Разработка алгоритмов скаффолдинга при помощи дополнительной геномной информации

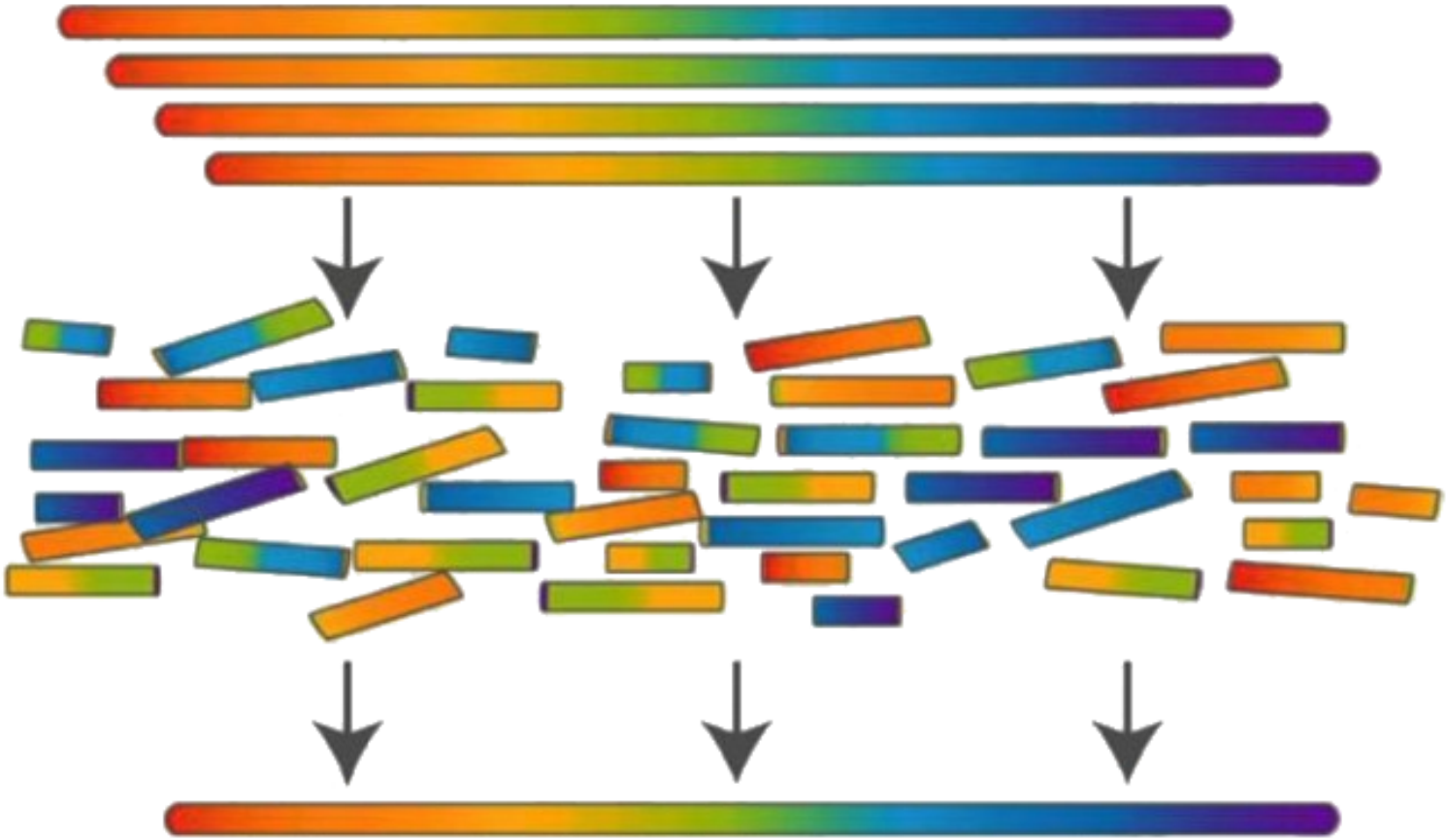
Клещин Антон Сергеевич, 16.Б10-мм

Научный руководитель: доц. каф. СП, к.т.н. Ю. В. Литвинов
Консультанты: доц. каф. стат. мод., к.ф.-м.н. А. И. Коробейников
Научный сотрудник Центра алгоритмической биотехнологии СПбГУ А. Д. Пржибельский
Рецензент: PhD кандидат Корнельского университета Д. А. Мелешко

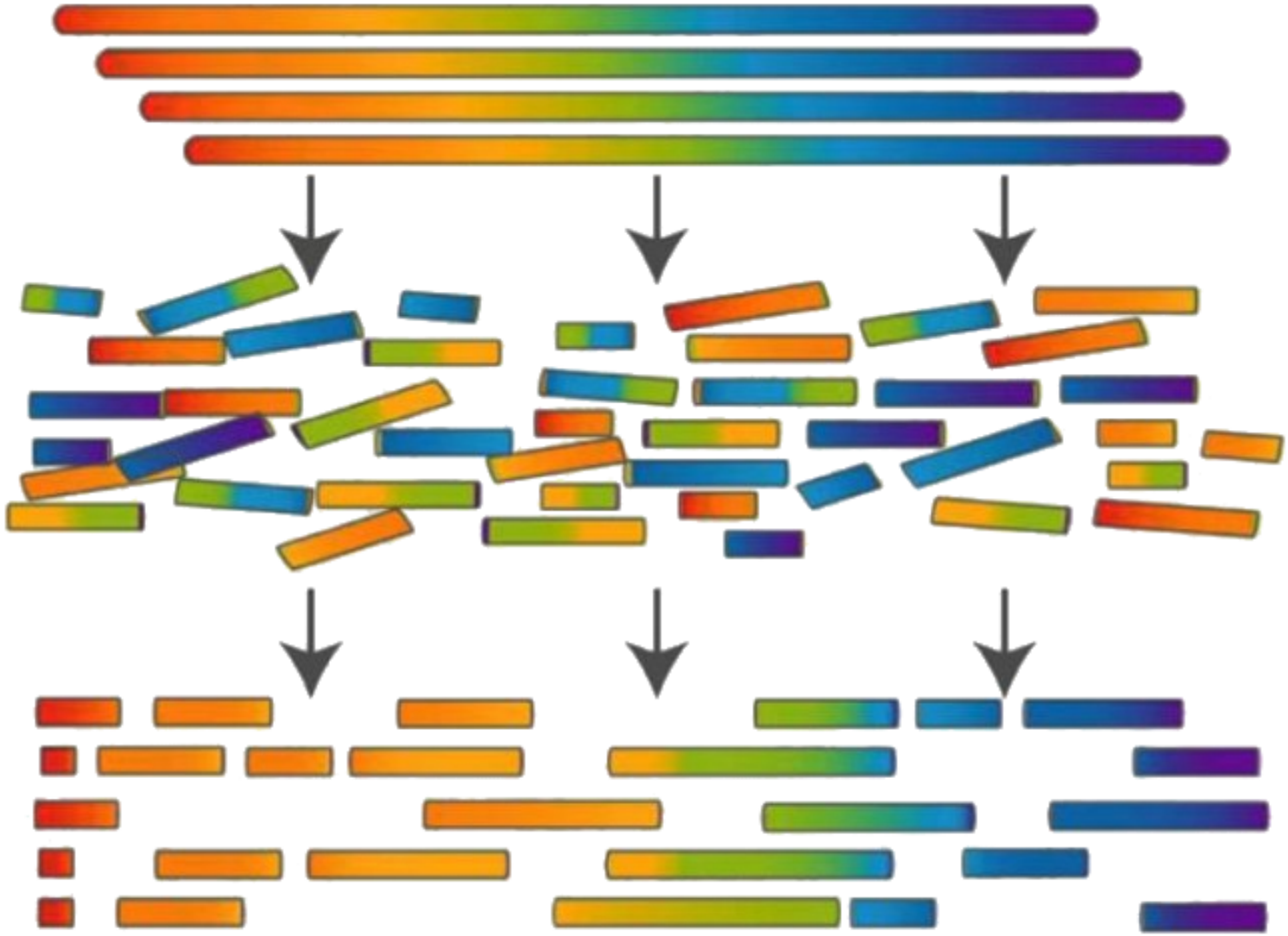
СПбГУ

9 июня 2020 г.

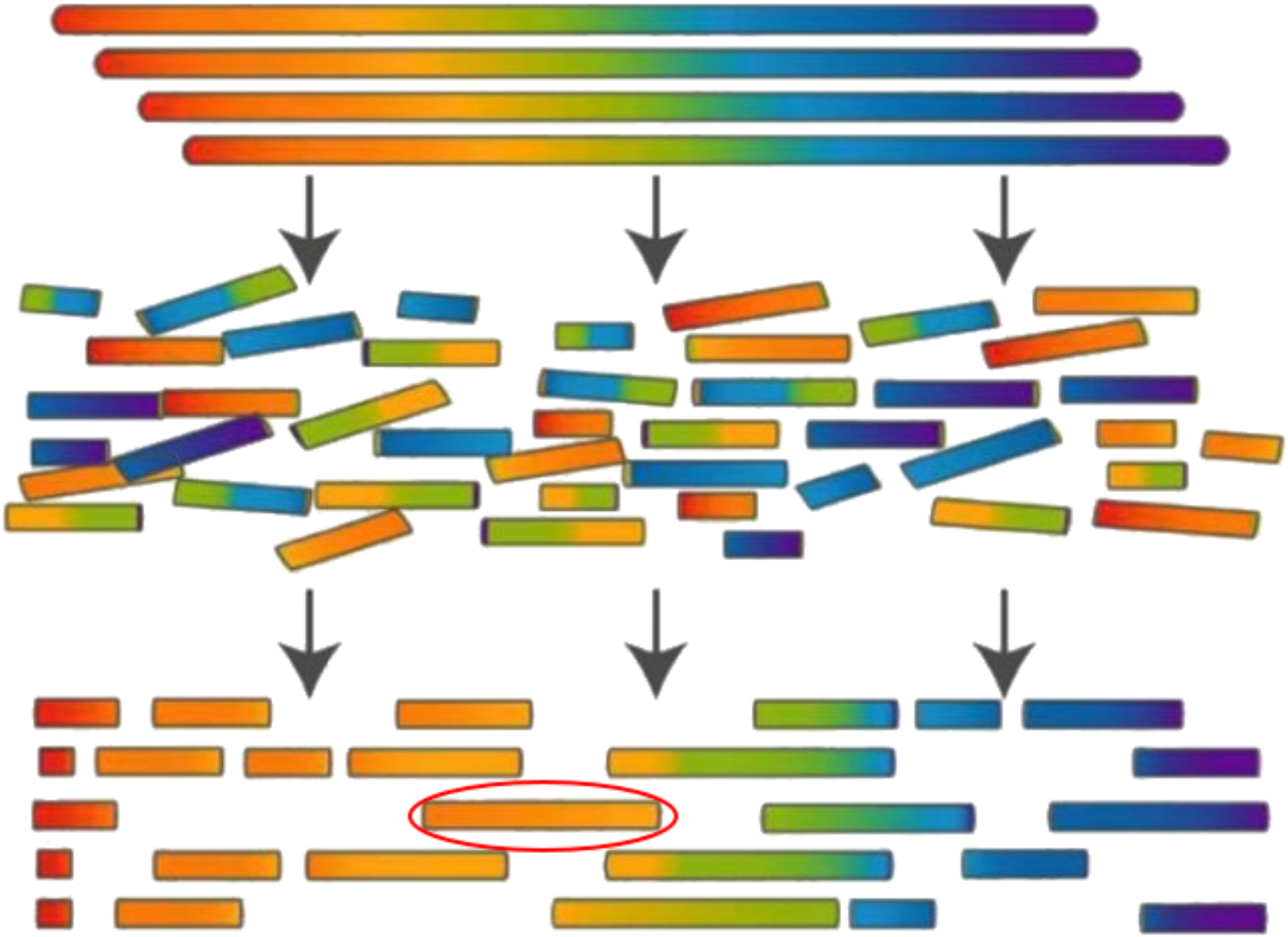
В идеальном мире



В реальности



В реальности



Мотивация

- Переиспользование накопленных результатов
- Использование результатов сторонних ассемблеров
- Использование похожих геномов
- Упрощение метагеномной сборки использованием референсных геномов

Постановка задачи

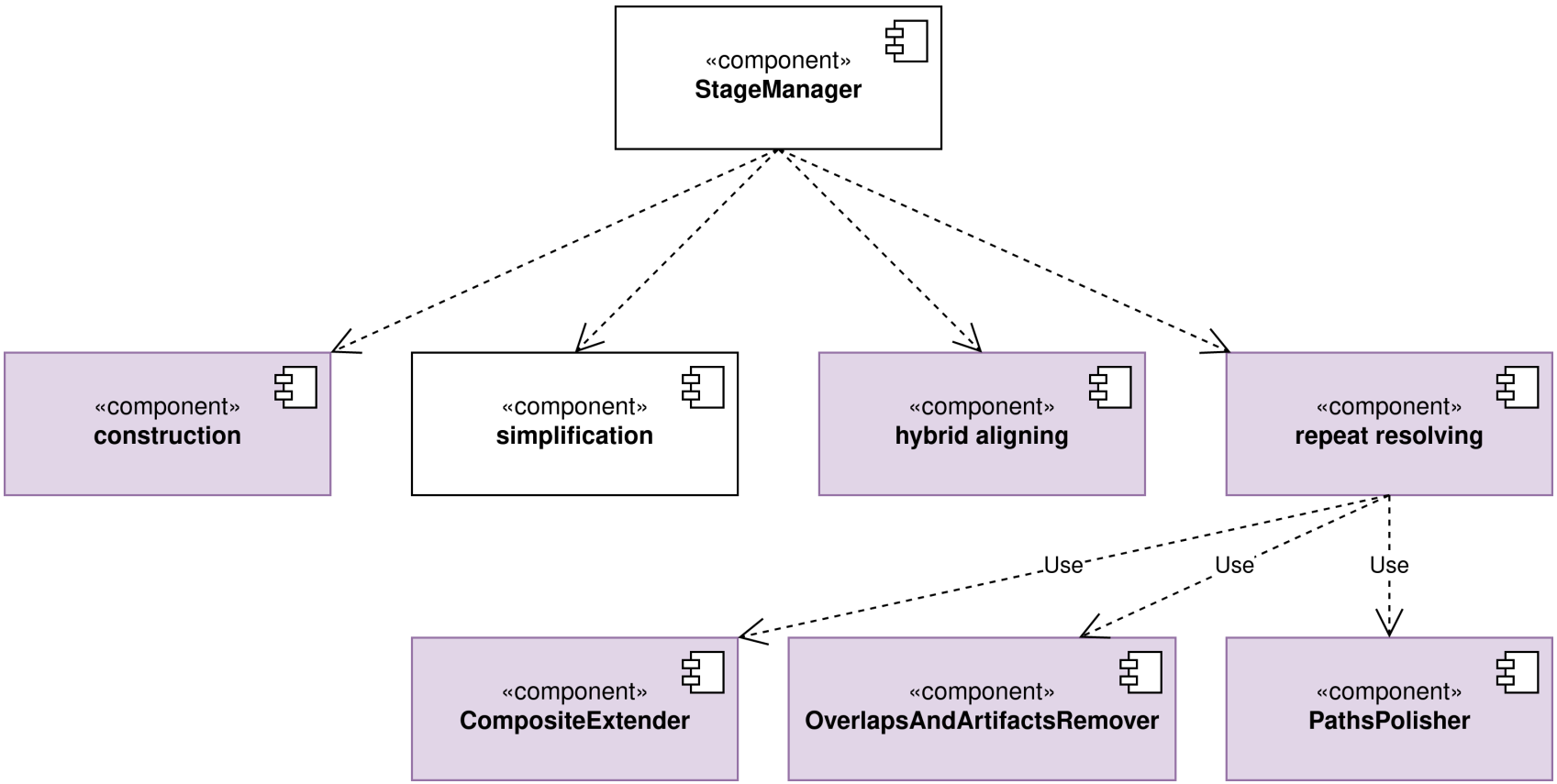
Цель — добавление поддержки контигов в качестве входных данных для геномного ассемблера SPAdes

- Разработка алгоритма скаффолдинга, использующего контиги
- Реализация расширения для геномного ассемблера SPAdes
- Тестирование алгоритма на известных геномах

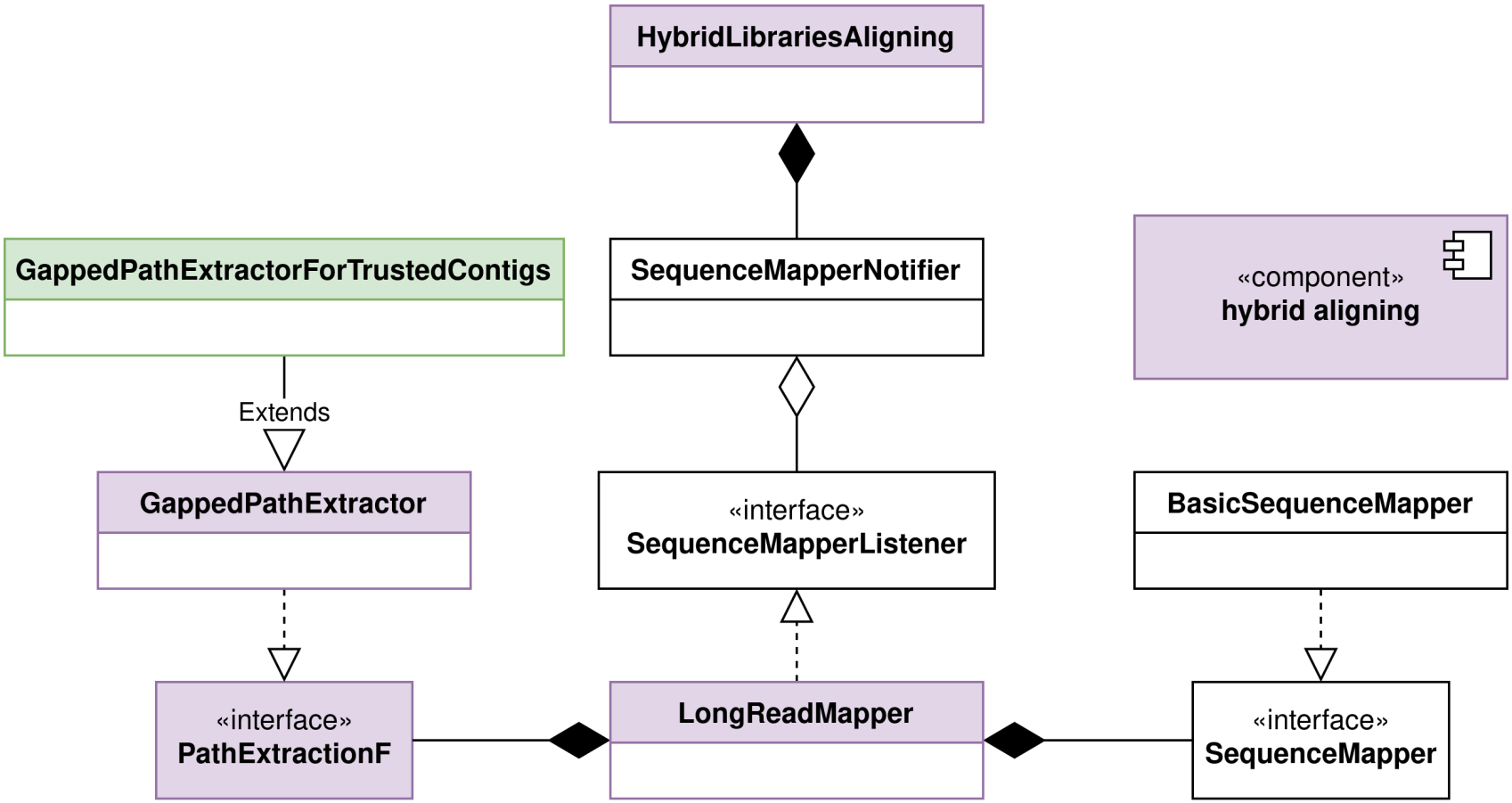
Алгоритм

- Построение графа сборки
  - Короткие риды
  - Контиги
- Упрощение графа
- Выравнивание контигов на граф
- Выращивание путей
  - Точное совпадение с путями выравненных контигов
  - Неточное совпадение
- Постобработка

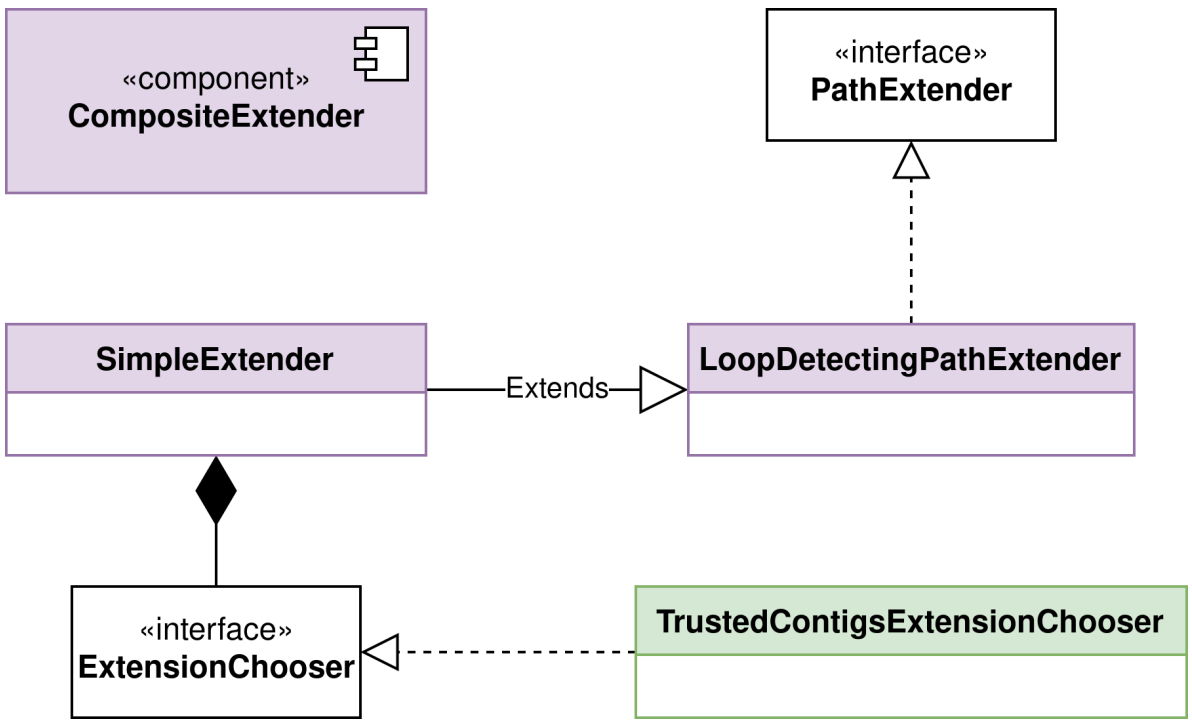
Архитектура



Архитектура



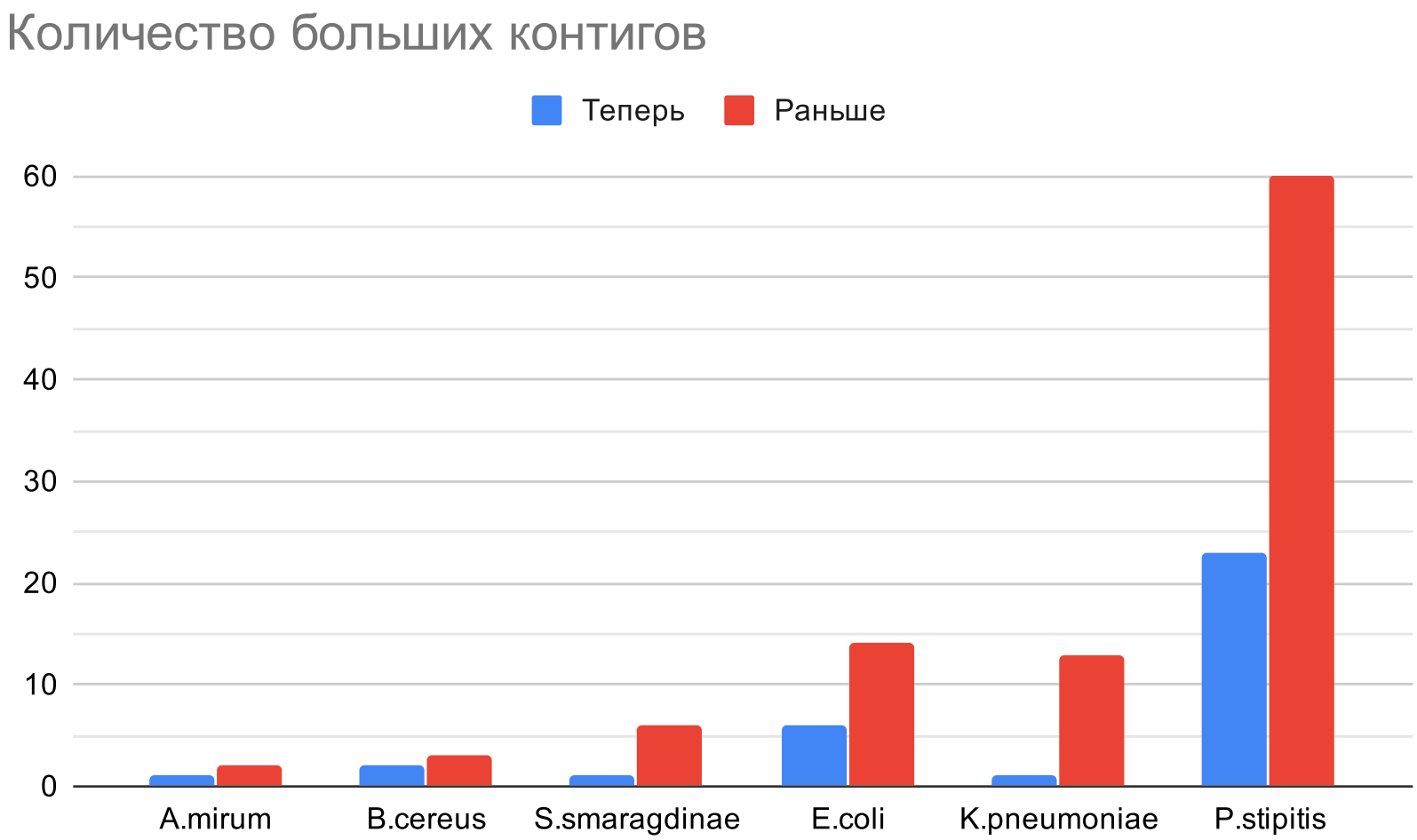
Архитектура



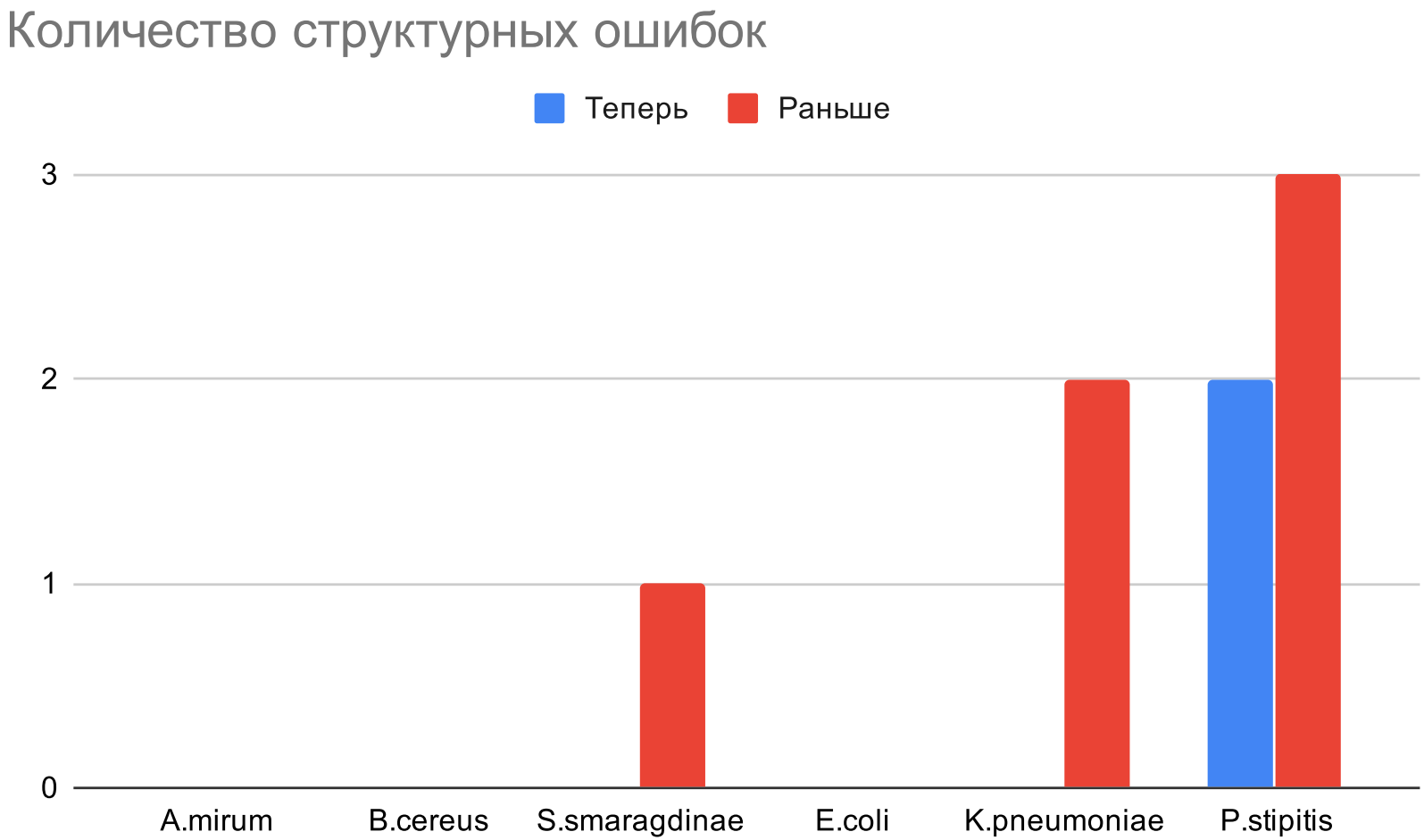
Метрики

- Количество больших контигов
- Количество структурных ошибок

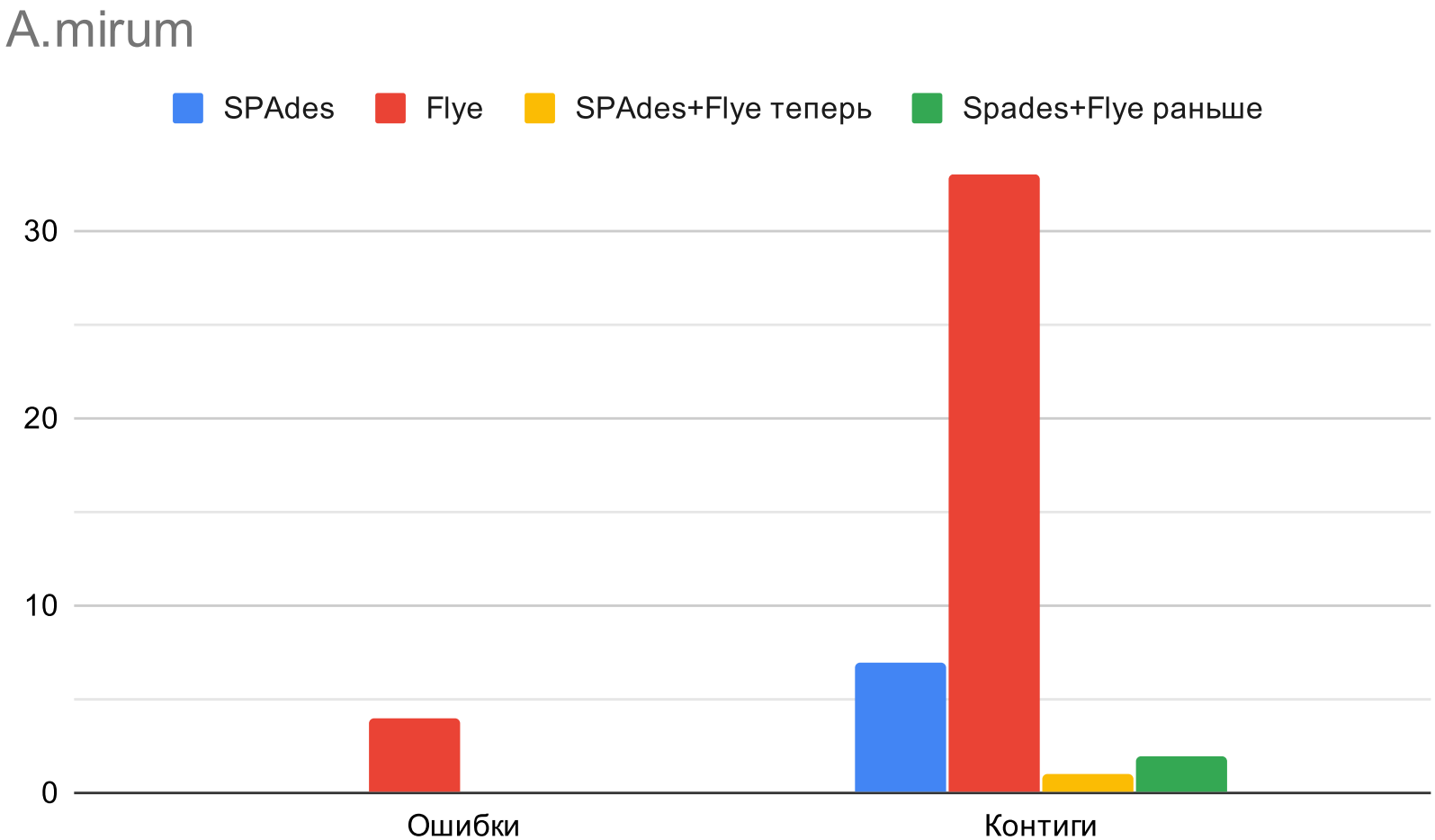
Тестирование



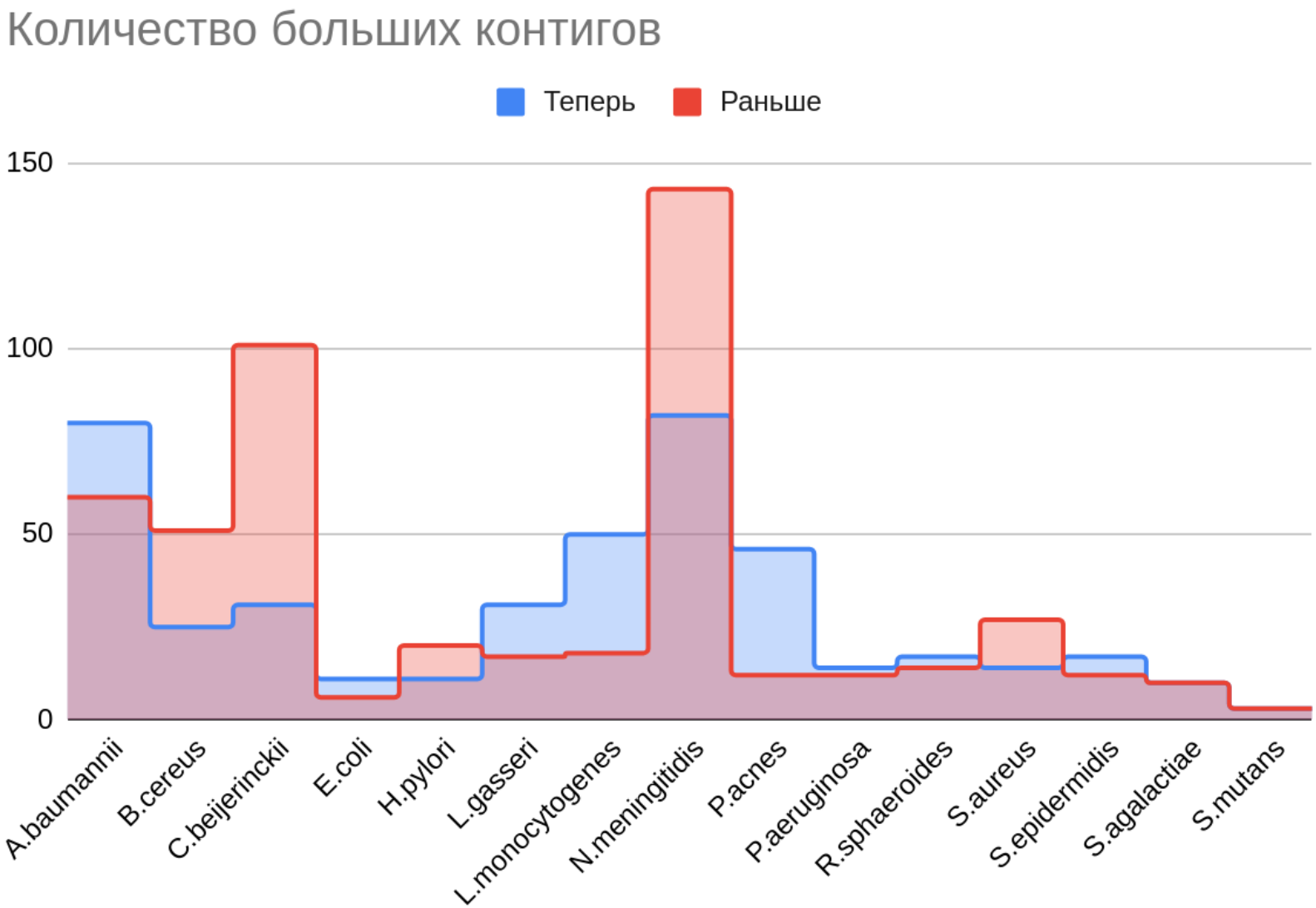
Тестирование



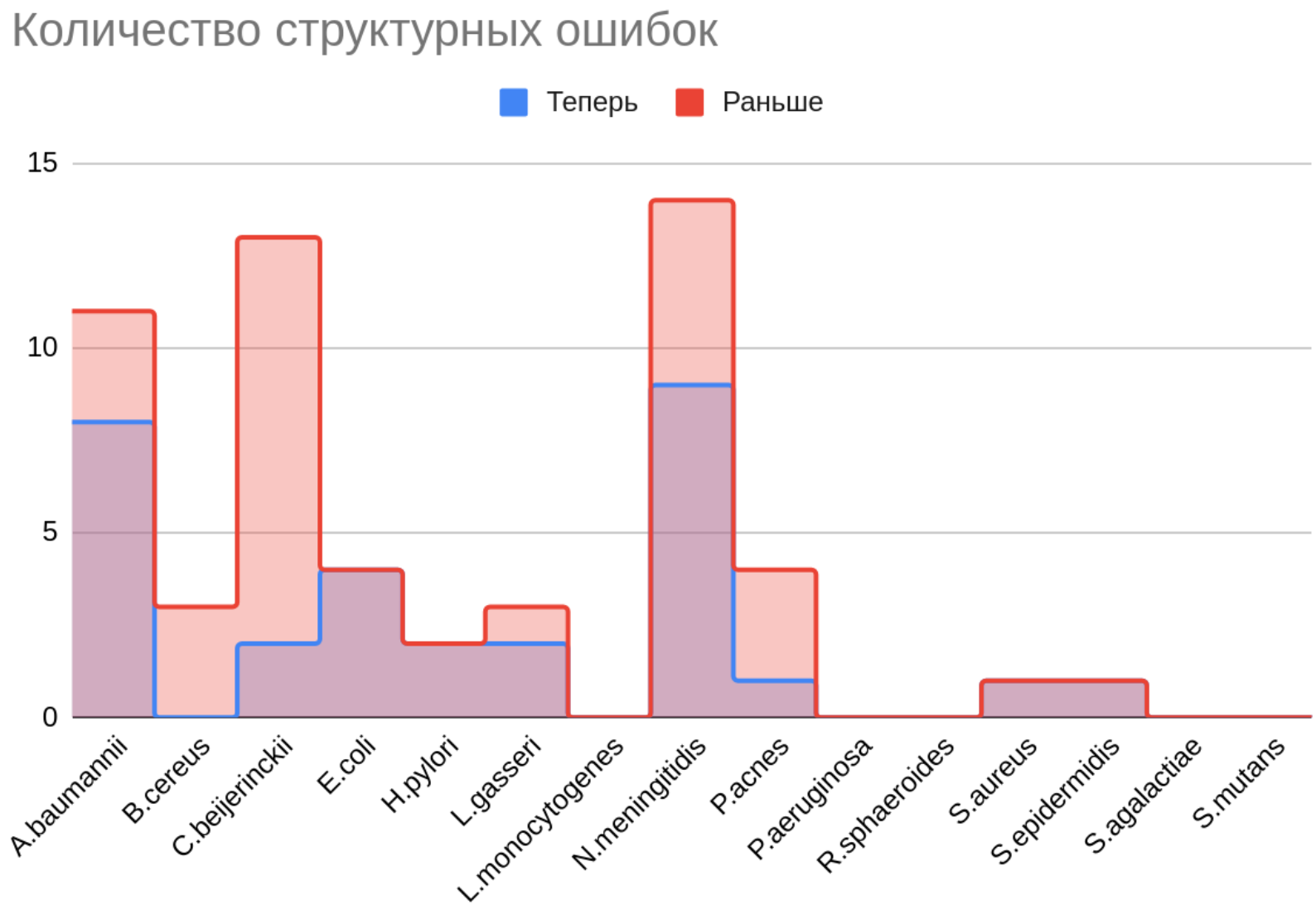
Тестирование, Flye



Тестирование, метагеном



Тестирование, метагеном



Заключение

- Разработан алгоритм скаффолдинга, использующего контиги
  - Алгоритм выравнивает контиги на граф сборки, а затем использует полученные пути при разрешении повторов
- Реализовано расширение для геномного ассемблера SPAdes
  - Реализовано на языке C++
  - Расширение позволяет эффективно использовать контиги в качестве входных данных
  - Исходный код SPAdes доступен по ссылке: `https://github.com/ablab/spades/`.
- Алгоритм протестирован на известных геномах
  - Протестировано на сборках одиночных геномов с высоким и низким качеством входных контигов, а также на метагеномной сборке
  - Теперь соединяется больше контигов с меньшим количеством ошибок по сравнению с предыдущим модулем SPAdes