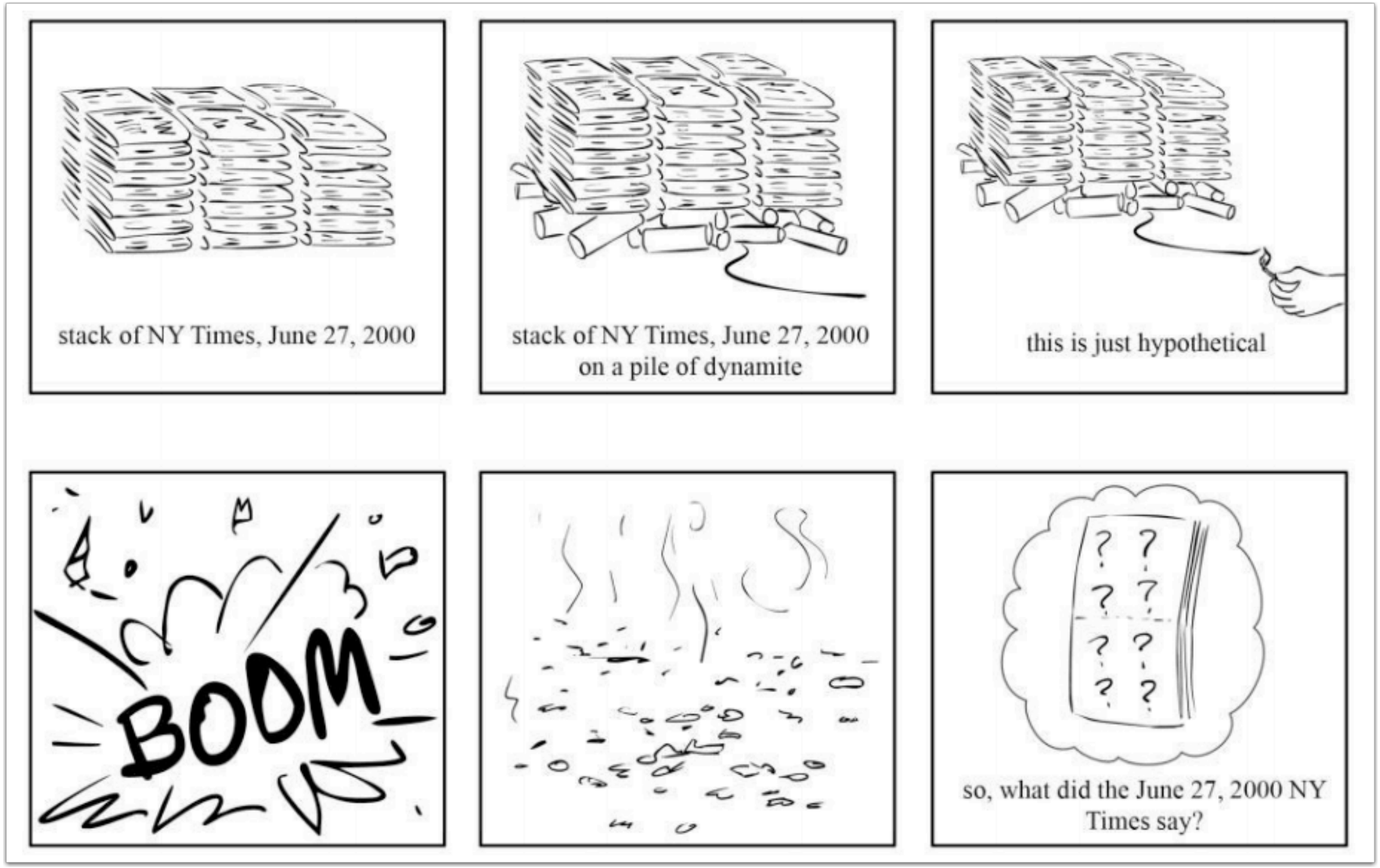
Исследование задачи поиска генов в метагеномных сборках de novo

Соса Екатерина, 445 группа
Руководитель: Минкин Илья Валериевич

Введение: определения

- геном – длинная строка в алфавите {A, C, G, T}
- сборка – множество длинных подстрок генома
- ген – короткая подстрока генома
- метагеномная сборка – перемешанные сборки разных геномов

Введение: иллюстрация



Особенности

- Метагеномная сборка подразумевает, что в ней находятся разные организмы, которые ранее изучались по-отдельности
- “de novo” означает, что ничего о рассматриваемых организмах заранее не известно

Постановка задачи

- Придумать и реализовать инструмент для сравнения существующих методов предсказывания генов
- Предложить и проанализировать метод кластеризации метагеномной сборки

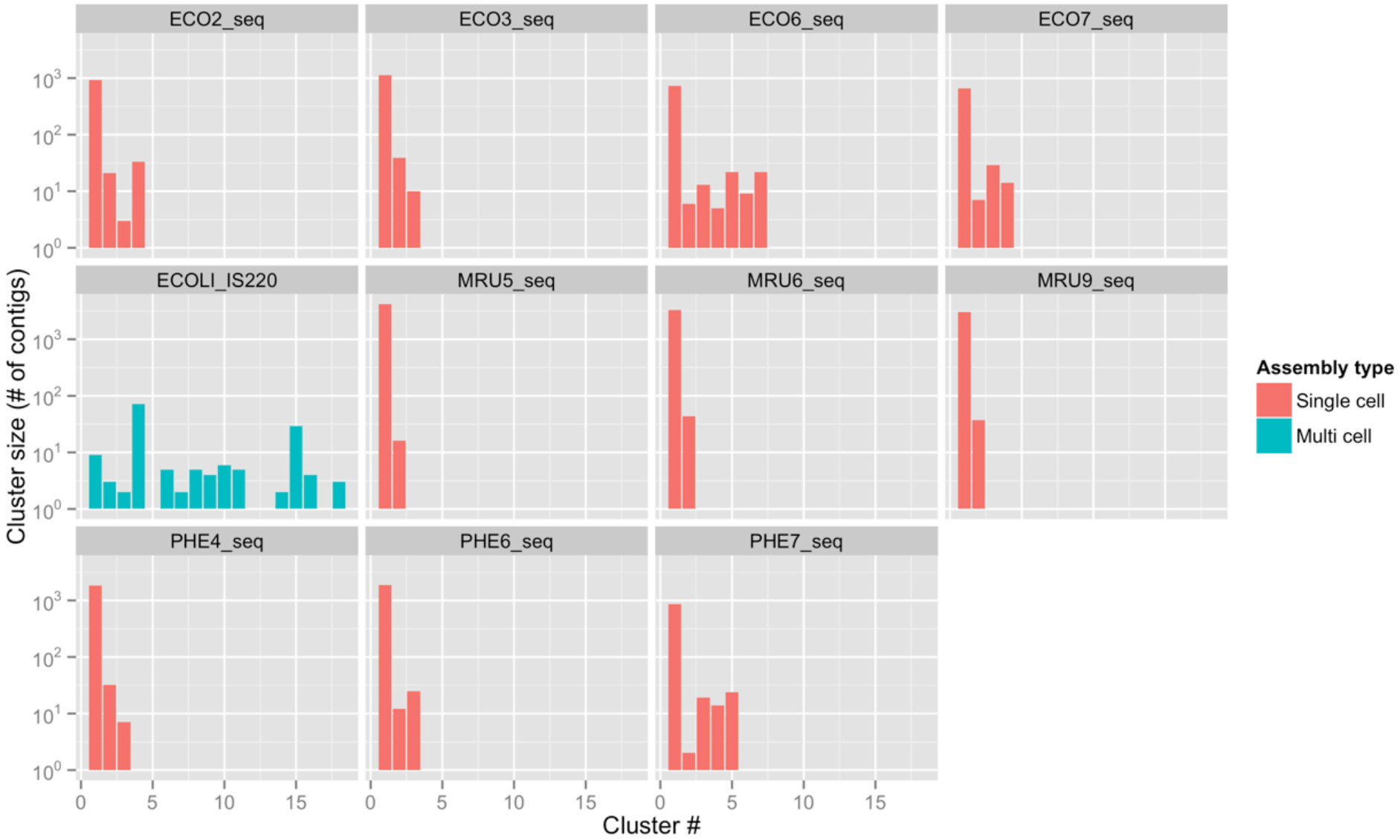
Реализация инструмента

- Понятность использования
- Расширяемость на аналогичные задачи
- Визуализация результатов
- Возможность распараллеливания

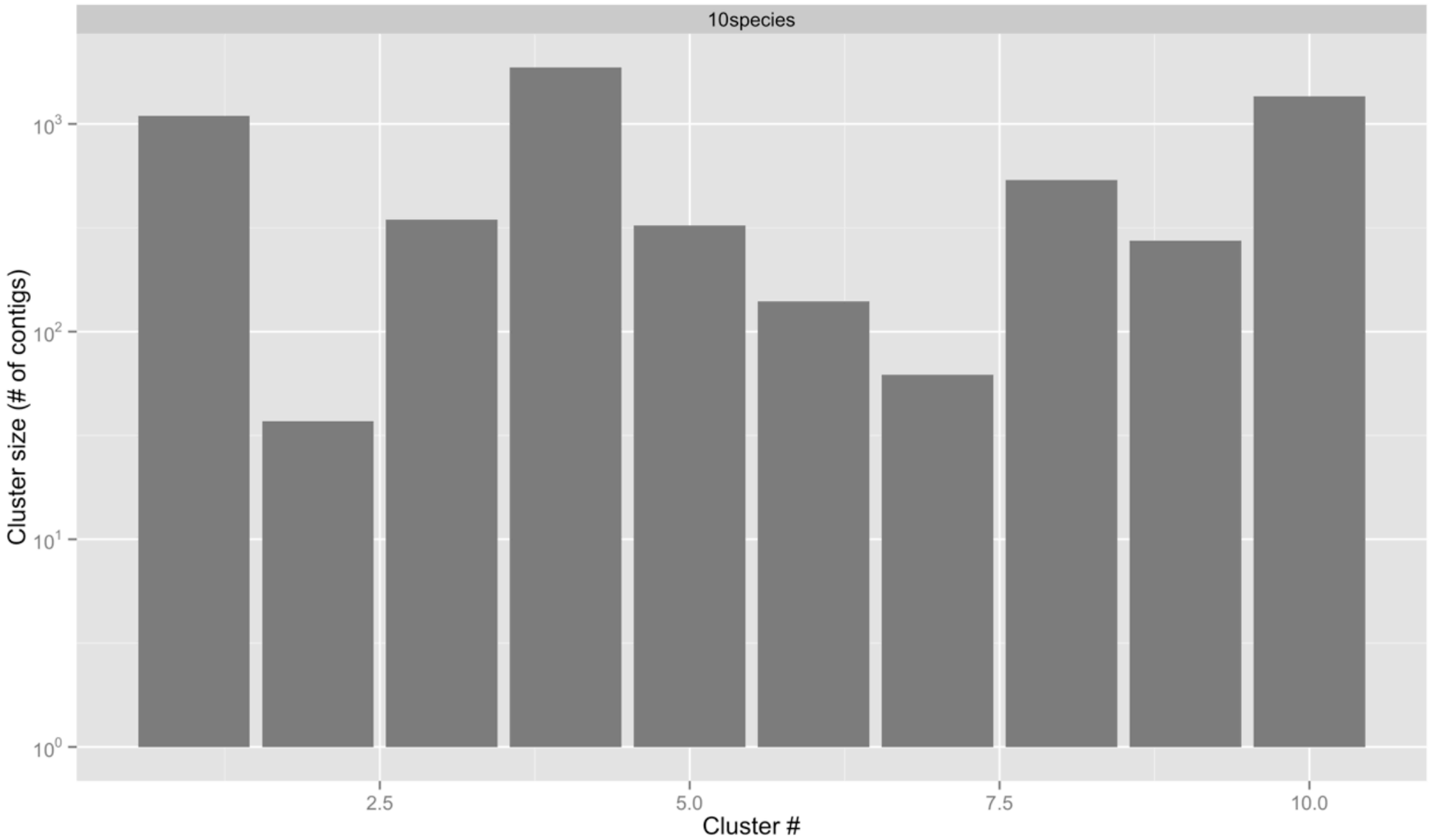
Кластеризация

- k-means, k-means++, mini-batch k-means
- байесовский информационный критерий для выбора k
- параметр для кластеризации – частоты n-мер

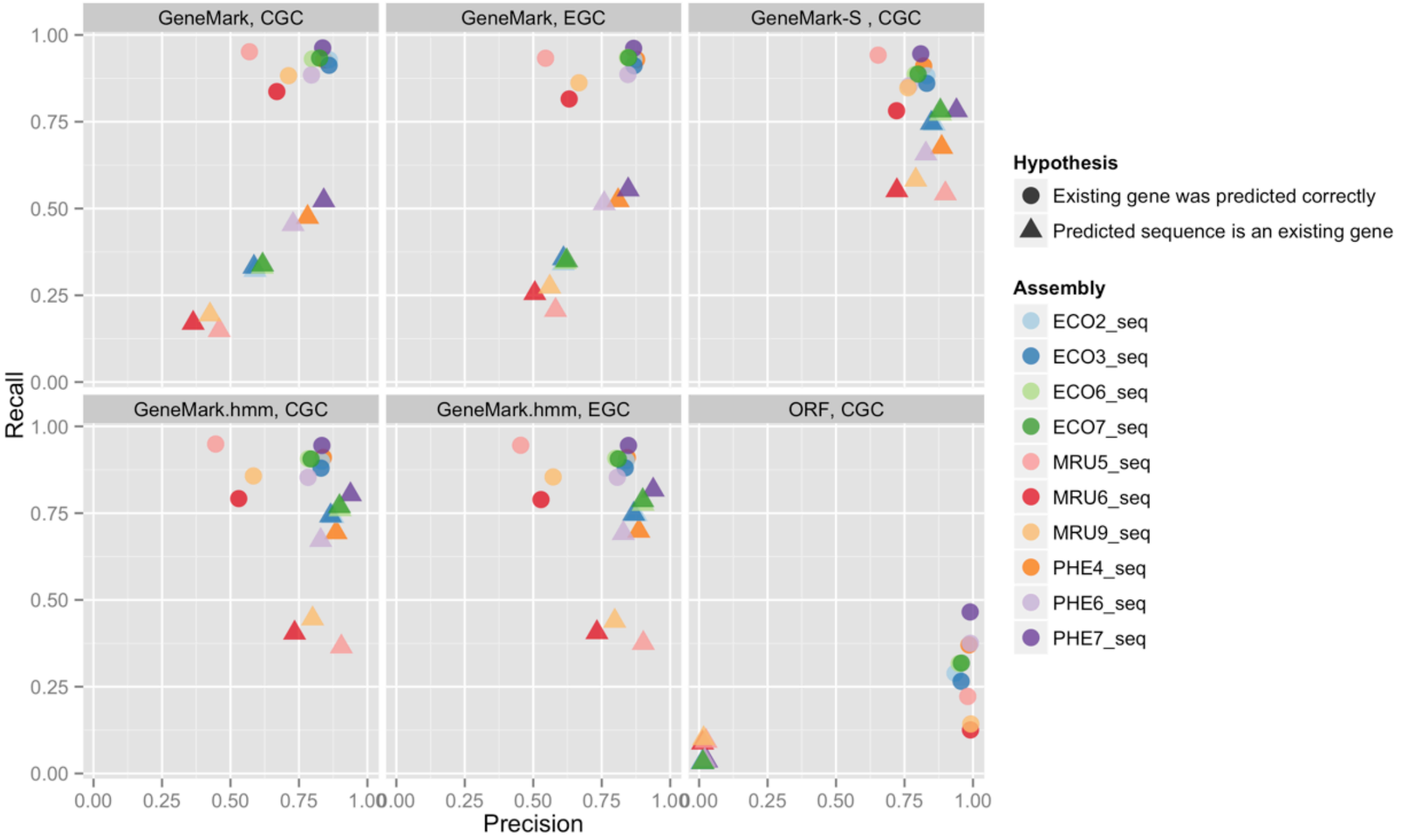
Результаты: кластеризация



Результаты: кластеризация



Результаты: поиск генов



Заключение

- Предложен и реализован метод кластеризации. Параметр выбран неверно, и его стоит заменить на метрику на основе IMM.
- Предложена и реализована оптимизация методов поиска генов на основе HMM, встроена в QUAST.
- Реализован инструмент для сравнения методов, проведено сравнение.