

Методы и алгоритмы автоматизированной сборки генома на вычислительном кластере

Романенков Кирилл Владимирович

Специальность 05.13.11

Научные руководители:

чл.-корр. РАН, д.ф.-м.н. Королев Л.Н.

чл.-корр. РАН, д.ф.-м.н. Якобовский М.В.

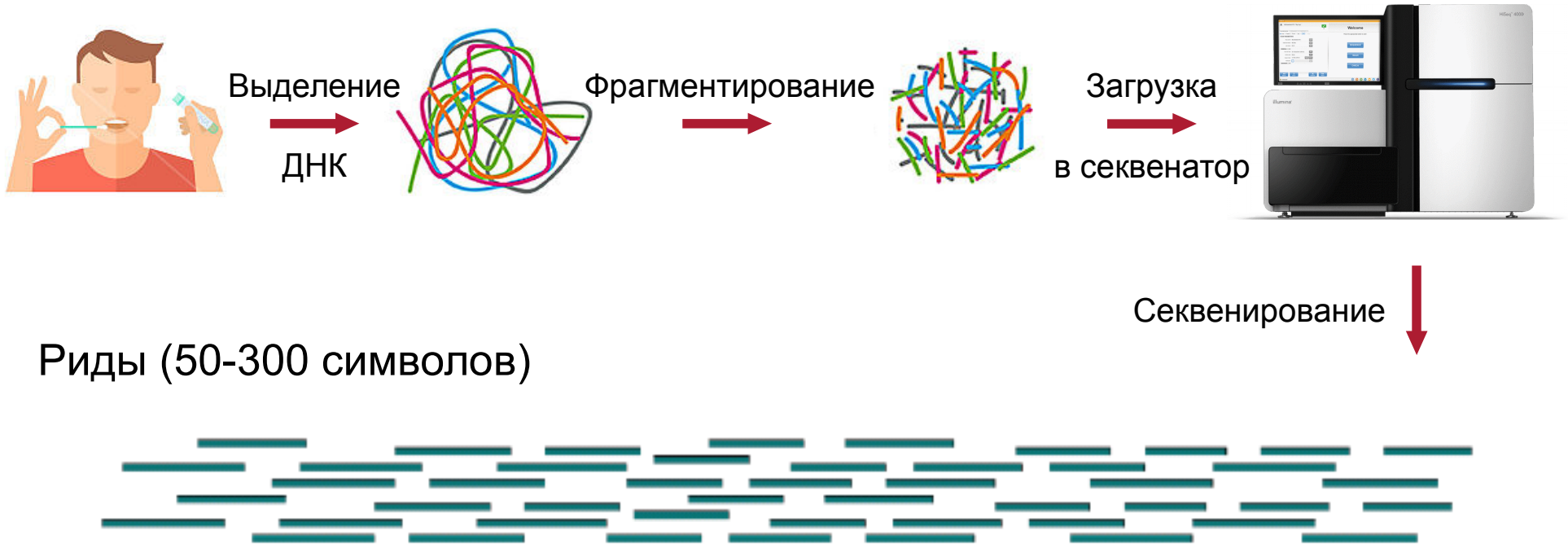
Сборка генома

Геном – это совокупность всех последовательностей ДНК организма,
геном человека $\sim 3 \cdot 10^9$ символов



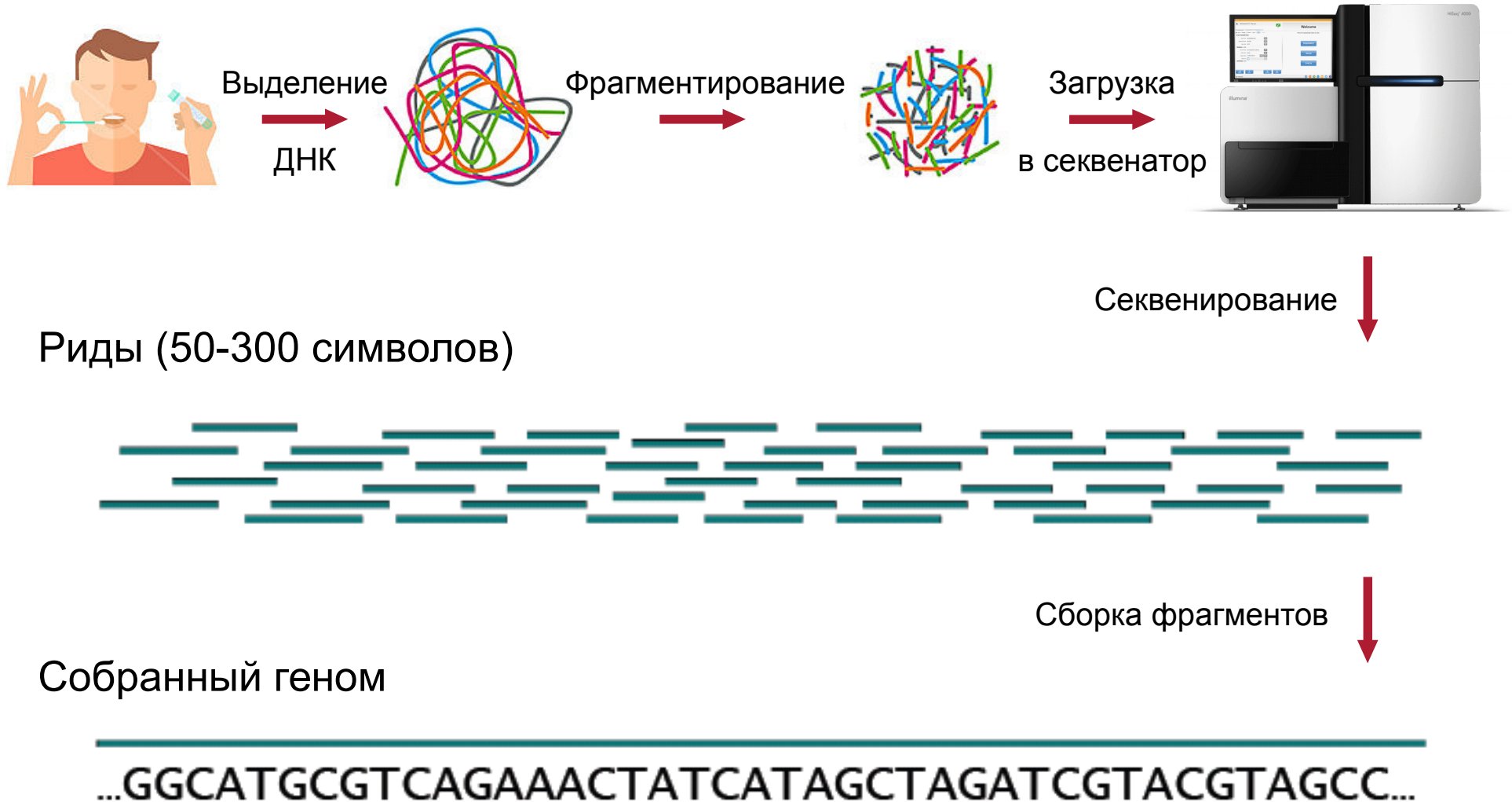
Сборка генома

Геном – это совокупность всех последовательностей ДНК организма,
геном человека $\sim 3 \cdot 10^9$ символов



Сборка генома

Геном – это совокупность всех последовательностей ДНК организма,
геном человека $\sim 3 \cdot 10^9$ символов



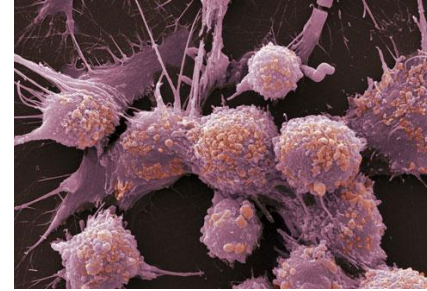
Сборка генома. Актуальность проблемы



Молекулярная
биология



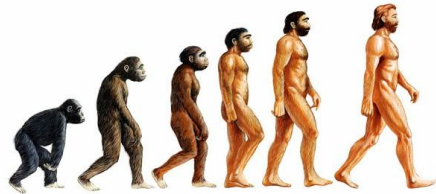
Поиск новых
лекарств



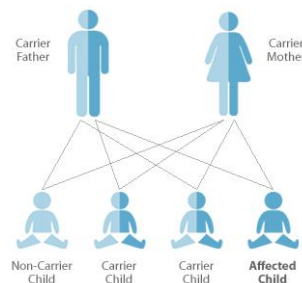
Лечение
рака



Экология



Теория эволюции



Генетические
заболевания



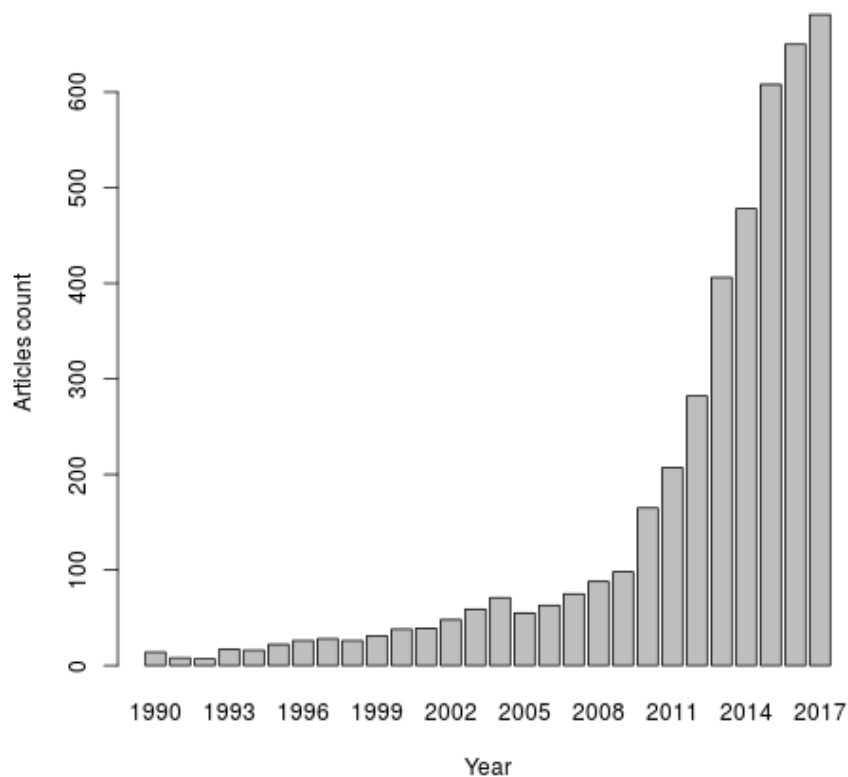
Популяционная
геномика



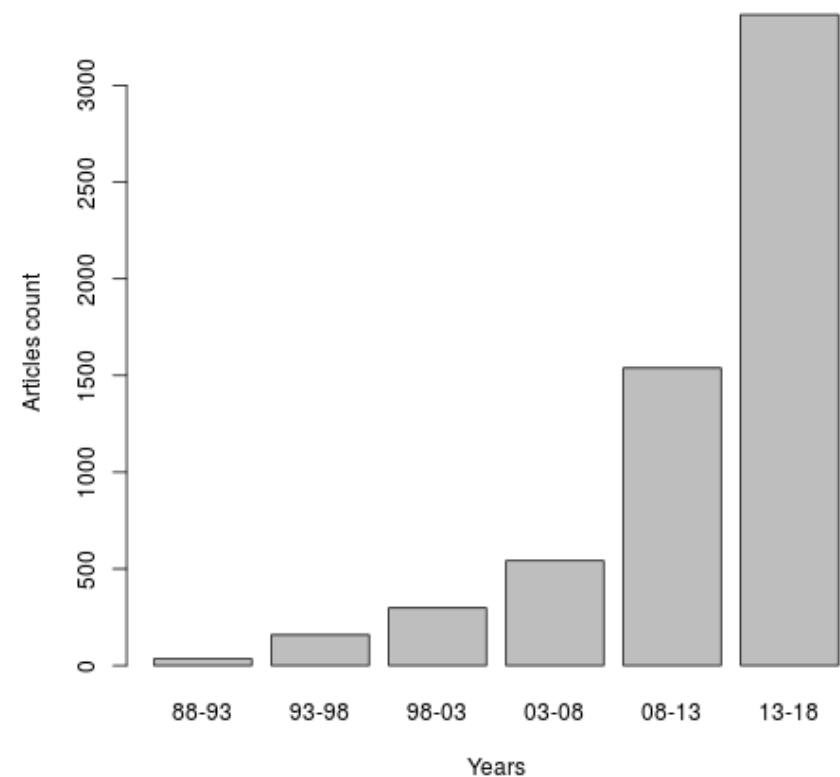
Палеогенетика

Сборка генома. Актуальность проблемы

- Большинство биоинформатических проектов начинаются со сборки генома
- Количество публикаций, содержащих ключевые слова "de novo assembly", стабильно растет



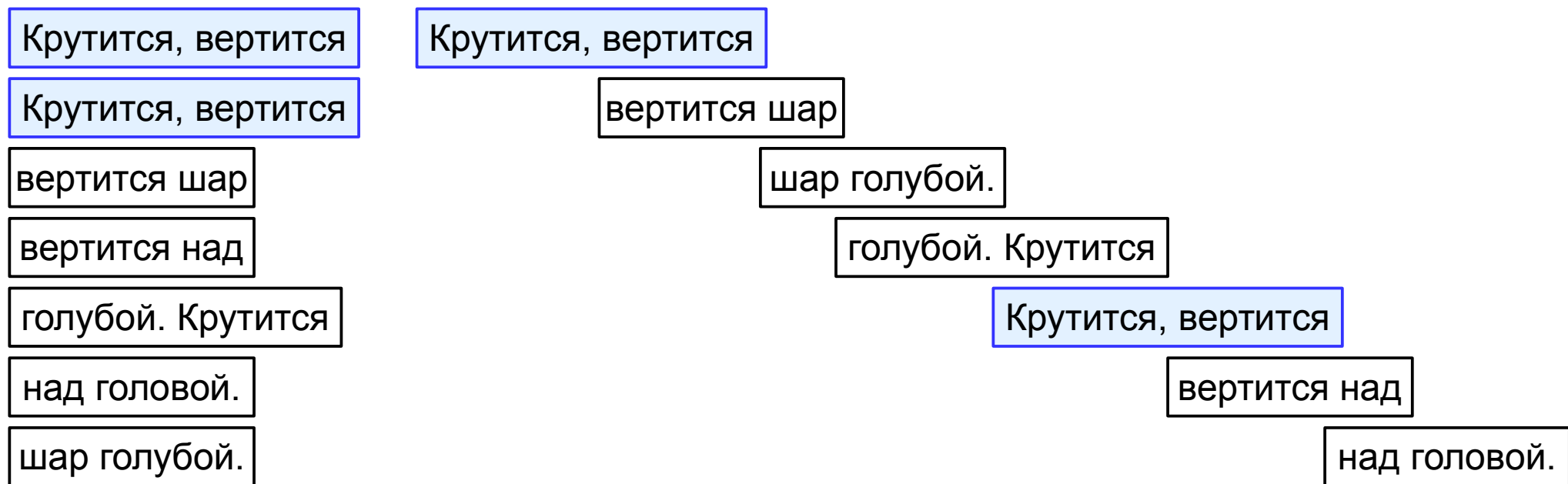
Pubmed



Web of Science

Что такое сборка генома?

Крутится, вертится	шар голубой.	Крутится, вертится	над головой.	
Крутится,	вертится шар	голубой. Крутится,	вертится над	головой.



- Задача сборки может быть формализована как задача поиска наименьшей общей надстроки
- Прочитываются фрагменты нескольких копий строки, фрагменты каждой копии смешиваются
- Некоторые фрагменты могут быть одинаковыми

Сложности задачи сборки генома de novo

- **Биологический аспект.** Гетерозиготность, повторы, загрязнение другими ДНК
- **Секвенирование.** Наличие ошибок секвенирования, вероятность ошибки растет вместе с длиной фрагмента
- **Вычислительный аспект.** Большие требования к памяти (сотни Гб для больших геномов)
- **Точность.** Нет четкого определения корректности сборки

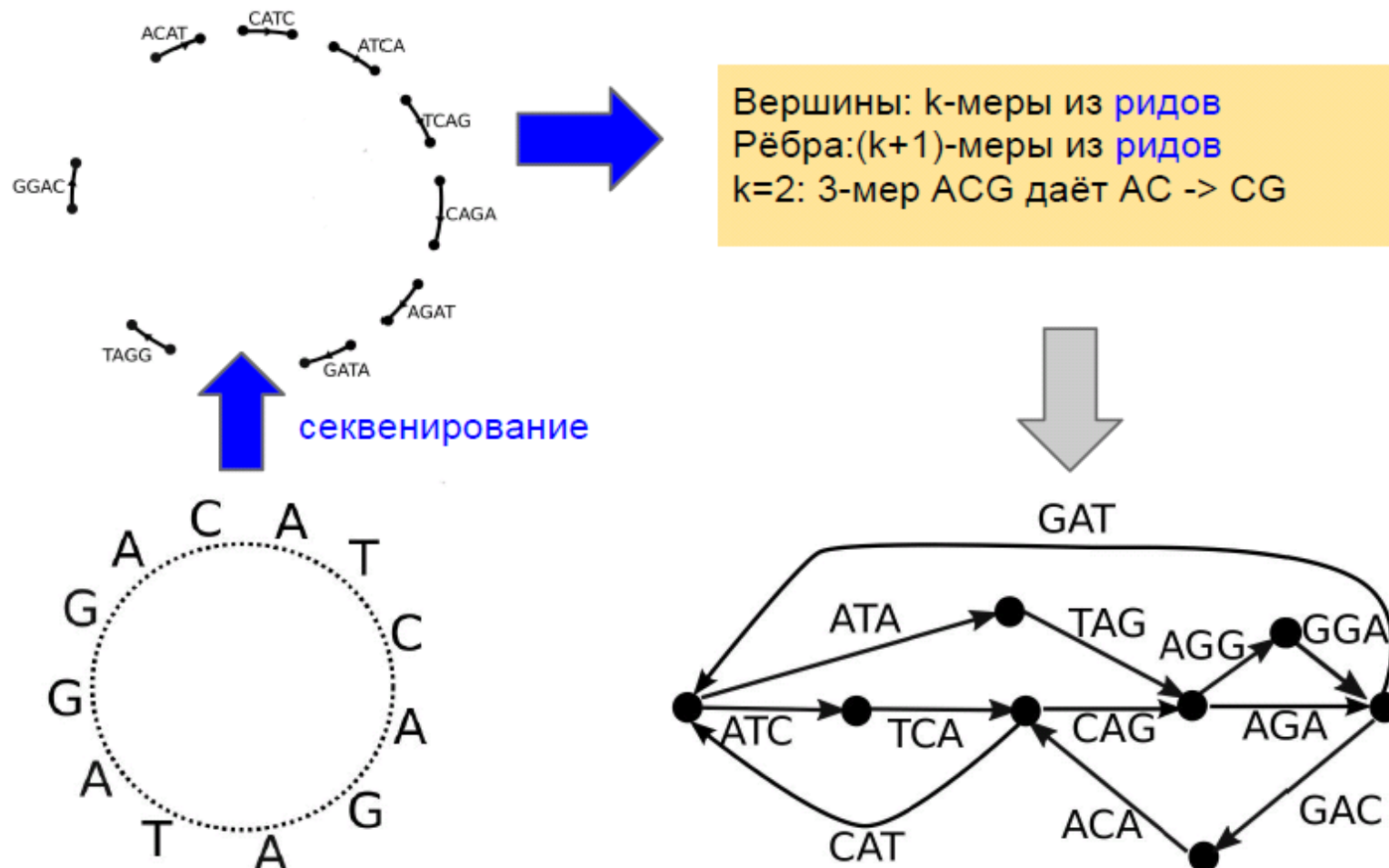
Актуальность задачи

EULER-SR WhatsHap GARM SOAPdenovo SSPACE HaploMerger mip
A5 Telescope Contrail fermi GABenchToB QSRA Opera RAMPART Arapan
SSAKE SWAP-Assembler Newbler AutoAssemblyD HapCompass Dazzler PCAP Forge SHEAR
TIGR Mapsembler 2 ALLPATHS-LG VICUNA Edena CLC PERGA KmerGenie
gapfiller CloudBrush REAPR Ray Tedna TIGRA Amos
MIRA dipSPAdes MetAMOS Geneious SeqMan NGen Nesoni ATAC
Arachne Celera VCAKE GAM PASHA bambus2 MetaVelvet-SL Quast
SCARPA IDBA Trinity MHAP Hapsembler GAML Sequencher BESST GGAKE
GRIT Phrap MaSuRCA HiPGA PADENA SeqPrep Phusion SGA PE-Assembler
CGAL Curtain SWiPS KGBAssembler Metassembler HGAP PRICE Pilon MSR-CA
Pipeline Pilot SHRAP Taipan SILP3 IDBA-MTP SR-ASM Velvet Enly
OMACC Anchor Omega SUTTA ABySS HyDA-Vista Atlas FRCBam
SOPRA iMetAMOS DNAnexus Ragout SPAdes SAT-Assembler image
DNA Dragon CABOG SAGE Cerulean Monument ngsShoRT ABBA
FALCON SuccinctAssembly SHARC GS GAGM Khmer GenoMiner ELOPER
GigAssembler Lasergene PBJelly DecGPU

- На одних и тех же данных дают совсем разные ответы – наборы контигов (длинные фрагменты генома)
- Основная алгоритмическая концепция – граф де Брюйна
- Причины разных ответов: эвристические алгоритмы упрощения графа и фильтрации ридов

Граф де Брюйна

- Риды разбиваются на k -меры, k -мер: последовательность из k символов
- Вершины графа де Брюйна: все k -меры, ребра графа де Брюйна: все $(k+1)$ -меры, ребро соединяет префикс и суффикс $(k+1)$ -мера
- Геном – Эйлеров путь в графе де Брюйна



Постановка задачи

- Провести анализ существующих программ по сборке генома и методов объединения их результатов
- Провести анализ существующих методов выбора параметра k для сборщиков, основанных на графе де Брюйна
- Разработать метод оценивания качества геномной сборки при отсутствии референса и выбора оптимального параметра k
- Разработать алгоритм объединения результатов работ различных геномных сборщиков при отсутствии референса

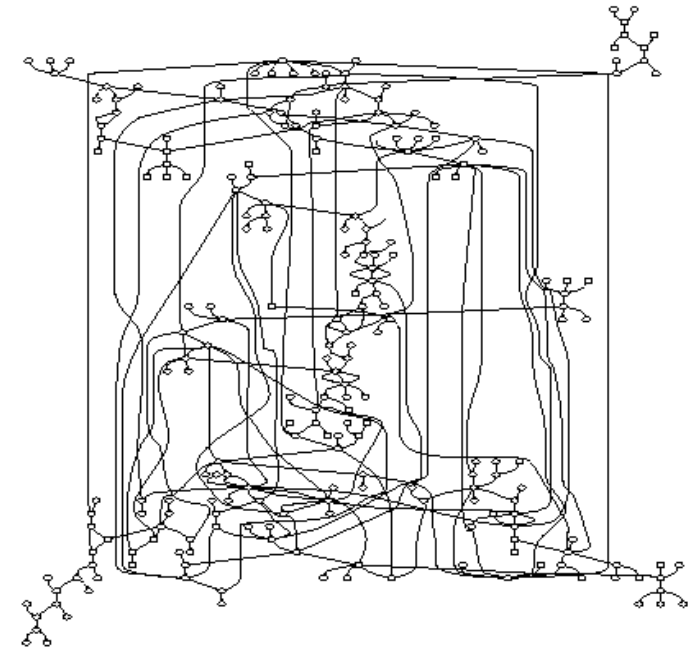
Выбранные геномные сборщики

- A5, Abyss, IDBA, Ray, SOAPdenovo, SPAdes, Velvet



- Фигурируют в статьях, посвященных сравнительному анализу качества получаемых сборок

- Используют концепцию графа де Брюйна



- Open source

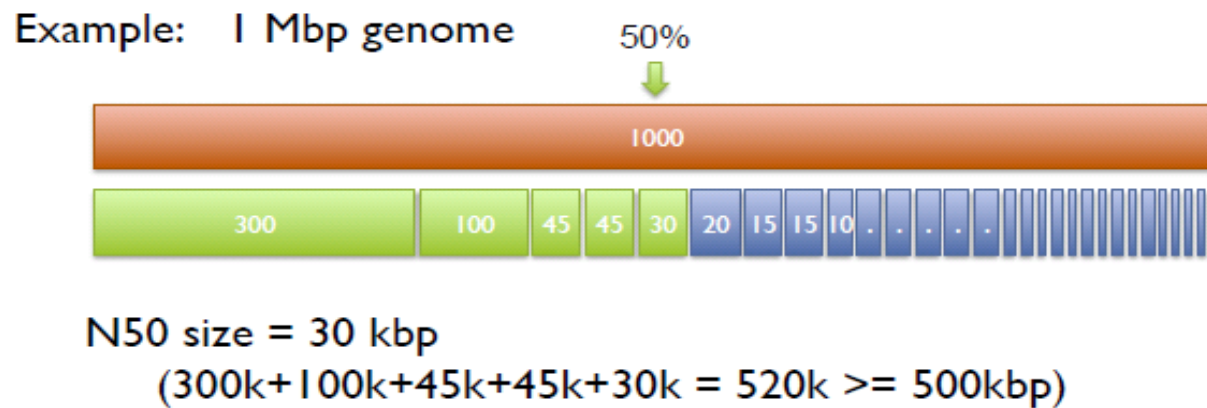


Наборы данных для тестирования

Организм	Encephalitozoon cuniculi	Candida parapsilosis	Saccharomyces cerevisiae S288C
Общая длина	2,497,519	13,078,488	12,157,105
Технология	Illumina HiSeq 2000 paired end	Illumina Genome Analyzer IIx	Illumina MiSeq
GC	47,31	38,68	38,15
Число последовательностей	11 хромосом	8 контигов	17 хромосом
N50	218,329	2,091,826	924,431

Существующие метрики оценивания качества геномной сборки

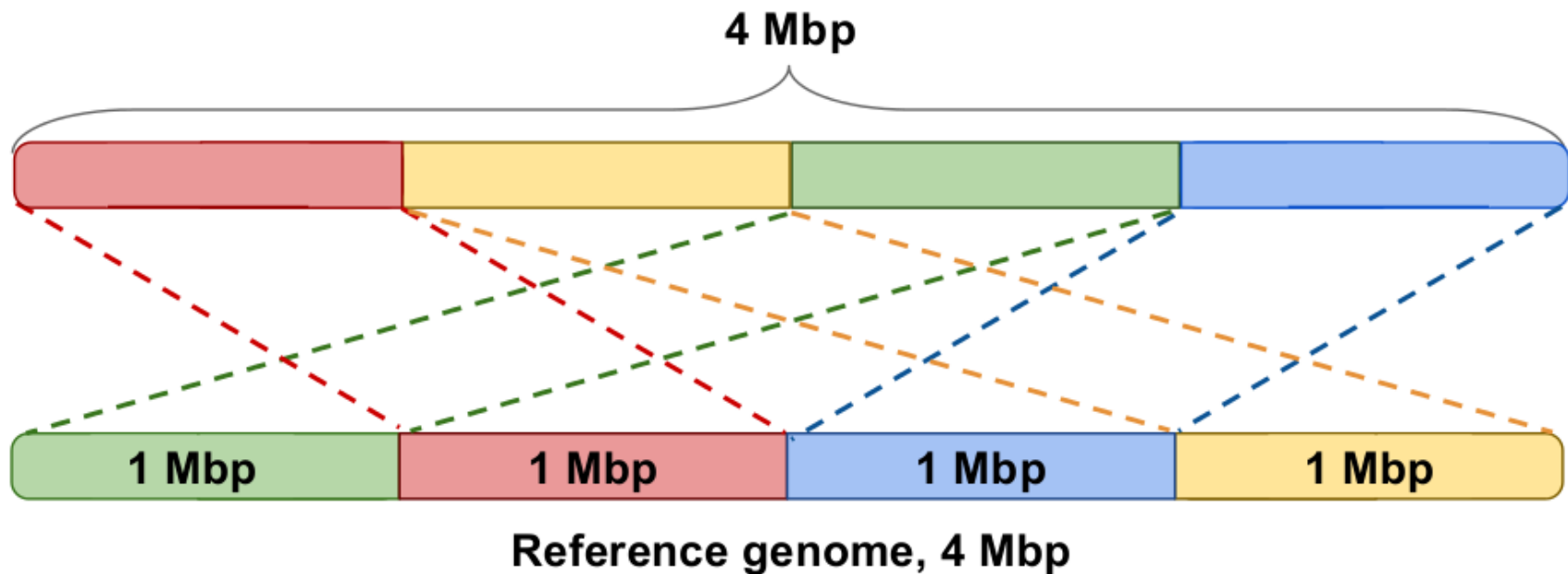
- Безреференсные метрики
 - Число контигов
 - Суммарная длина контигов
 - N50. Максимально возможная длина контига, такая что контиги длиннее этого значения составляют по крайней мере половину сборки.



- L50. Минимальное число контигов, составляющих по крайней мере половину сборки. В вышеприведенном примере L50 = 5.

Существующие метрики оценивания качества геномной сборки

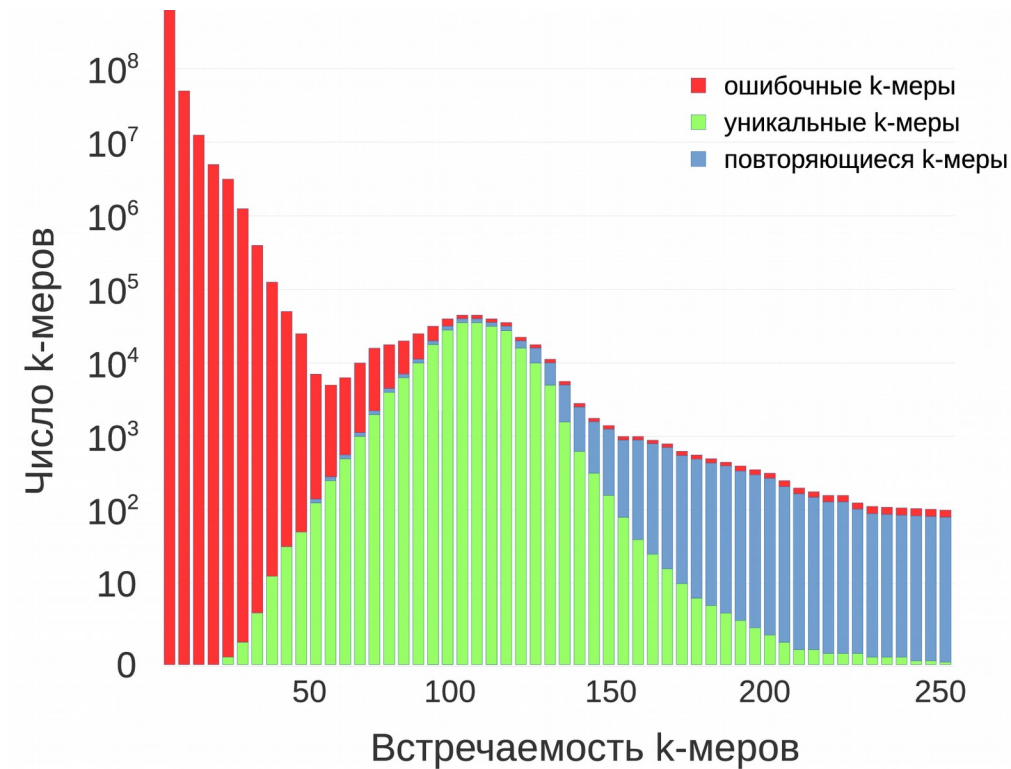
- Метрики, использующие референсный геном
 - Ошибки сборки
 - NGA50. «Скорректированный» N50



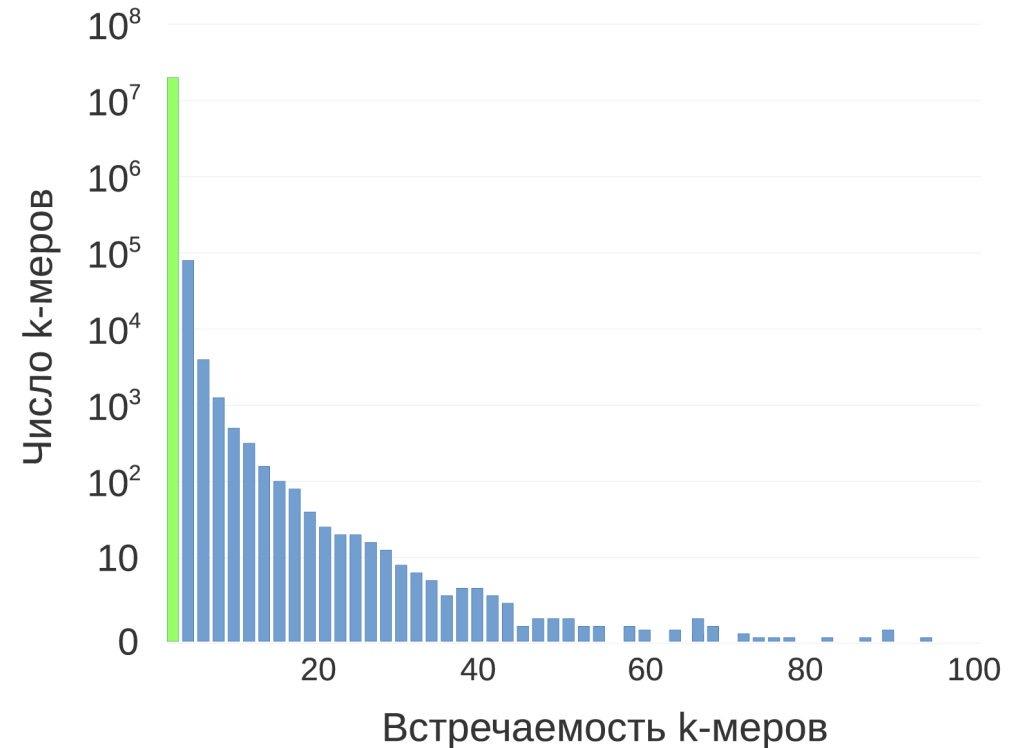
- Процент собранного генома
- Показатель дублирования

Предложенный метод оценивания качества геномной сборки при отсутствии образца

Основная идея предложенного метода заключается в установлении соответствия между уникальными k-мерами в собранном геноме и k-мерами в ридх.



Встречаемость 21-меров в ридх
Encephalitozoon cuniculi



Встречаемость 21-меров в собранном геноме
Encephalitozoon cuniculi

Предложенный метод оценивания качества геномной сборки при отсутствии образца

- Предлагается вычислять долю различных k-меров, взятых из некоторой окрестности второго пика на гистограмме встречаемости k-меров в чтениях, среди уникальных k-меров для собранного генома по следующей формуле:

$$Q = \frac{\sum_{i=1}^{|K_{uniq_reads}|} [k^r(i) \in K_1^g]}{|K_{uniq_reads}|},$$

$$K_1^g = \{k^g : k^g \in K^g, abundance(k^g) = 1\}$$

$[a_{p_l}; a_{p_r}]$ - окрестность из второго пика на гистограмме чтений

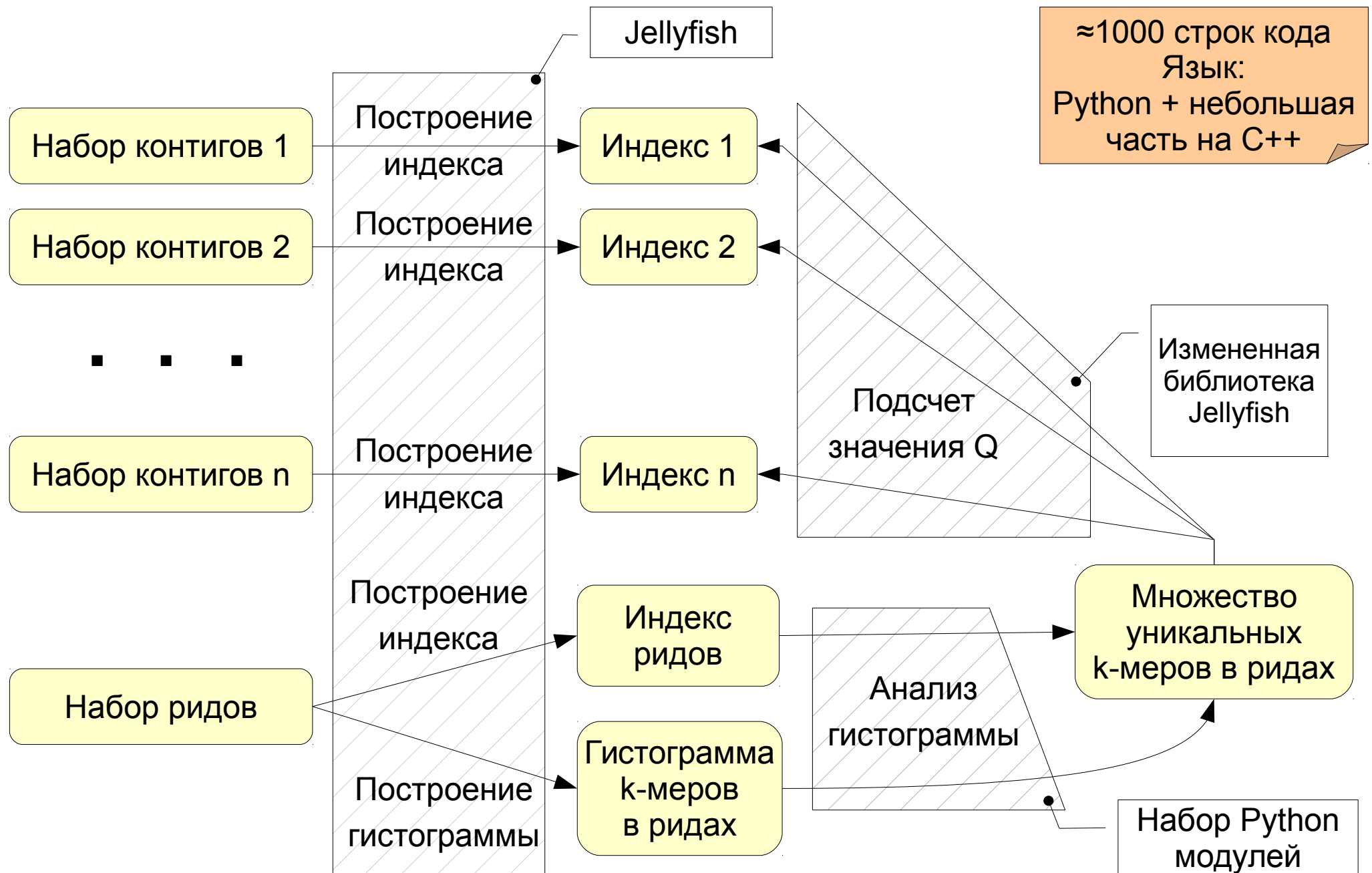
$$K_i^r = \{k^r : k^r \in K^r, abundance(k^r) = i\}$$

$abundance(k)$ – частота встречаемости k-мера k в множестве k-меров

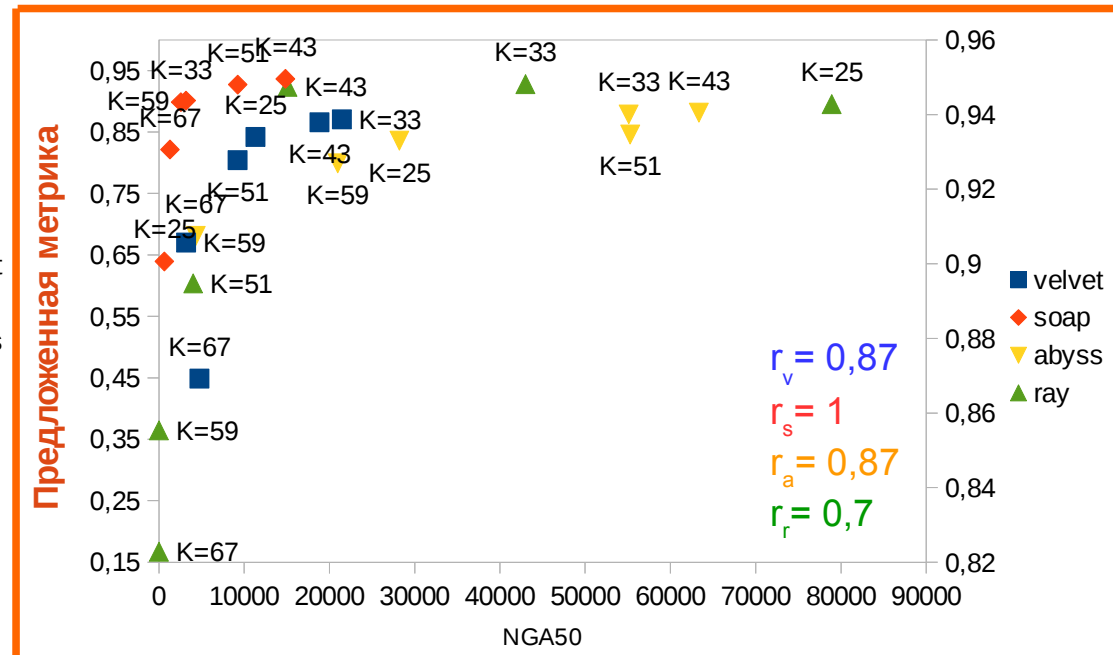
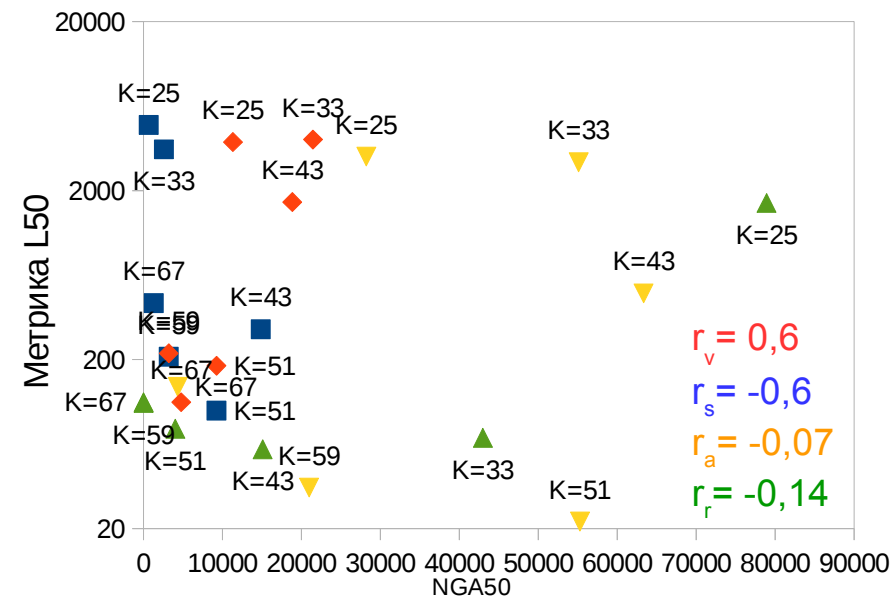
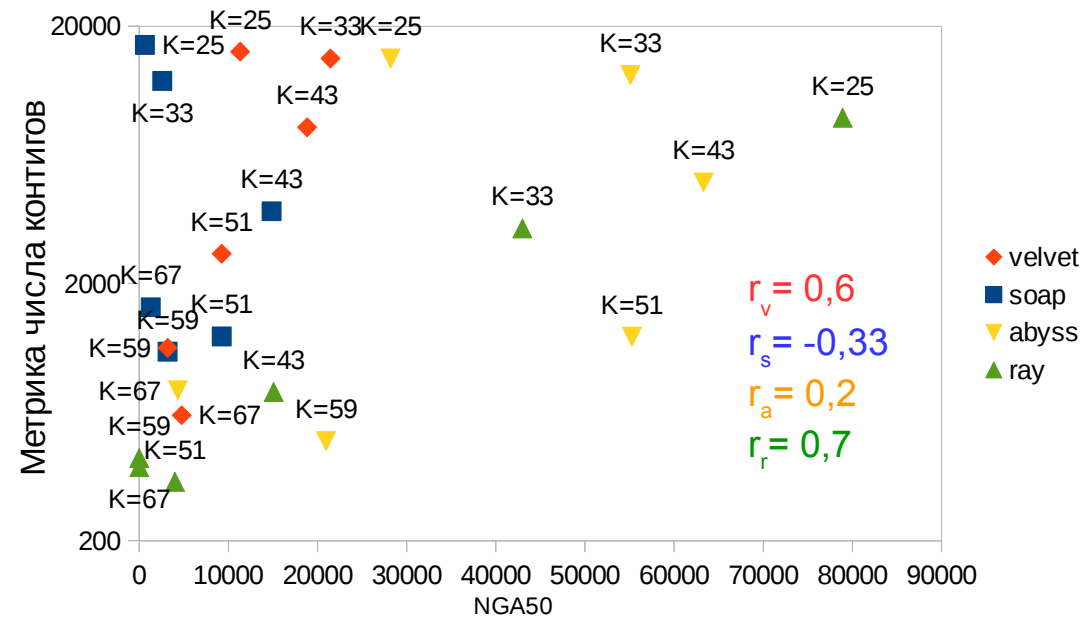
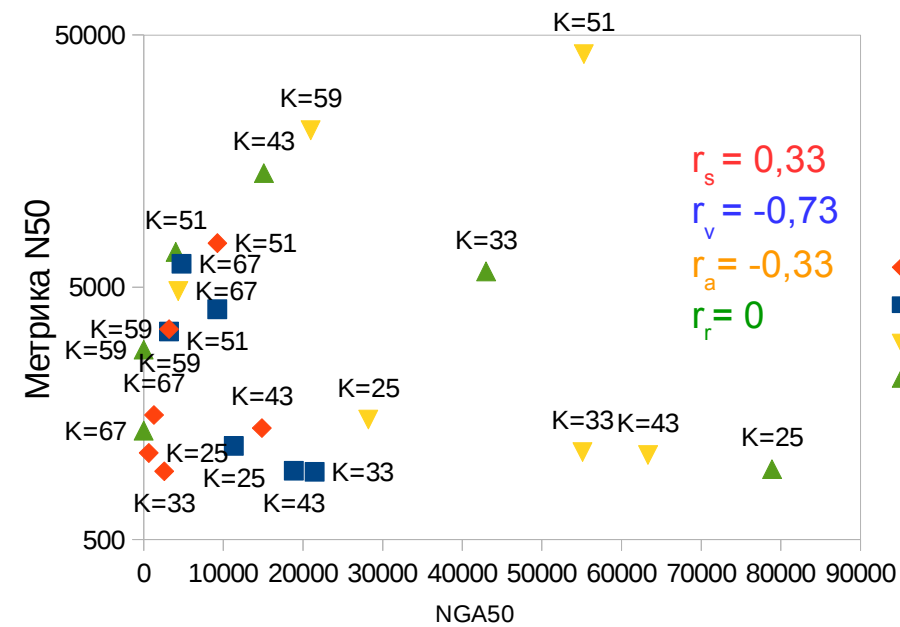
$$K_{uniq_reads} = \bigcup_{i=a_{p_l}}^{a_{p_r}} K_i^r$$

K^r – множество всех k-меров ридов, K^g – множество всех k-меров собранного генома из контигов длиннее порогового значения

Схема взаимодействия программных модулей



Проверка предложенного метода оценивания качества геномной сборки при отсутствии образца



Объединение результатов работ геномных сборщиков. Существующие решения

- CISA , MIX



- Слабomasштабируемы: последовательные программы, не применимы для геномов больших размеров
- Ресурсоемкость: на ряде входных данных время работы может достигать нескольких суток.
- Качество итоговой сборки может оказаться хуже чем у отдельно взятой сборки, участвующей в объединении.

Формальная постановка задачи

- Пусть даны k множеств слов: $C_1, C_2, \dots, C_k$:

$$\forall c_i \in C_i \exists c_j \in C_j : o(c_i, c_j) > 0, i \neq j \quad \forall c_i \in C_i \nexists c_j \in C_j : isSubstr(c_i, c_j), i \neq j$$

- На основе этих слов строится граф перекрытий $G(V, E)$:

$$V = \bigcup_{i=1}^k C_i, E = \{ v_i \rightarrow v_j : o(v_i, v_j) = p, p > 0;$$

$$v_i[|v_i| - p; |v_i|] = v_j[1; p], v_i \in C_n, v_j \in C_m, n \neq m \}$$

- Пусть также дана функция $Err : E \rightarrow \{0, 1\}$
- Необходимо построить минимальное покрытие графа G независимыми путями, не содержащими ошибочных ребер:

$$P = \{ P_i \}_{i=1}^n, \text{ где } P_i = \{ v_{i1}, v_{i2}, \dots, v_{il} \} : P_i \cap P_j = \emptyset, i \neq j \text{ и}$$

$$\left\{ \begin{array}{l} \forall (v_{ij}; v_{ij+1}) \in P_i \quad Err(v_{ij}; v_{ij+1}) = 0; \\ \forall v \in V \exists P_i \in P : v \in P_i; \\ |P| \rightarrow \min \end{array} \right.$$

$$\left\{ \begin{array}{l} \forall v \in V \exists P_i \in P : v \in P_i; \\ |P| \rightarrow \min \end{array} \right.$$

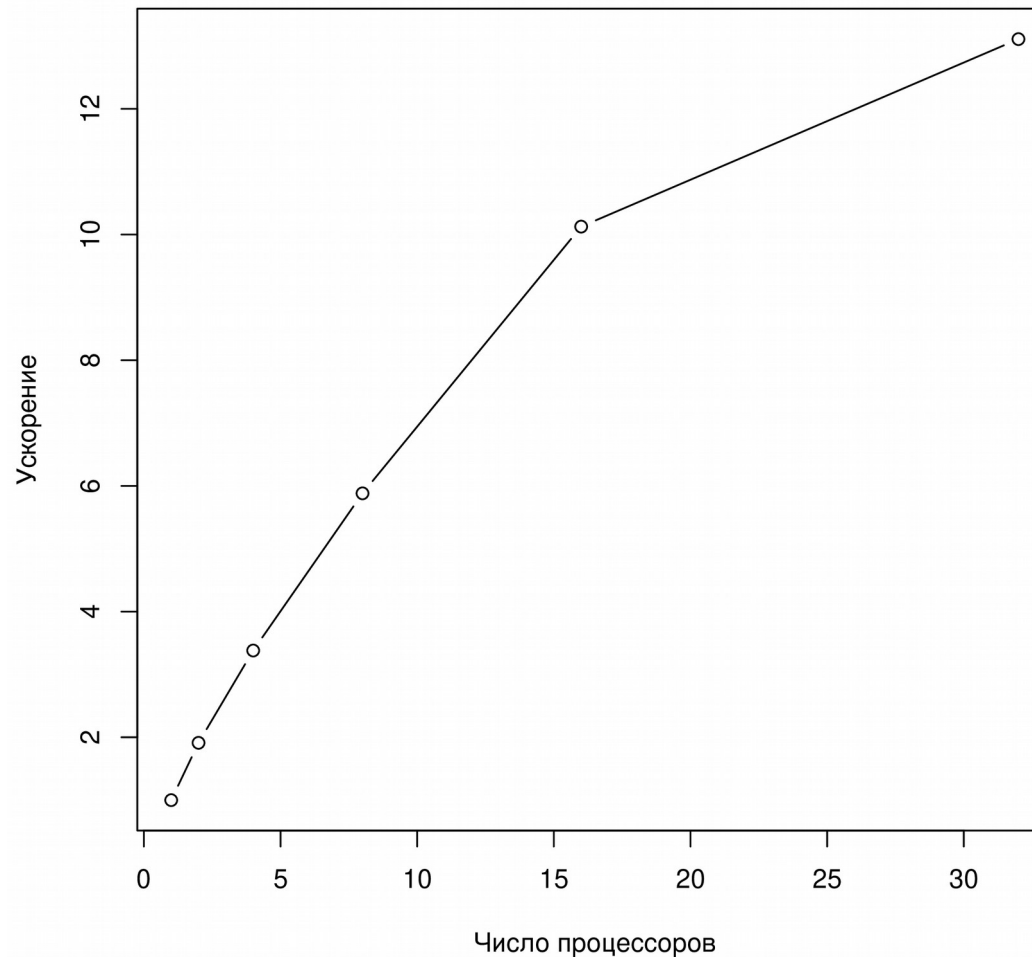
$$|P| \rightarrow \min$$

Объединение результатов работ геномных сборщиков. Оценка сложности алгоритма

- Пусть в общей сложности имеется n слов, с максимальной длиной L
- Попарное выравнивание слов программой Megablast – $O(n*L)$
- Построение графа перекрытий – $O(|E|)$, где E - множество ребер в графе, заданном матрицей попарного выравнивания, $|E| \sim n$
- Кластеризация графа LabelRank алгоритмом – $O(|E|)$
- Объединение слов в каждом кластере жадным алгоритмом - $O(|E|)$
- Корректирование объединенных контигов - $O(|E|)$
- **Общая сложность** – $O(n*L) + O(|E|)$

Параллельная реализация

- Распределение контигов по процессорам, сбор статистики
- Каждый MPI-процесс производит попарное выравнивание программой MEGABLAST множества своих контигов
- Полученные разреженные матрицы отсылаются мастер-процессу

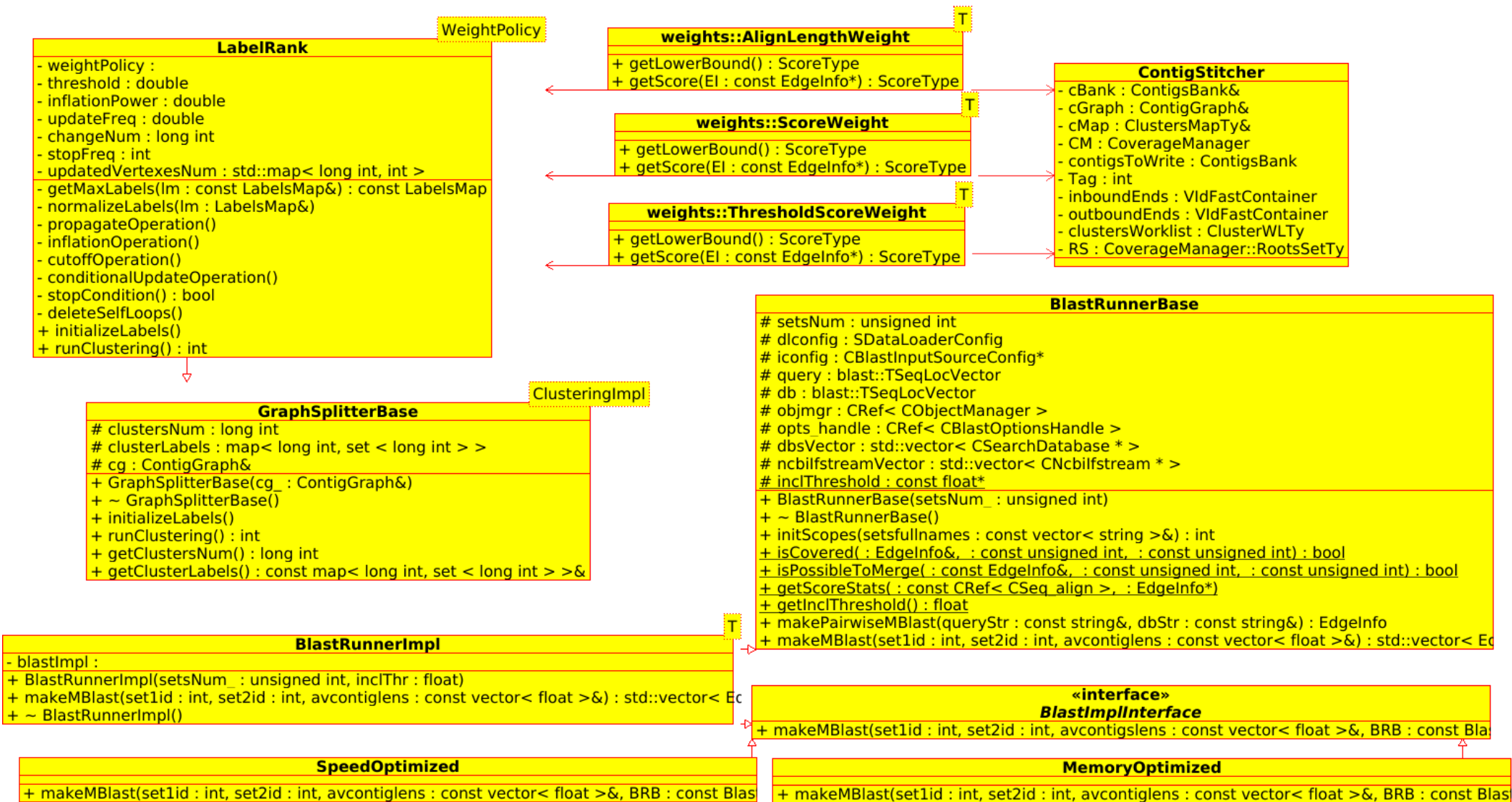


≈4500 строк кода;
язык: C++ плюс небольшая
часть на Python



UML-диаграмма классов

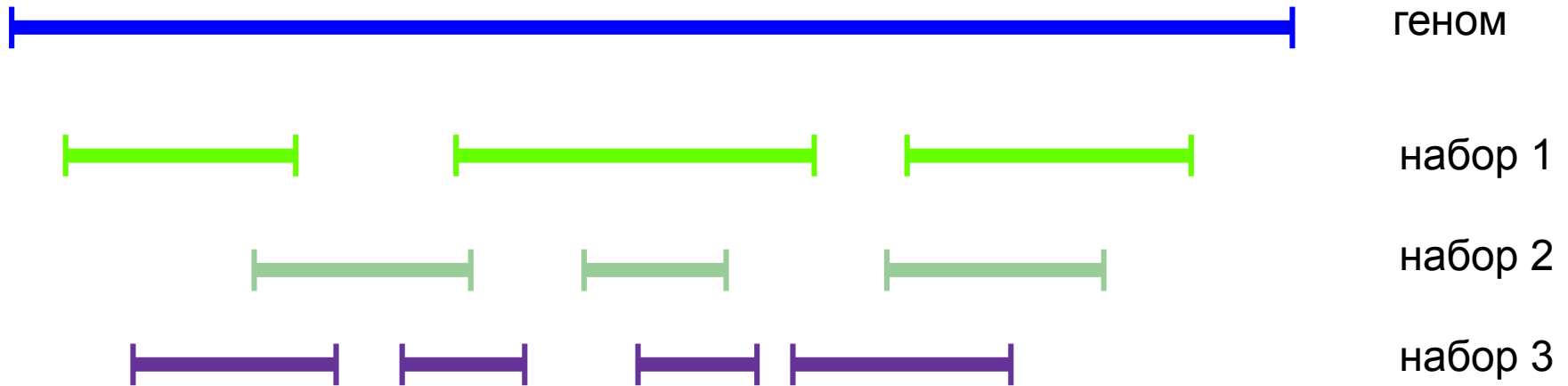
≈4500 строк кода;
язык: C++ плюс небольшая
часть на Python



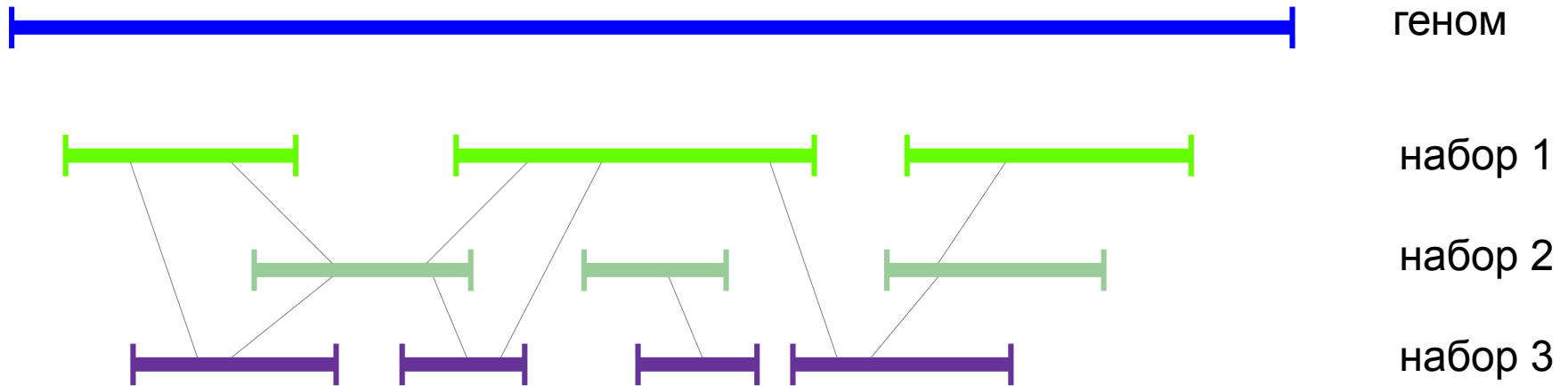
Объединение результатов работ геномных сборщиков. Предлагаемый метод

- Накладываемые ограничения:
 - Собираемый геном содержит один набор хромосом
 - В коротких фрагментах содержится геном только одного организма
 - Не производится коррекция ошибок в исходных наборах контигов

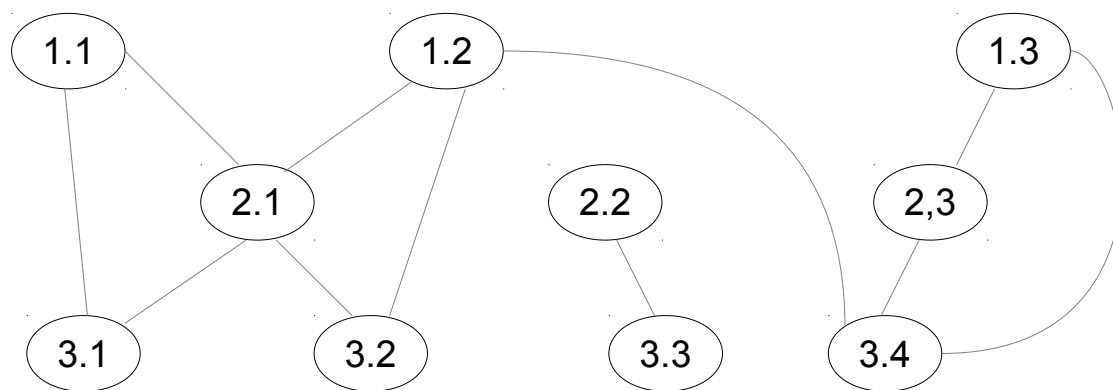
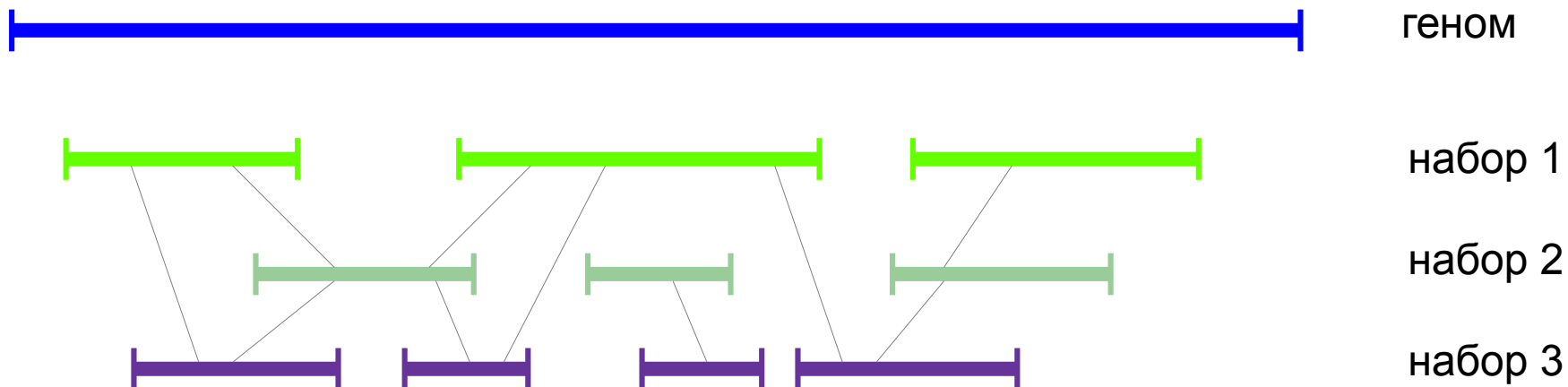
Объединение результатов работ геномных сборщиков. Пример



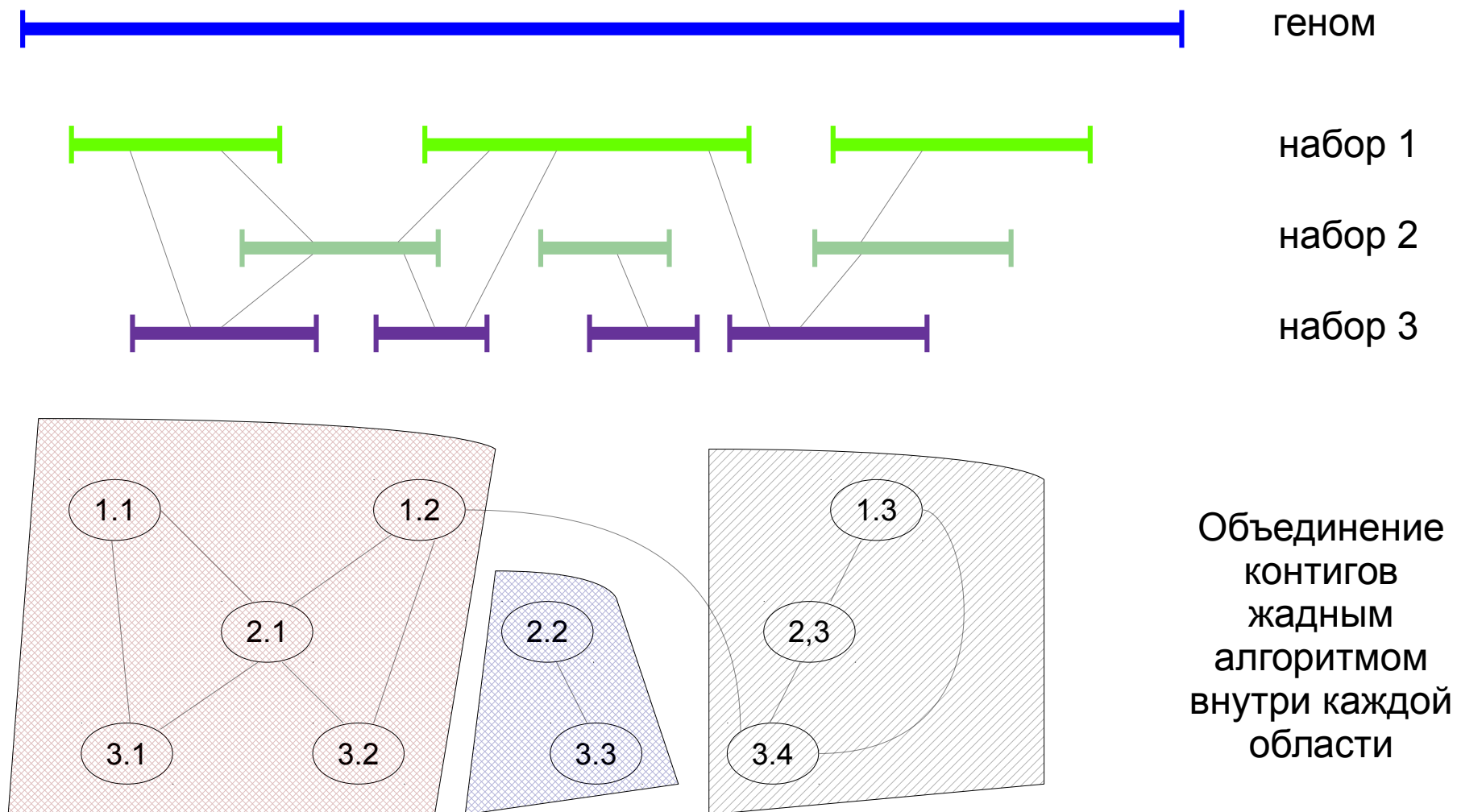
Объединение результатов работ геномных сборщиков. Пример



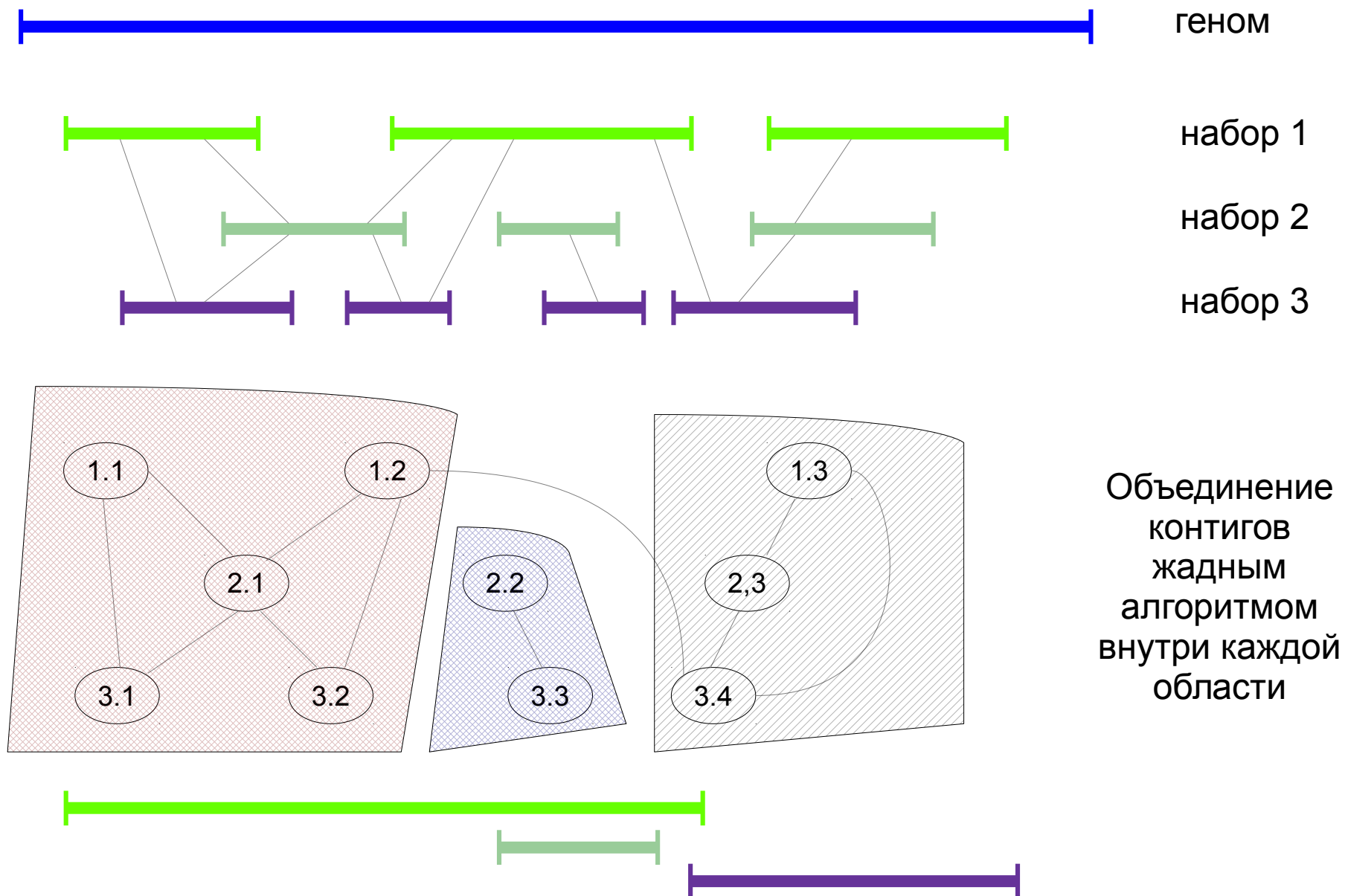
Объединение результатов работ геномных сборщиков. Пример



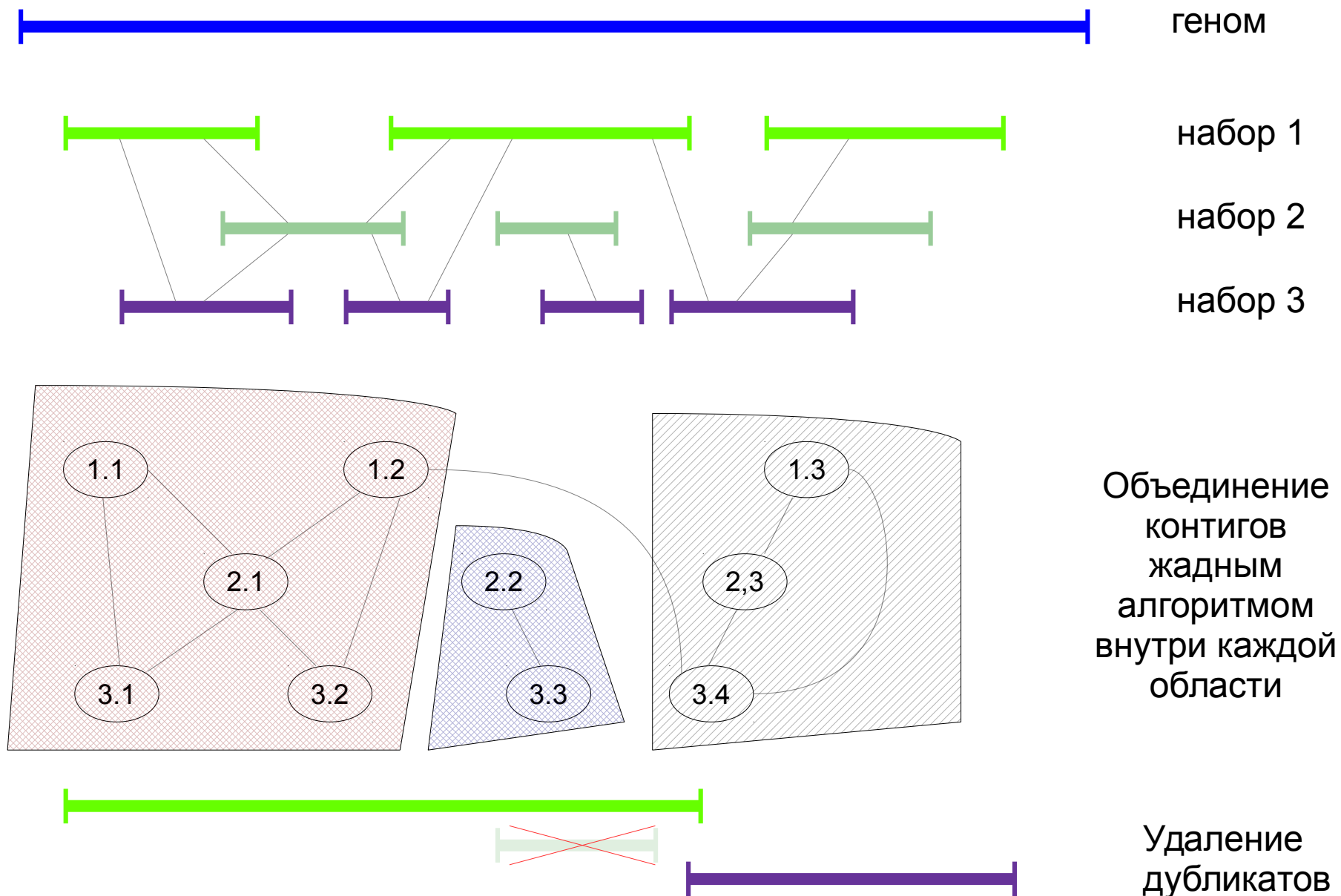
Объединение результатов работ геномных сборщиков. Пример



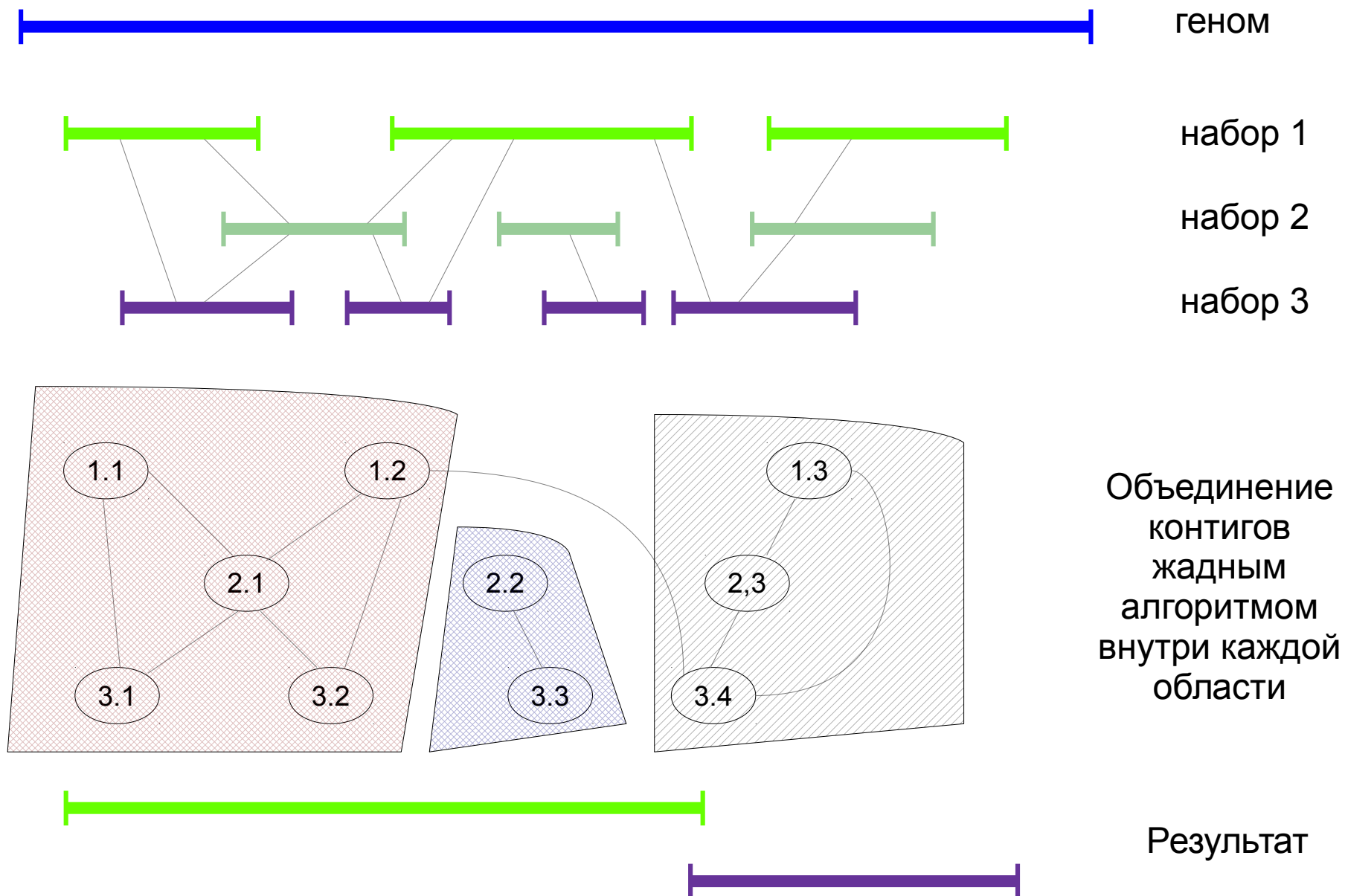
Объединение результатов работ геномных сборщиков. Пример



Объединение результатов работ геномных сборщиков. Пример



Объединение результатов работ геномных сборщиков. Пример



Предлагаемый метод. Результаты *Encephalitozoon cuniculi*

Программа	NGA50	Ошибки сборки	Число контигов (>=500 символов)	Показатель дублирования (%)	Покрывтие генома (%)
A5	134246	7	10737	0.2	90.965
Abyss	8767	5	549	0.7	90.971
Ray	6655	5	650	0.9	91.082
SOAPdenovo	1295	1	1617	1.9	81.646
Spades	134106	4	14929	0.2	90.817
IDBA	133909	4	1240	0.2	90.965
Velvet	3189	1	1111	1.3	88.492
MIX	133909	14	1635	10.4	91.635
CISA	136410	17	1446	2.1	92.341
Предложенный метод	136670	17	6787	2	92.358

Программа	Время	ОП	Место на диске
Mix	3 ч 40 м	23 Гб	5 Гб
CISA	10 ч 50 м	177 Мб	142 Мб
Предложенный метод	8 м	687 Мб	135 Мб

AMD FX-8350
1.4Ghz 8-core
32 GB RAM

Предлагаемый метод. Результаты

Candida parapsilosis

Программа*	NGA50	Ошибки сборки	Число контигов (>=500 символов)	Показатель дублирования (%)	Покрытие генома (%)
Abyss	104678	13	346	2.2	99.406
Ray	24137	13	1092	0.8	97.490
SOAPdenovo	16128	5	1533	0.3	97.353
IDBA	38332	20	757	0.2	97.596
Velvet	17131	4	1421	0.2	97.029
Mix	110568	47	297	5	99.246
CISA	95331	26	323	3.9	99.249
Предложенный метод	110597	25	318	3.8	99.330

* Сборщики A5 и SPAdes не завершили свою работу

Программа	Время	ОП	Место на диске
Mix	9 м	2,5 Гб	485 Мб
CISA	2 ч 45 м	156 Мб	214 Мб
Предложенный метод	1 м	196 Мб	93 Мб

AMD FX-8350
1.4Ghz 8-core
32 GB RAM

Предлагаемый метод. Результаты *Saccharomyces cerevisiae* S228C

Программа*	NGA50	Ошибки сборки	Число контигов (>=500 символов)	Показатель дублирования (%)	Покрытие генома (%)
Abyss	25177	58	757	0.3	92.633
Ray	13033	32	358	0.5	57.669
SOAPdenovo	15771	64	1633	0.6	92.813
Spades	73166	50	2584	0.3	93.703
IDBA	45419	58	1708	0.3	93.480
Velvet	18811	62	1011	0.4	89.529
MIX	73233	55	2515	0.5	93.654
CISA	74268	59	1869	0.9	93.752
Предложенный метод	79654	58	287	2.2	93.777

* Сборщик A5 не завершил свою работу

Программа	Время	ОП	Место на диске
Mix	40 м	1,9 Гб	388 Мб
CISA	15 ч	219 Мб	217 Мб
Предложенный метод	2 м	244 Мб	100 Мб

AMD FX-8350
1.4Ghz 8-core
32 GB RAM

Апробация

- Метод объединения результатов работ различных геномных сборщиков был интегрирован в информационно-вычислительную среду Mathcell: www.mathcell.ru/model8.php
- Создан Docker образ для предложенного метода объединения геномных сборок:
`docker pull kromanenkov/gar`

Основные результаты, выносимые на защиту

- Предложен новый критерий оценки качества сборки геномов небольших размеров при отсутствии референса с помощью анализа частот k -меров. Показано, что предложенная метрика коррелирует с референс-зависимыми характеристиками генома в отличие от существующих аналогов.
- Разработан метод выбора наилучшей сборки с точки зрения предложенного метода оценки качества геномной сборки.
- Разработан и реализован алгоритм объединения результатов работ различных геномных сборщиков при отсутствии референса.

Разработанный метод объединения превосходит существующие аналоги по большинству референс-зависимых оценок качества сборки.

Публикации

- Список ВАК:

- Романенков К.В. Метод оценки качества сборки генома на основе частот k-меров // Препринты ИПМ им. М.В.Келдыша. - 2017. - No 11. - 24 с.
- Романенков К.В., Сальников А.Н., Алексеевский А.В. Параллельный метод объединения результатов работы программ по сборке генома // Вестник ЮУрГУ, Серия Вычислительная математика и информатика. — 2016. — Том 5, No 1. — С. 24–34.

- Scopus

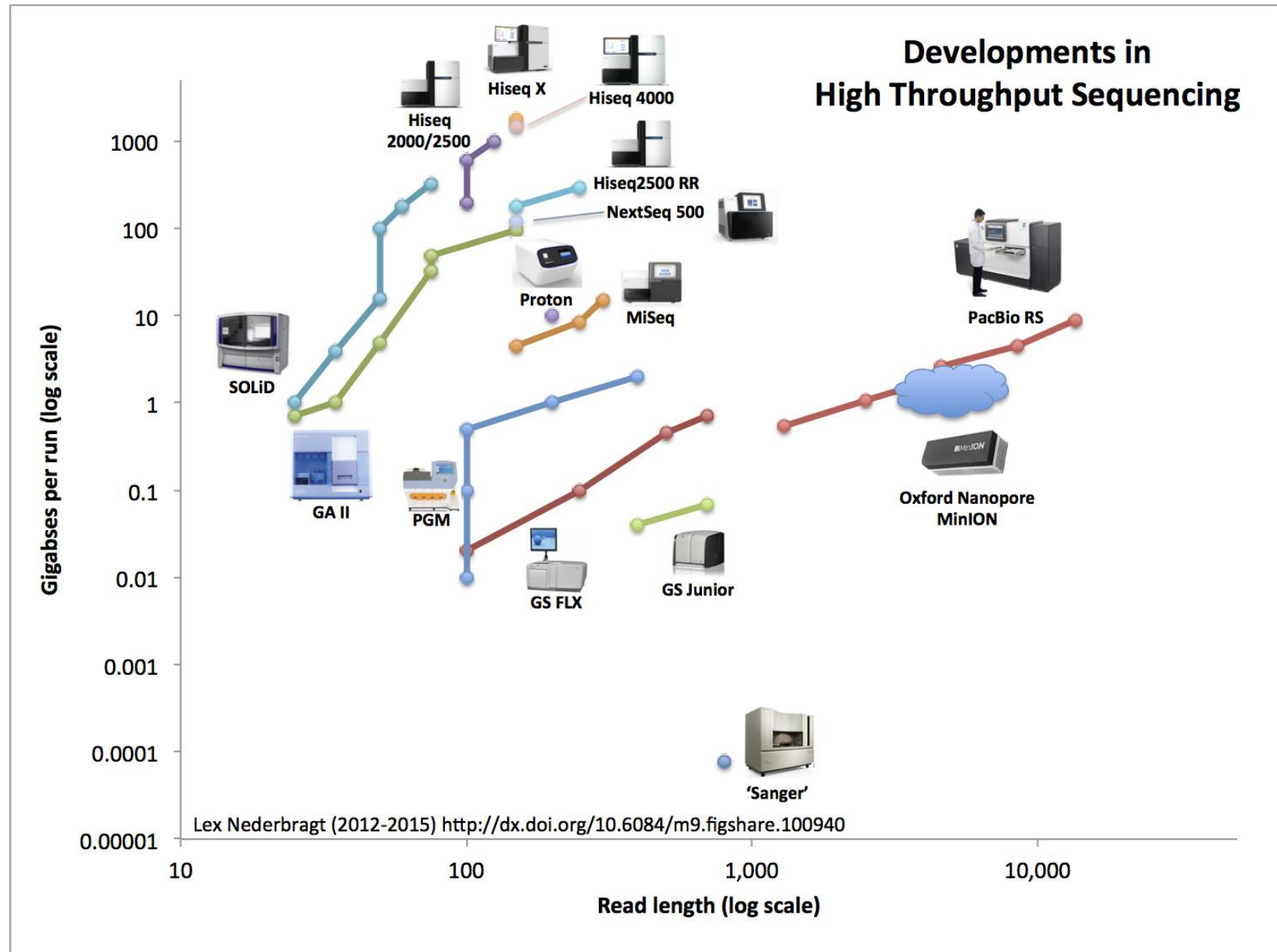
- Romanenkov K.V. Parallel merging method to integrate different genome assemblies. Proceedings of IEEE International Conference on Bioinformatics and Biomedicine (BIBM). — 2015. — P. 1461–1464.

Спасибо за внимание!

Сборка генома. Определения

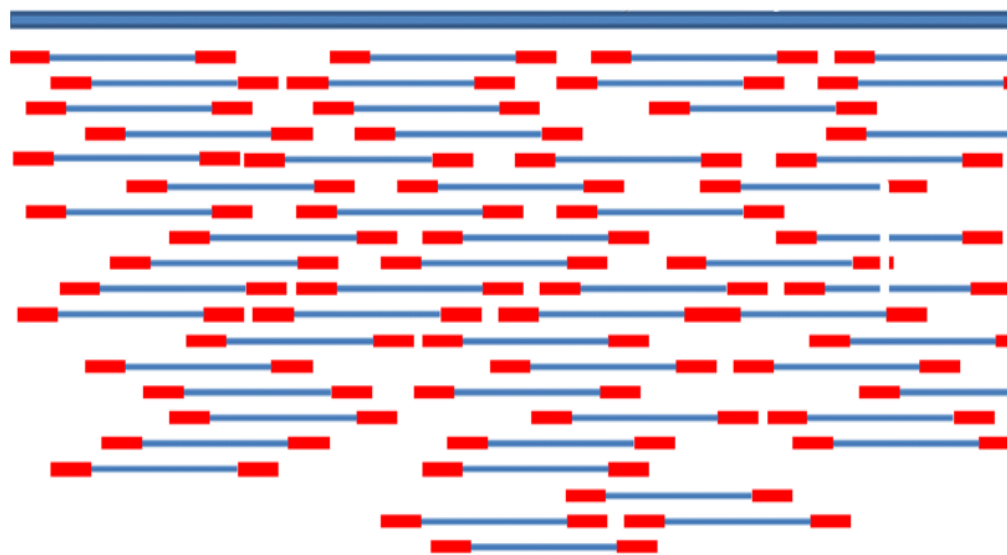
- **Assembly(сборка)** – результат работы сборщика, набор контигов (длинных фрагментов генома)
- **Read(чтение)** – короткий фрагмент генома
- **Paired read(парное чтение)** – $read_1$, пропуск ≤ 500 bp, $read_2$
- **Mate-pair(парное чтение с большим интервалом)** – $read_1$, пропуск ≥ 1 kbp, $read_2$
- **k-mer** – последовательность длины k
- **Contig(контиг)** – последовательность, полученная в процессе сборки, не содержащая gaps(N)
- **Scaffold(скаффолд)** – последовательность, полученная в процессе сборки, содержащая gaps(N)

Развитие технологий секвенирования



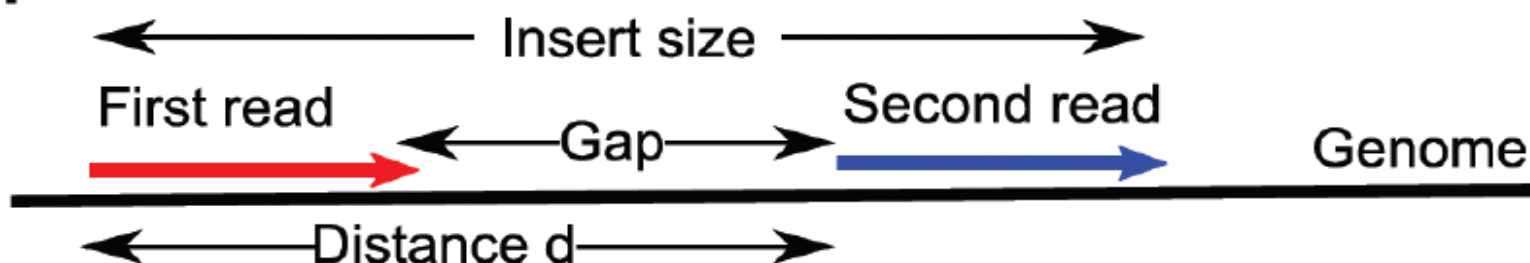
Типы ридов. Парные риды

- Paired-end (200-500)
- Mate pair (2-5k)

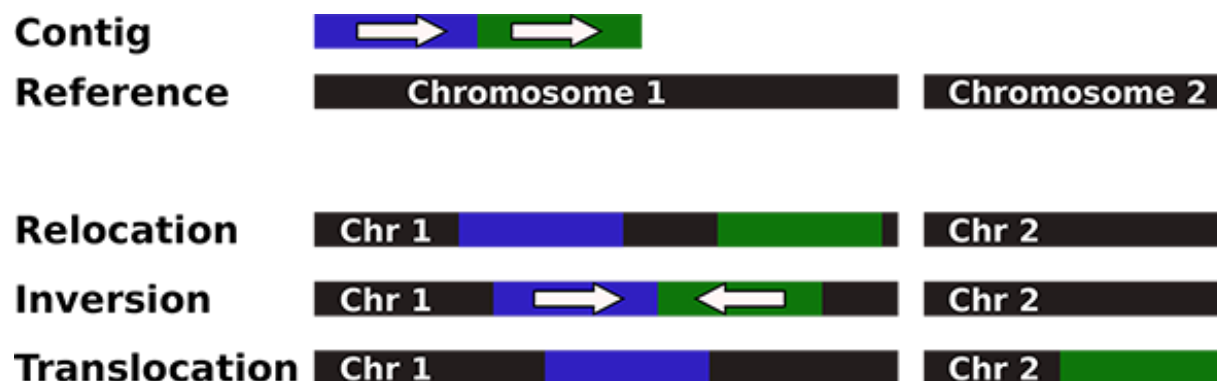


Для технологии Illumina, типичная длина ридов 100-200bp

Input

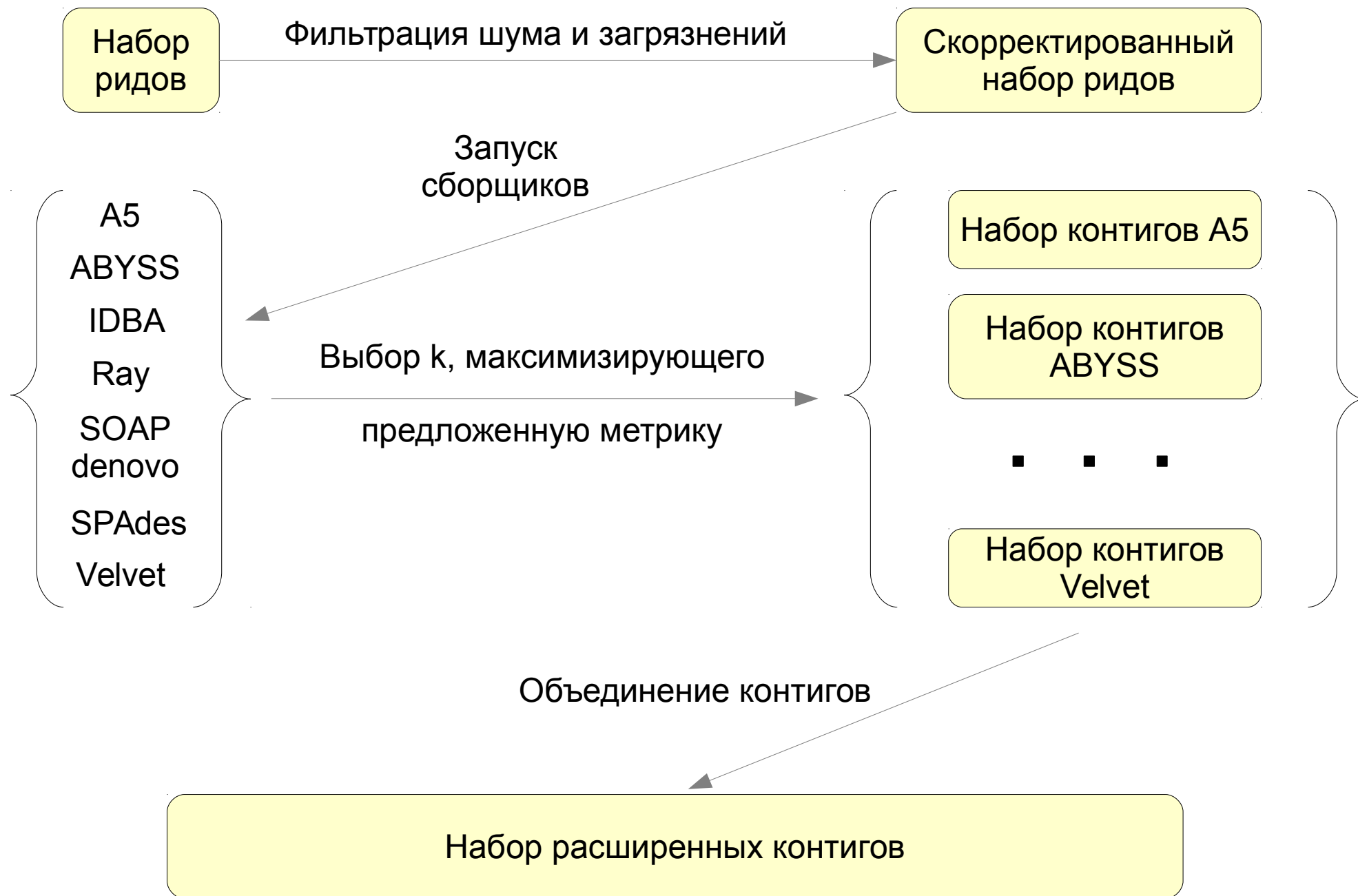


Ошибки сборки



- Relocation – картирование двух частей контига на референс либо с промежутком ≥ 1000 символов, либо с перекрытием ≥ 1000 символов
- Inversion – картирование двух частей контига на разные нити ДНК
- Translocation – картирование двух частей контига на разные хромосомы

Схема программного комплекса



Метод выбора наилучшей сборки с точки зрения предложенной метрики

- Построение гистограммы частот k-меров для ридов и каждой из полученных сборок (Jellyfish)
- Выбор окрестности пика уникальных k-меров на гистограмме встречаемости k-меров в ридах:

$$a_{p-r} = 2(p-10) - \alpha : \alpha \in [0; 8], 10 | a_{p-r}, \quad \text{где } p - \text{число встречаемости для пика уникальных k-меров}$$

$$G = \frac{S(L-k+1)}{Lp}, \quad \text{где } S - \text{общий объем коротких фрагментов, } L - \text{длина ридов, } k - \text{размер k-мера}$$

K_1^g – количество уникальных k-меров в сборке, общей размер контигов в которой ближе всего к оцененному размеру генома G

$$a_{p-l} = \max \left\{ \begin{array}{l} \max_x : \left| \bigcup_{i=x}^{a_{p-r}} K_i^r \right| \geq 0.85 |K_1^g| \\ m - \alpha : \alpha \in [1; 10], 10 | (m - \alpha) \end{array} \right., \quad \begin{array}{l} \text{где } m - \text{локальный минимум между} \\ \text{пиками на гистограмме встречаемости} \\ \text{k-меров в чтениях} \end{array}$$

- Для каждой сборки каждый k-мер из множества уникальных k-меров ридов проверяется на вхождение во множество уникальных k-меров сборки (Jellyfish) и подсчитывается значение Q
- Сборка с максимальным значением Q объявляется наилучшей

Объединение результатов работ геномных сборщиков. Предлагаемый метод

- Функция Err определялась следующим образом:

$$Err(e)=0, e=(v_i; v_j), \text{ если } \begin{cases} K m n e^{-\lambda S(e)}=0, & S(e) - \text{вес перекрытия } v_i \text{ и } v_j \\ v_i \text{ и } v_j \text{ принадлежат разным кластерам} \end{cases}$$

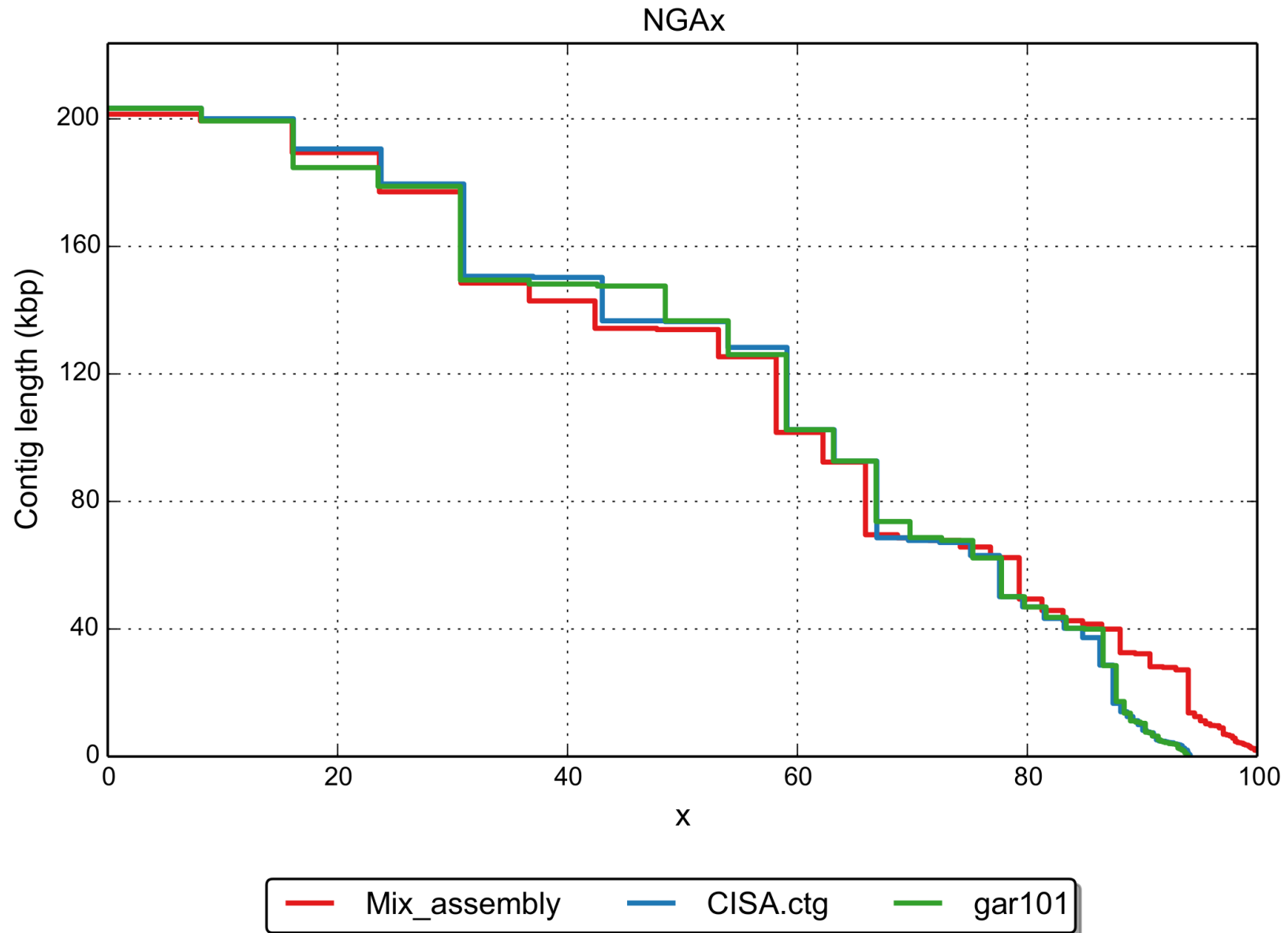
- Разбиение на кластеры проводилось приближенным алгоритмом оптимизации модулярности графа:

$$Q = \sum_{i=1}^c (e_{ii} - a_i^2), \quad e_{ij} = \sum_{vw} \frac{A_{vw}}{2m} 1_{v \in c_i} 1_{w \in c_j}, \quad a_i = \sum_j e_{ij}$$

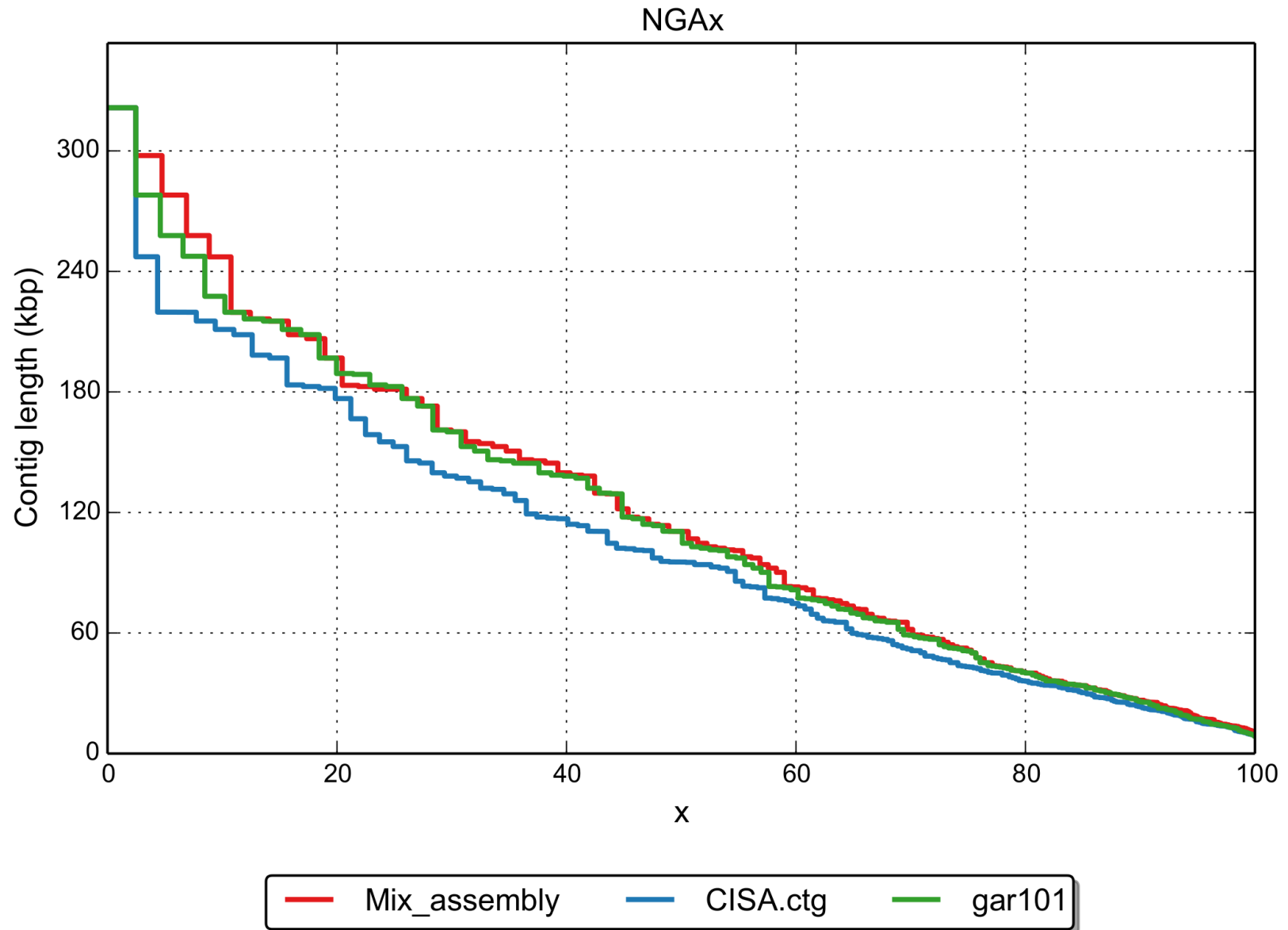
Объединение результатов работ геномных сборщиков. Предлагаемый метод

- Кластеризация взвешенного графа контигов:
- Label Propagation – подобный итеративный алгоритм
- Для каждой вершины добавляется петля, взвешенная средним весом её ребер. Каждой вершине ставится в соответствие пара <метка; вероятность>
- На каждой итерации для каждой вершины вычисляется новое множество меток
- Критерий остановки – количество обновлений меток либо равно 0, либо повторилось более чем t раз, t - параметр алгоритма

Предлагаемый метод. Результаты *Encephalitozoon cuniculi*



Предлагаемый метод. Результаты *Candida parapsilosis*



Предлагаемый метод. Результаты *Saccharomyces cerevisiae* S228C

