3) Сжатие графа
- 3) Проведено тестирование на сгенерированных и реальных метагеномных данных.
- 4) Участие на конференциях:
  - 4.1) XI Конгресс молодых ученых в Университете ИТМО
  - 4.2) 51-ая научная и учебно-методическая конференция Университета ИТМО.
- 5) Акт о внедрении (центр физико-химической медицины).



Настоящий акт подтверждает то, что результаты выпускной квалификационной работы Шостиной А.Д. на тему “Разработка методов построения и визуализации геномного контекста в метагеномных данных с использованием Hi-C связей на графах де Брейна” успешно внедрены в проекте “Поиск воспроизводимых метагеномных маркеров микробиоты кишечника ассоциированных с ответом на иммунотерапию рака”, реализуемом в ФГБУ Федеральный научно-клинический центр Физико-Химической медицины Федерального медико-биологического агентства России. Данная работа позволила усовершенствовать поиск воспроизводимых микробных маркеров в наборах кишечных метагеномов из открытых источников (см. Приложение 1).

Заместитель генерального директора  
по научной работе



В.Н. Лазарев

# Спасибо за внимание!

[www.ifmo.ru](http://www.ifmo.ru)

IT'sMO<sup>re</sup> than a  
UNIVERSITY