Оценивание качества сборки геномных последовательностей

Алексей Гуревич, гр. 604 (SE)

Кафедра математических и информационных технологий

Научные руководители:
Вяххи Н.И.
Лесин В.М.

Рецензент:
Алексеев М.А.

Предметная область

Геном — длинные строки над алфавитом {A, C, G, T}.

Нуклеотид — буква в данном алфавите.

Секвенирование — процесс «чтения» генома, результат которого — фрагменты длиной $100 - 150$ символов.

Ассемблирование — процесс сборки генома по считанным фрагментам.

Контиг — последовательность нуклеотидов, являющаяся участком генома.

Референс — последовательность нуклеотидов генома, считающаяся достоверной для данного вида.

Постановка задачи

Цель

Создание научной базы и программного продукта, проводящего оценивание сборок.

Задачи

- Изучить существующие методы оценивания.
- Выделить ключевые метрики качества и добавить новые.
- Провести анализ существующих инструментов.
- Реализовать программный продукт, проводящий анализ по выбранным метрикам.

Источники метрик

- Assemblathon[1]
  - N50, NG50,
  - поиск генов и др.
- Описания существующих программ
  - метрики Mauve[2],
  - метрики Plantagora[3] и др.
- Разработки лаборатории
  - NA50, NGA50.

[1] Earl et al. "Assemblathon 1: A competitive assessment of de novo short read assembly method", Genome Research, 2011

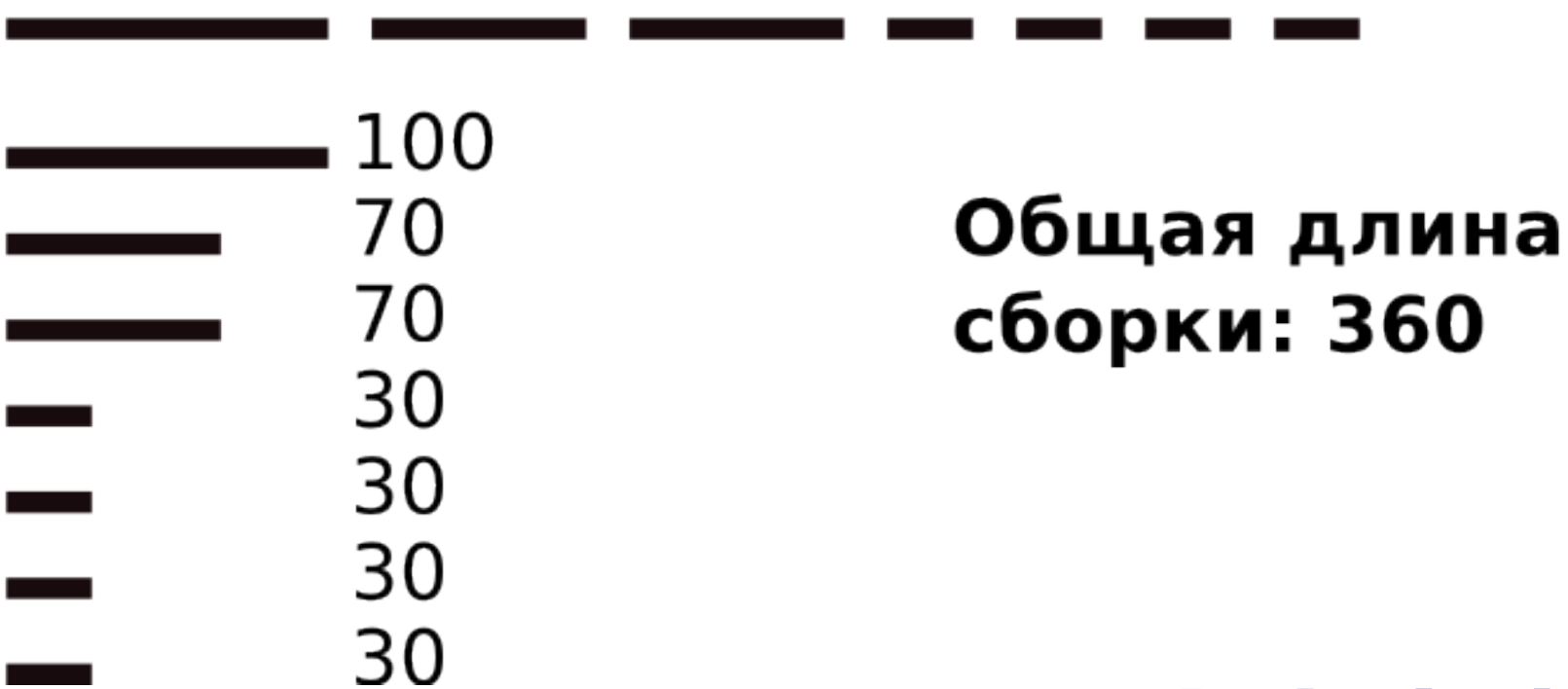
[2] Aaron E. Darling et al. "progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss, and Rearrangement", PLoS ONE, 2010

[3] Barthelson et al. "Plantagora: Modeling Whole Genome Sequencing and Assembly of Plant Genomes", PLoS ONE, 2011

Метрики: Nx

N50 (Nx)

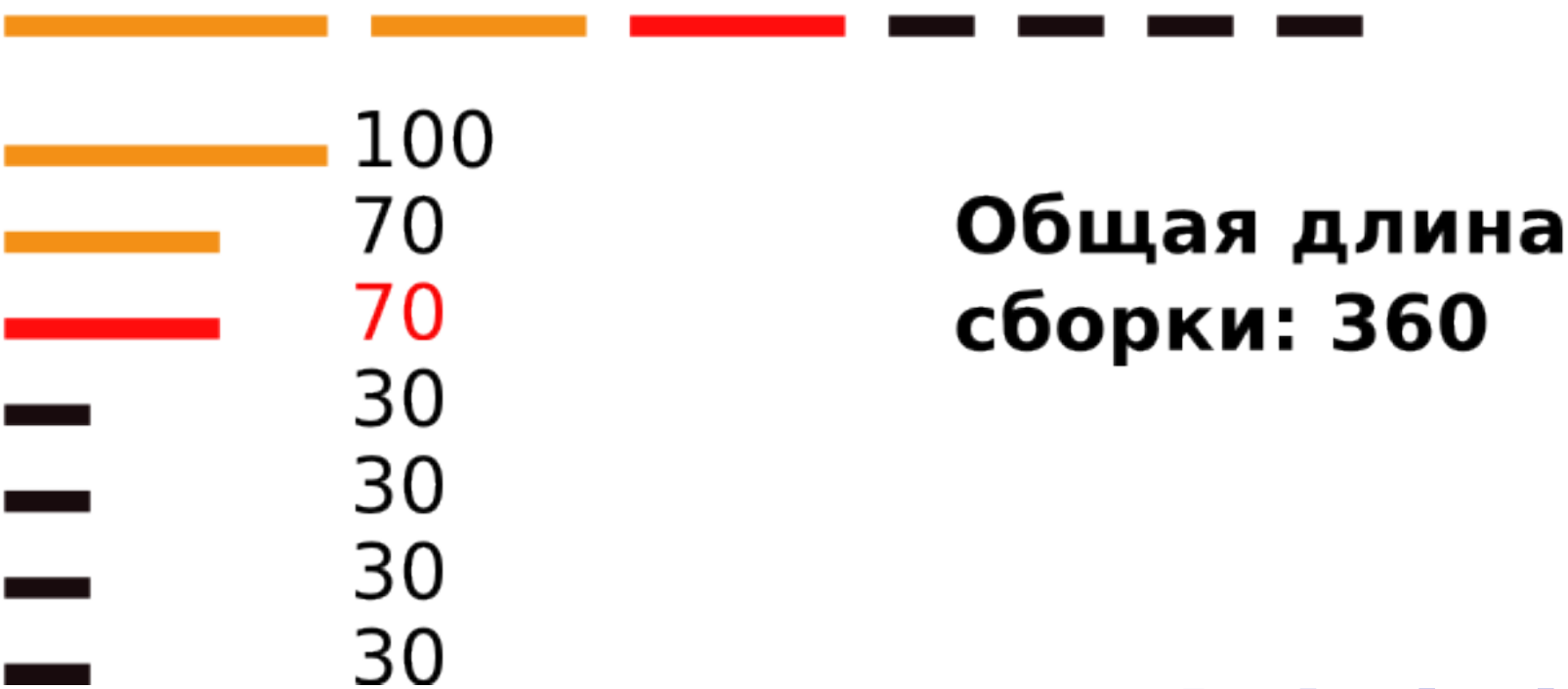
такая длина контига, что все контиги большей и равной длины составляют 50% ($x\%$) длины сборки.



Метрики: Nx

N50 (Nx)

такая длина контига, что все контиги большей и равной длины составляют 50% ($x\%$) длины сборки.



Метрики: Plantagora

Barthelson et al. “Plantagora: Modeling Whole Genome Sequencing and Assembly of Plant Genomes”, PLoS ONE, 2011

Метрики: NAx

Вычисление **NAx**:

1. Разбиваем контиги на блоки, имеющие выравнивание на референс.
2. Для блоков считаем Nx.

Необходимость собственного инструмента

- Существующие программы работают только с определенными входными данными.
- Возможность получать все необходимые метрики в одном приложении.
- Реализация своих метрик.

QUAST

- Приложение на Python, с использованием модулей на Java и Perl.
- Имеет иерархическую модульную структуру, содержащую:
  - 4 модуля на основе сторонних инструментов,
  - 9 модулей — собственная реализация.
- Поддерживает режимы оценивания:
  - при наличии референсного генома,
  - без него.
- Автоматическая генерация:
  - графиков,
  - текстовых отчетов,
  - сводных таблиц.

QUAST

Режим работы **при наличии референса**:

- Подсчет 49 метрик.
- Построение 12 графиков.

Режим работы **при отсутствии референса**:

- Подсчет 5 стандартных метрик.
- Анализ потенциальных генов:
  - поиск открытых рамок считывания (ORF),
  - использование GeneMark[4].
- Совместный анализ двух сборок на мисассемблы.

[4] Ter-Hovhannisyan V. et al. "Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training", Genome Research, 2008

Результаты

Реализованное приложение:

- Активно используется для разработке и отладке SPAdes.
- Использовано при написании статьи о SPAdes в Journal of computational biology[5].
- используется в совместных проектах с NIH, JCVI, JGI, Yale University.
- входит в состав геномного ассемблера SPAdes версии 2.1.*.

[5] Bankevich A., Nurk S. et al. "SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing", Journal of computational biology, 2012

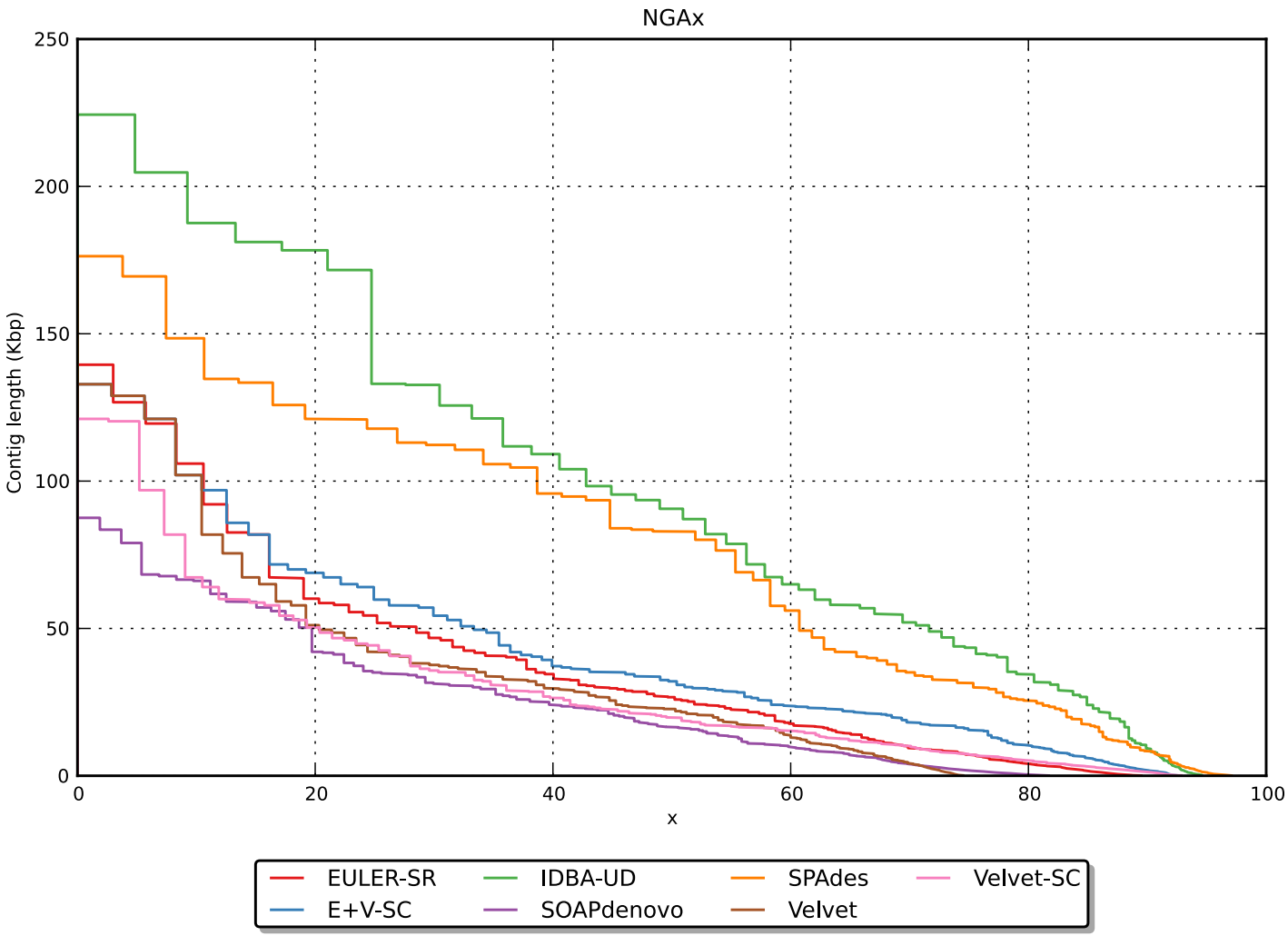
Примеры вывода

Сводная таблица

Assembler	# contigs	NG50 (bp)	Largest (bp)	Total (bp)	Mapped genome (%)	Misassemblies	Complete genes
E+V-SC	501	32051	132865	4570583	93.83	**2**	3809
EULER-SR	1344	26662	140518	4369634	87.87	17	3457
IDBA-UD	**283**	**90607**	**224018**	4734432	95.82	7	4026
SOAPdenovo	1240	18468	87533	4237595	82.51	10	3059
SPAdes	950	82949	176326	4980646	**96.94**	6	**4048**
Velvet	428	22648	132865	3533351	75.80	**2**	3117
Velvet-SC	872	19791	121367	4589603	93.81	**2**	3654

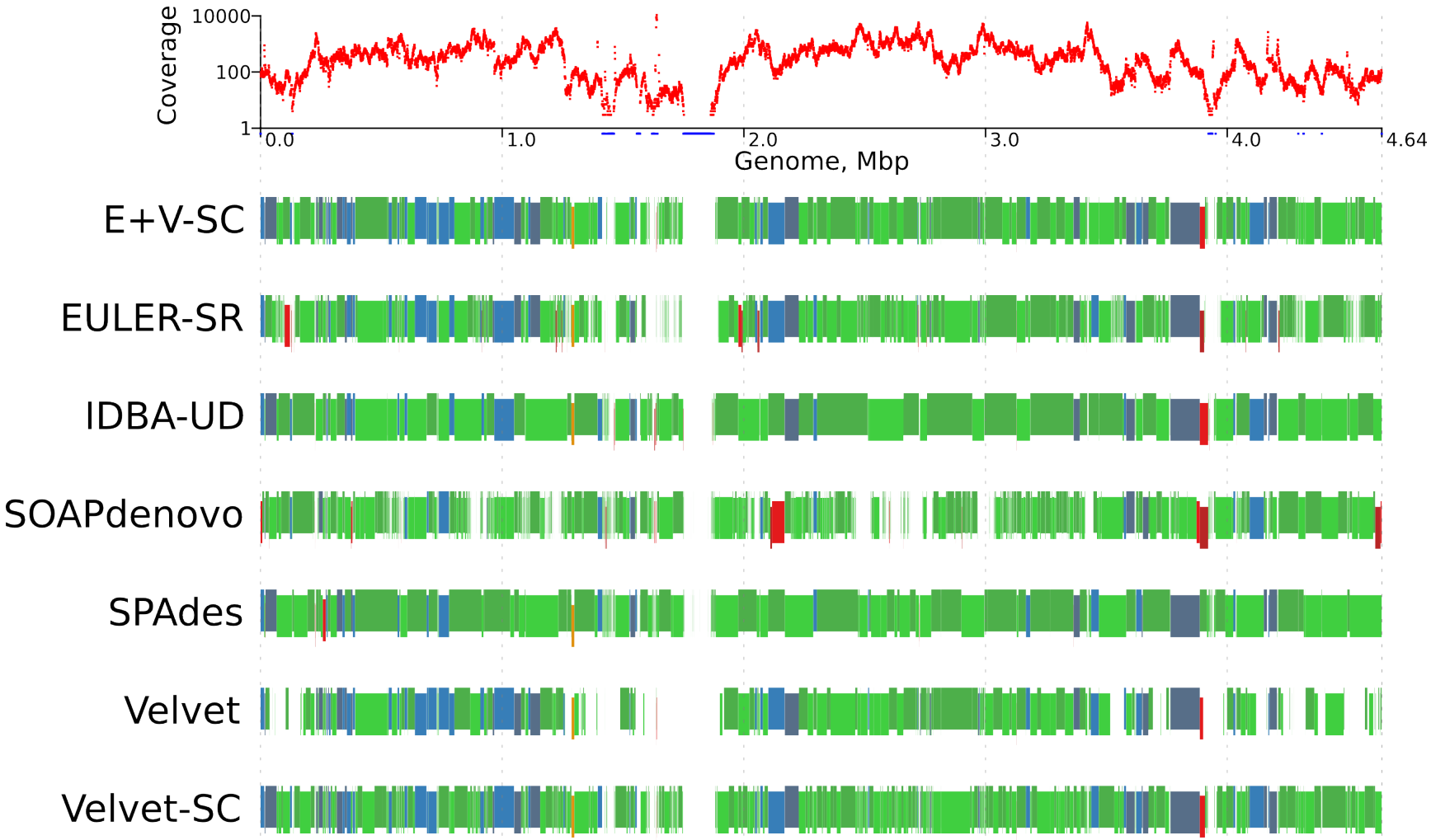
Примеры вывода

График зависимости NGAx от x



Примеры вывода

Сравнение контигов



Благодарности

Благодарности

Спасибо за внимание!