Подбор оптимальных параметров rnaSPAdes для сборки транскрипотомов

МИХАИЛ НИКОНОВ

НАУЧНЫЙ РУКОВОДИТЕЛЬ: АНДРЕЙ ПРЖИБЕЛЬСКИЙ

SPAdes

- Геномный ассемблер
- Граф де Брёйна, **k**-меры
- rnaSPAdes - сборка РНК

rnaQUAST

Инструмент для оценки качества сборки транскриптома.

Метрики:

- Database coverage
- Misassemblies
- 95%-covered isoforms
- 95%-covered genes
- ...

Задача

Посмотреть, как изменяется качество сборки в зависимости от k-меры.

- Перебор k-меры и организмов
- Представление результатов в удобном виде
- Анализ

Результаты

Python script:

SPAdes and rnaQUAST test results

For s_cerevisiae

K-mer size	39	43	47	49	51	53	55	57	59
Genes	7126	7126	7126	7126	7126	7126	7126	7126	7126
Avg. number of exons per isoform	1.06	1.06	1.06	1.06	1.06	1.06	1.06	1.06	1.06
Transcripts	11805	12891	12530	12111	11403	10443	10103	10159	10284
Transcripts > 500 bp	3590	3694	3820	3877	3959	4007	4041	4132	4205
Transcripts > 1000 bp	2947	3016	3114	3170	3218	3255	3291	3354	3392
Aligned	6744	7133	7799	8003	9791	9384	9261	9365	9545
Uniquely aligned	6434	6774	7407	7620	9264	8889	8789	8884	9064
Multiply aligned	251	285	305	301	451	423	398	391	384
Unaligned	5061	5758	4731	4108	1612	1059	842	794	739

Результаты

- Для разных организмов результаты могут очень сильно отличаться
- Есть корреляция со сложностью генома
- К-мера по умолчанию не оптимальная

Ссылки

- Скрипт: https://github.com/Karma-Police/SPbAU-5th-term-project
- Почта: nikonov.m.i@yandex.ru

СПб АУ РАН, 2017