Федеральное государственное бюджетное автономное образовательное учреждение высшего образования
«Новосибирский национальный исследовательский государственный университет»
Кафедра информационной биологии

Федеральное государственное бюджетное научное учреждение
«Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук»

Федеральное государственное бюджетное учреждение науки
«Институт систематики и экологии животных Сибирского отделения Российской академии наук»

Сборка *de novo* и аннотация генома *Dendrolimus sibiricus*

Якимова Мария Евгеньевна,
студентка КИБ ФЕН НГУ, группа 18410,
лаборатория экологической физиологии ИСиЭЖ СО РАН

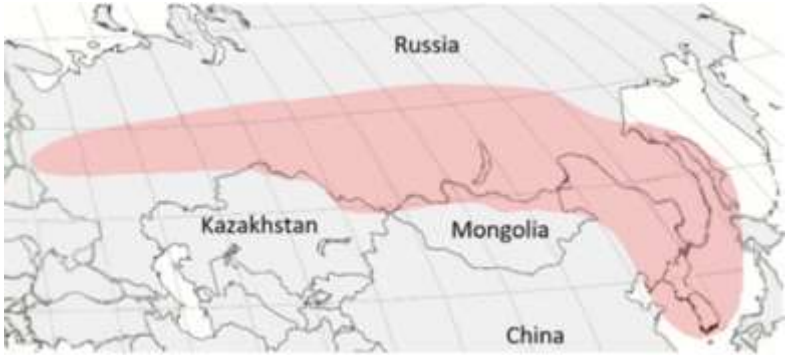
Научные руководители:
Мартемьянов Вячеслав Викторович,
к.б.н., зав. лаборатории экологической физиологии ИСиЭЖ СО РАН

Ершов Никита Игоревич,
к.б.н., с.н.с. лаборатории регуляции экспрессии генов ИЦиГ СО РАН

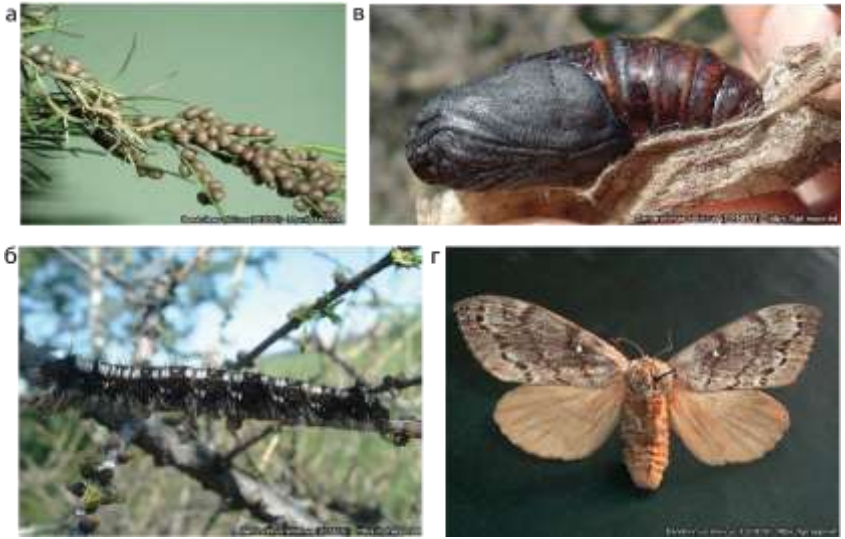
Новосибирск - 2022

Сибирский шелкопряд (*Dendrolimus sibiricus*)

В Сибирском регионе в отдельные годы площади ущерба исчисляются миллионами гектаров, где гибель деревьев может достигать 50%



Ареал Сибирского шелкопряда



Стадии развития:
а – яйца,
б – личинка,
в – куколка,
г – имаго

Секвенирование генома Сибирского шелкопряда

Наличие референсного генома является необходимой основой для современных генетических и физиологических исследований.

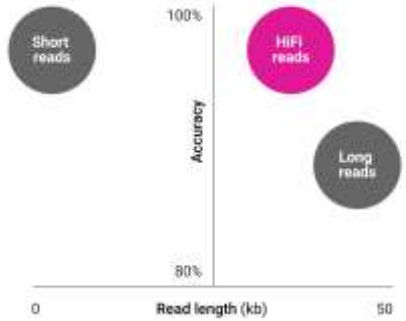
Но существуют проблемы в получении высококачественной сборки генома:

1. Повторы, которые могут быть намного длиннее прочтений секвенирования нового поколения
2. Высокая гетерозиготность естественных популяций насекомых
3. Небольшой размер некоторых насекомых, недостаточный для выделения ДНК из одной особи

→ Сильно фрагментированная сборка генома, ошибки сборки →

Решение – секвенирование третьего поколения

(длина прочтений 10-20 кб)



Цель и задачи

Цель работы: сборка de novo ядерного генома на основе данных секвенирования PacBio HiFi и функциональная аннотация генов *D. sibiricus*.

Задачи:

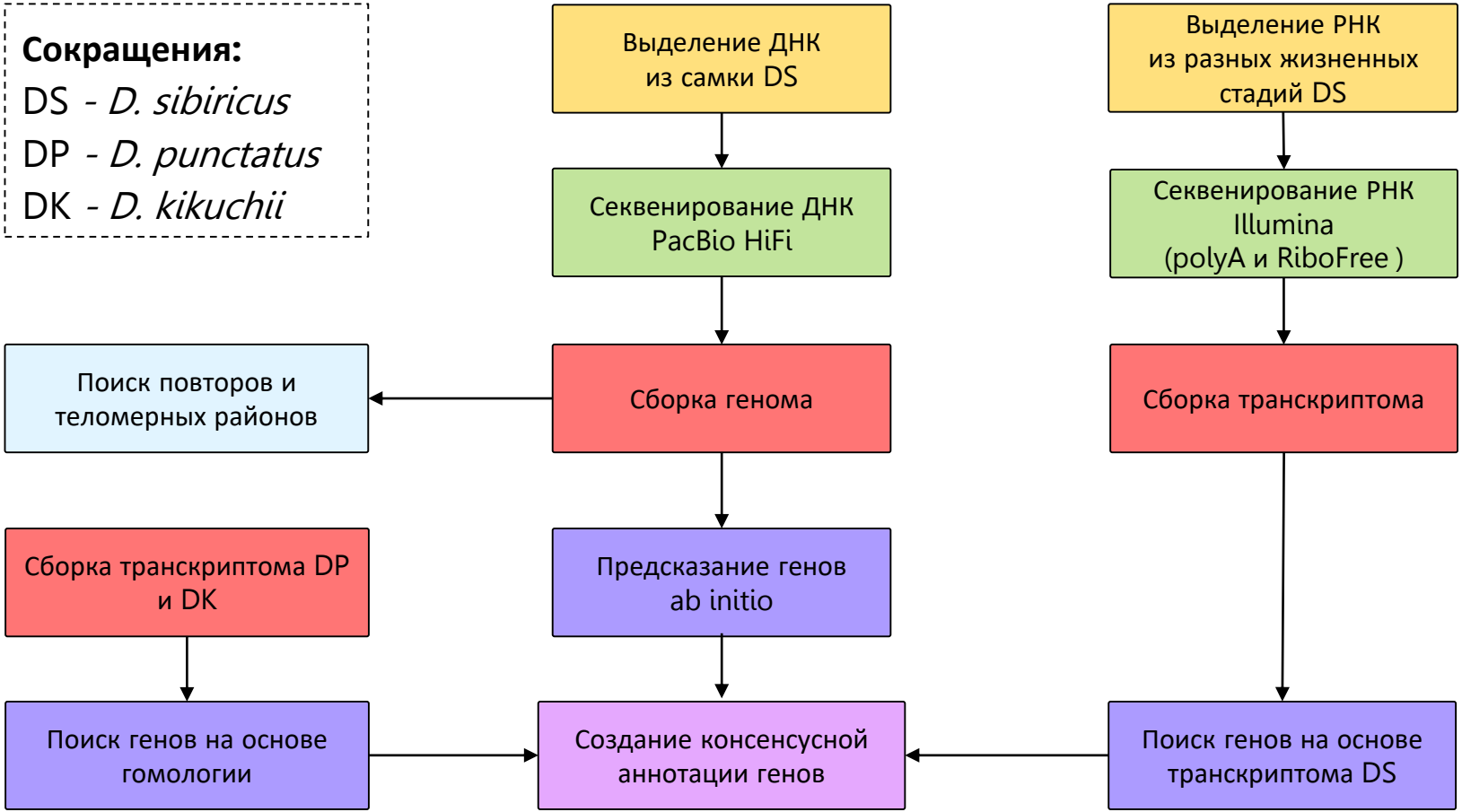
1. Выполнить сборку генома на основе данных секвенирования PacBio HiFi геномной ДНК единичной особи с использованием различных специализированных сборщиков.

2. Провести анализ качества и сравнение полученных сборок; выбрать оптимальную первичную сборку генома и оценить финализацию сборки.

3. Выполнить предсказание генов в геноме с помощью данных секвенирования репрезентативного транскриптома *D. sibiricus*, полученных двумя различными протоколами, общедоступных данных mRNA-seq близкородственных видов, а также с помощью алгоритмов предсказания *ab initio*.

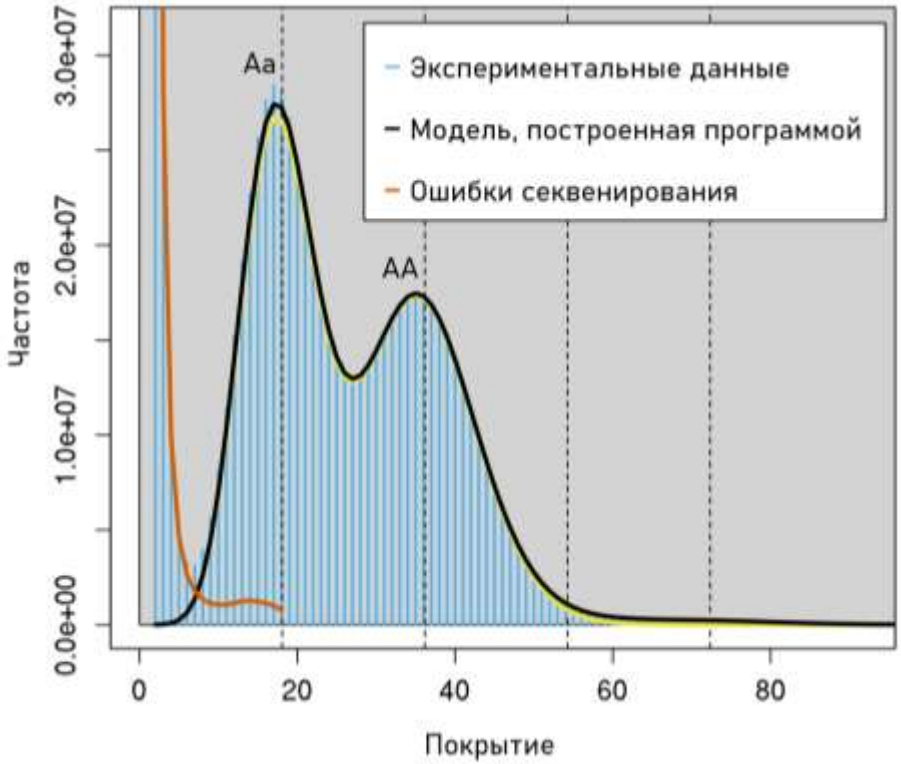
Этапы работы

Сокращения:
DS - *D. sibiricus*
DP - *D. punctatus*
DK - *D. kikuchii*



Характеристика полученных экспериментальных данных

Гистограмма распределения частот 45-меров, полученная по исходным данным PacBio HiFi



Библиотека прочтений PacBio HiFi:
N50 = 17933
Глубина секвенирования = 38

Оценочный размер генома = 507 Мб

Плоидность = 2

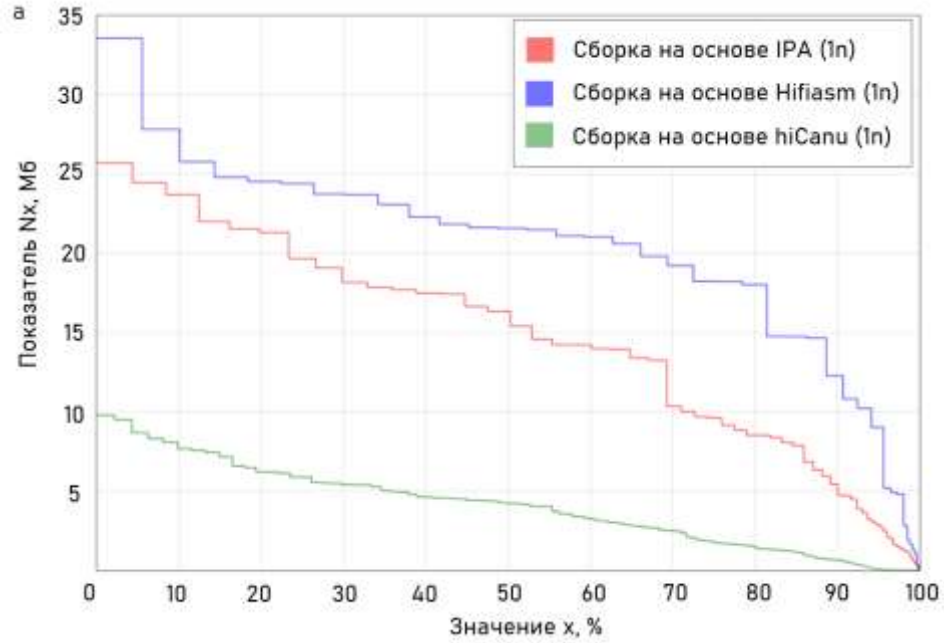
Гетерозиготность ~1%

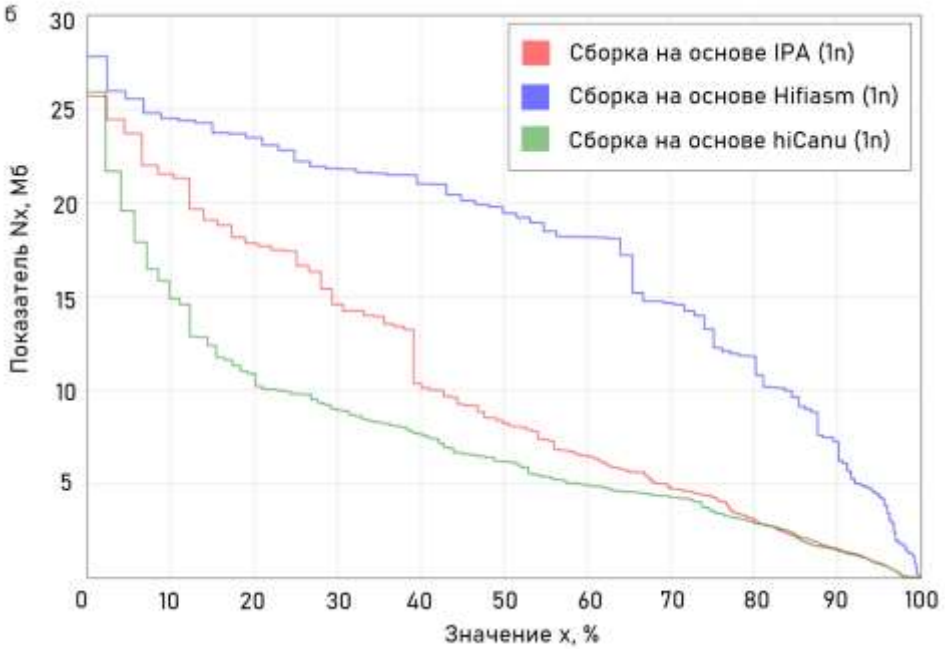
K-меры - нуклеотидные последовательности длины k, полученные разбиением прочтений путем сдвига окна размера k
Прочтение: T T C T
k-меры: T T C
T C T

Сравнение программ-сборщиков

Распределение метрики Nx для гаплоидных (а) и диплоидных (б) сборок

Каждая “ступенька” отображает контиг сборки





Контиг - набор перекрывающихся сегментов ДНК, которые в совокупности представляют собой консенсусную область ДНК, полученную в процессе сборки

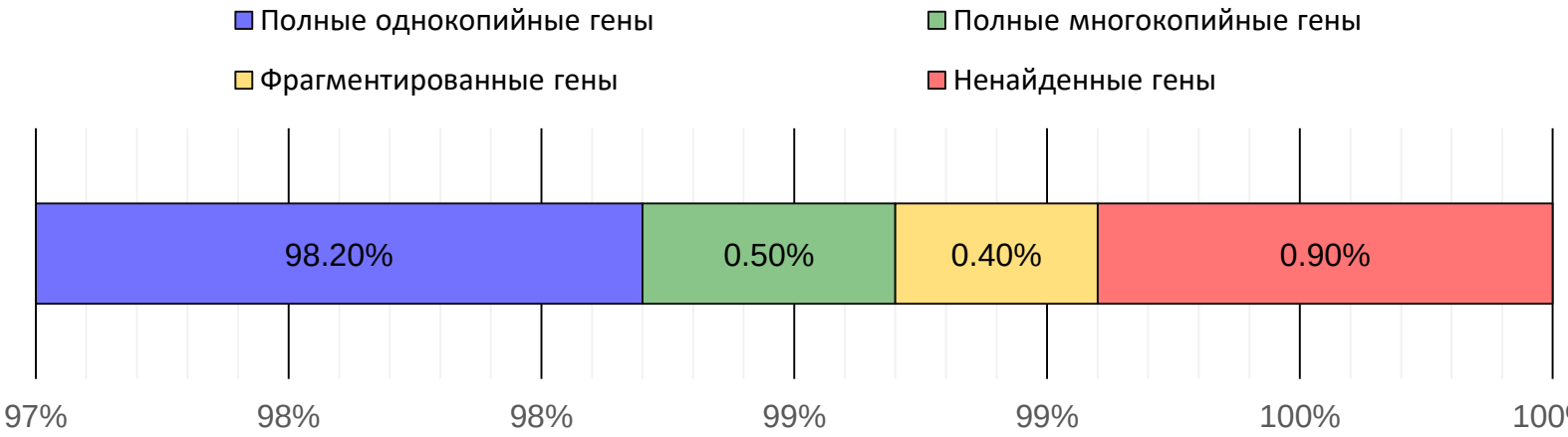
Качество полученной сборки генома

Фрагментированность сборки

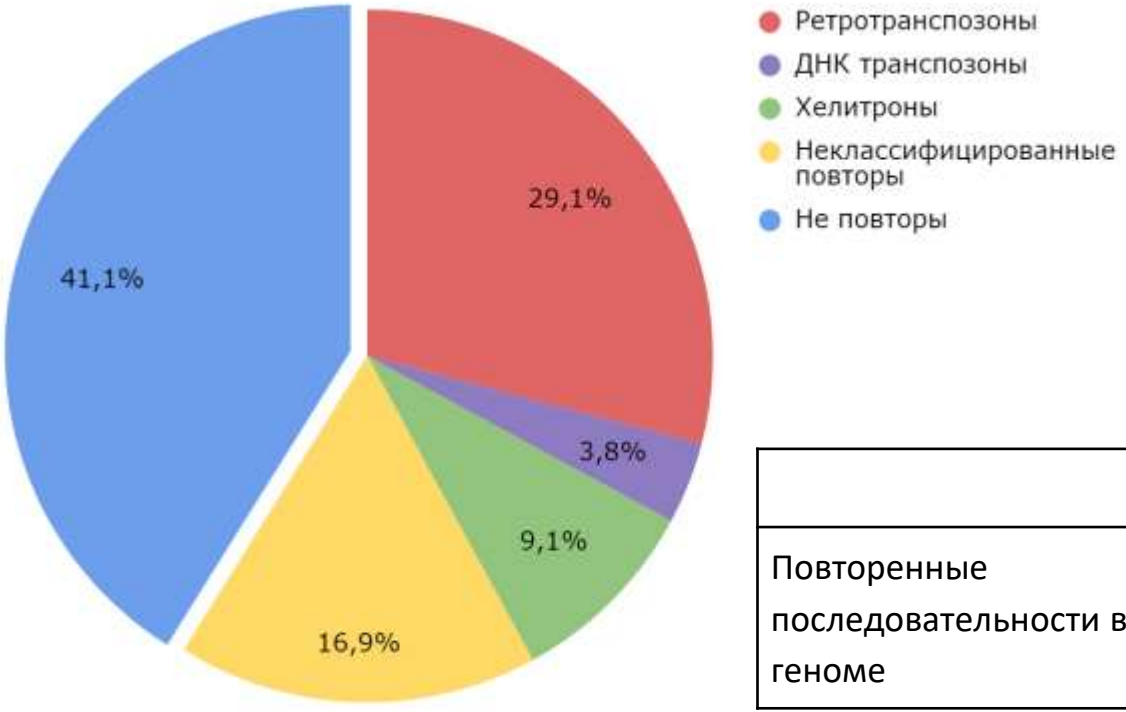
Организм	*D. sibiricus*	*D. punctatus*	*D. kikuchii*
Размер генома	600 Мб	614 Мб	705 Мб
N50	21,5 Мб	22,15 Мб	24,73 Мб

N50 — это длина, при которой совокупность всех контигов такой или большей длины покрывает не менее 50% сборки

Полнота и неизбыточность сборки



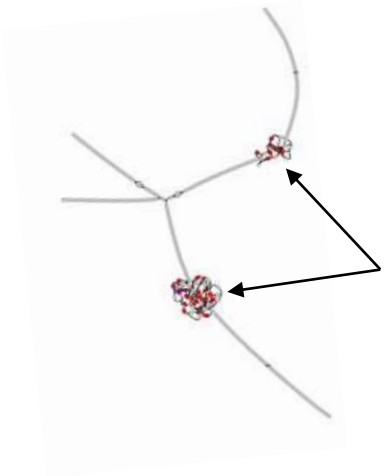
Аннотация повторов в геноме



	D. sibiricus	*D. punctatus*	*D. kikuchii*
Повторенные последовательности в геноме	59.82%	56.16%	63.44%

Анализ контигов сборки генома

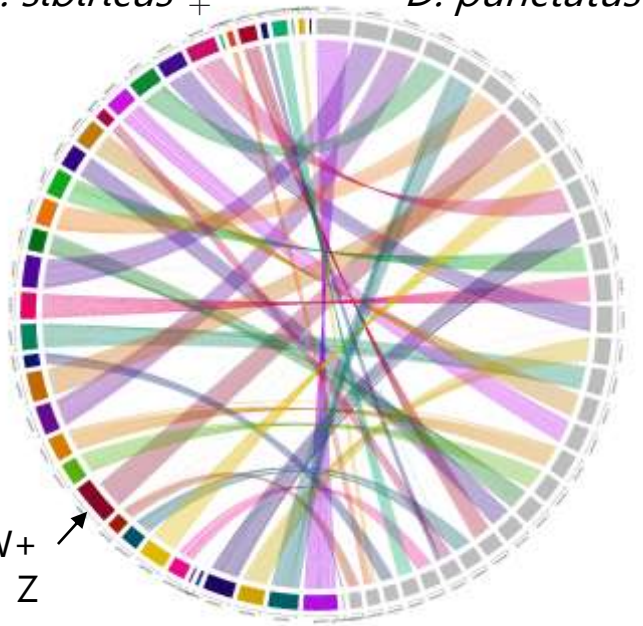
Два теломерных участка Один теломерный участок	→	28 контига 4 контига	→	28 полных хромосом 2 фрагментированные хромосомы
Нет теломерных участков	→	23 контига	→	Повторы рРНК, гистонов, фрагментированная половая хромосома W
Не ядерный геном *D. sibiricus*	→	4 контига	→	1 контиг – митохондриальный геном D. sibiricus 2 контига – геном Wolbachia 1 контиг – геном бактерии рода Bacillus



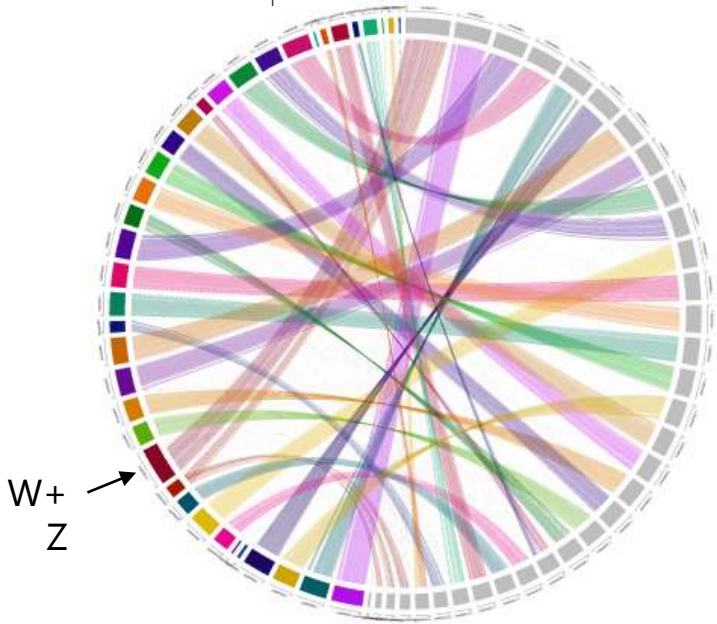
Визуализация графов юнитигов
(обозначены случайным цветом)

Синтения геномов внутри рода *Dendrolimus*

D. sibiricus ♀ *D. punctatus* ♂



D. sibiricus ♀ *D. kikuchii* ♀



Синтения - ситуация, когда расположение каких-либо локусов на одной и той же хромосоме наблюдается в разных наборах хромосом (например, у разных видов)

Сравнение количества хромосом

Организм	*D. sibiricus* (самка)	*D. punctatus* (самец)	*D. kikuchii* (самка)
Количество хромосом	29 + 2 половые (ZW)	29 + 1 половая (Z)	28 + 1 половая (Z)

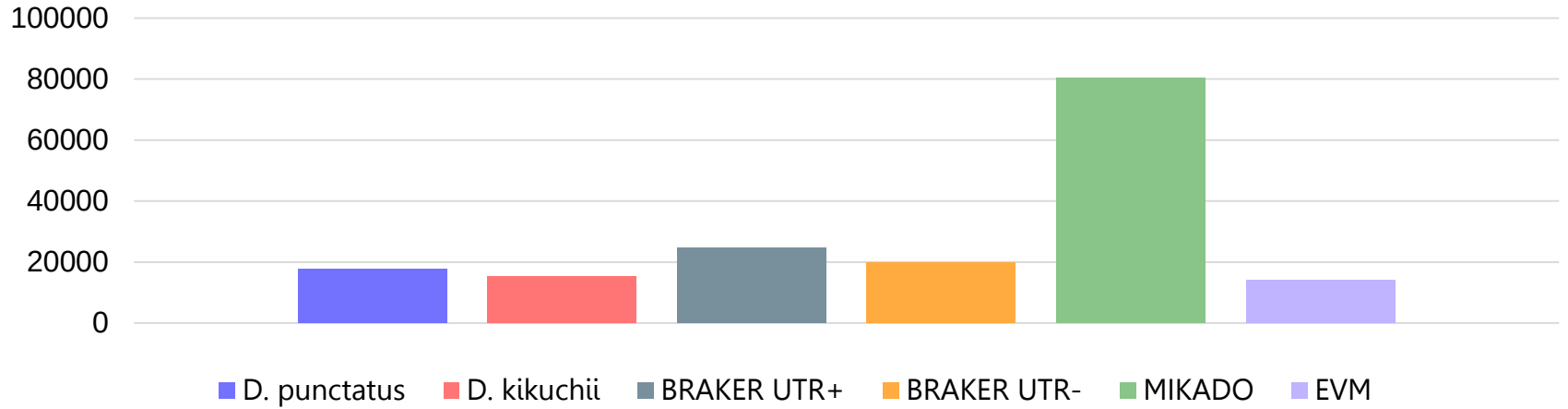
У чешуекрылых самки гетерозиготны по половым хромосомам (Z и W или Z и 0), самцы – гомозиготны (ZZ)

Аннотация генов

Модели транскриптов

	BRAKER		De novo		Genome guided	
Источник	UTR+	UTR-	mRNA	mRNA + RiboRree	mRNA	mRNA +RiboRree
Число транскриптов	28046	22641	214988	711786	216688	345448

Сравнение количества генов в полученной аннотации с близкородственными видами



Выводы

1. Исключительно на основе данных PacBio HiFi была получена сборка генома отдельной особи D. sibiricus, группы сцепления в которой собраны в подавляющем большинстве до уровня "от теломеры до теломеры". Основной причиной остаточной фрагментации сборки являются два высокоповторенных гомогенных кластера генов гистонов и рРНК, а также исключительная насыщенность повторами половой хромосомы W.
2. Из трех опробованных сборщиков, специализированных для сборки данных PacBio HiFi, оптимальные показатели качества сборки гаплоидной и частично фазированной диплоидной сборки генома D. sibiricus продемонстрировал алгоритм hifiasm.

Выводы

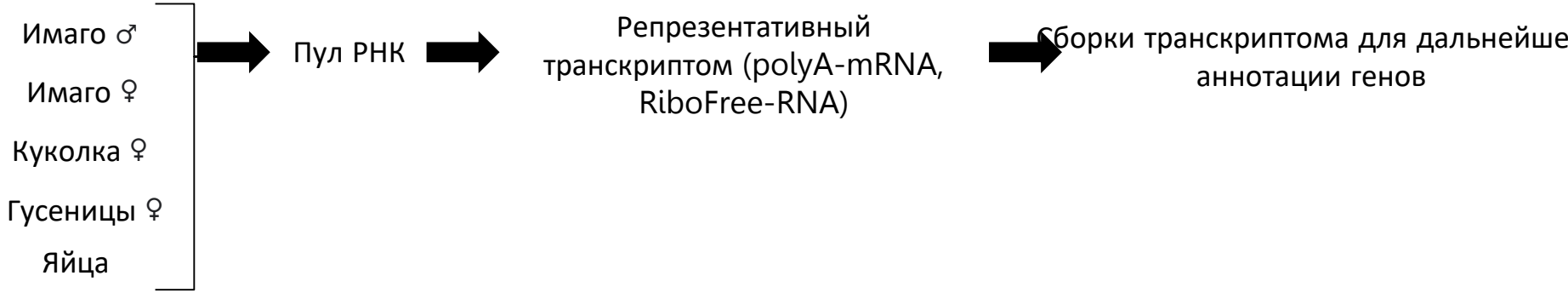
3. По результатам анализа синтении трех видов Dendrolimus гаплоидный набор хромосом D. sibiricus включает в себя 29 аутосом и две половые хромосомы Z и W, ошибочно слитые при сборке в одну группу сцепления. Число и состав полученных групп сцепления, а также спектр повторенных элементов в геноме D. sibiricus в целом соответствует таковым у родственных видов D. punctatus и D. kikuchii.

4. Побочным продуктом сборки ядерного генома D. sibiricus оказались кольцевой митохондриальный геном D. sibiricus и полноразмерный кольцевой геном его эндосимбионта Wolbachia, широко распространенного в естественной популяции сибирского шелкопряда.

5. С помощью экспериментальных данных RNA-seq получена репрезентативная сборка транскриптома D. sibiricus, на основе которой была создана аннотация генов в геноме D. sibiricus. Расширенный вариант аннотации, полученный с помощью инструмента MIKADO содержит более 80000 генов, включая как белок-кодирующие так и гены некодирующих РНК. Из их числа с помощью инструмента EVM получен набор из 14000 высокодостоверных моделей белок-кодирующих генов, что немногим отличается от числа таковых у близкородственных видов.

Спасибо за внимание!

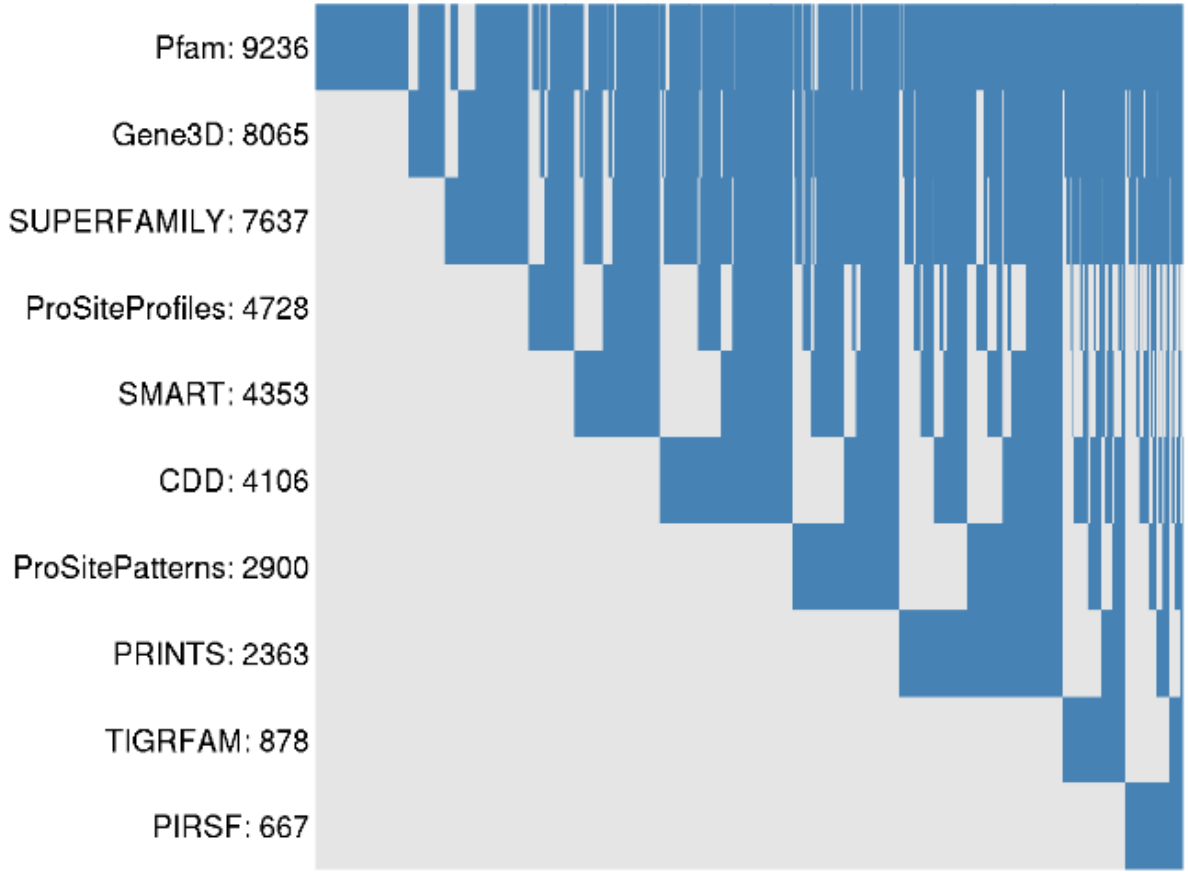
Сборка транскриптома



Характеристика транскриптомов, полученных двумя способами – de novo и на основе референсного генома (genome guided)

Вариант сборки	De novo			Genome guided		
	mRNA	RiboFree	mRNA + RiboFree	mRNA	RiboFree	mRNA + RiboFree
Общее количество транскриптов	178989	623401	609558	138327	535683	187277
Средняя длина контига	1003.80	641.55	575.18	1318.26	708.22	1152.30
N50	1945	956	766	3200	1166	3142

Функциональная аннотация генов

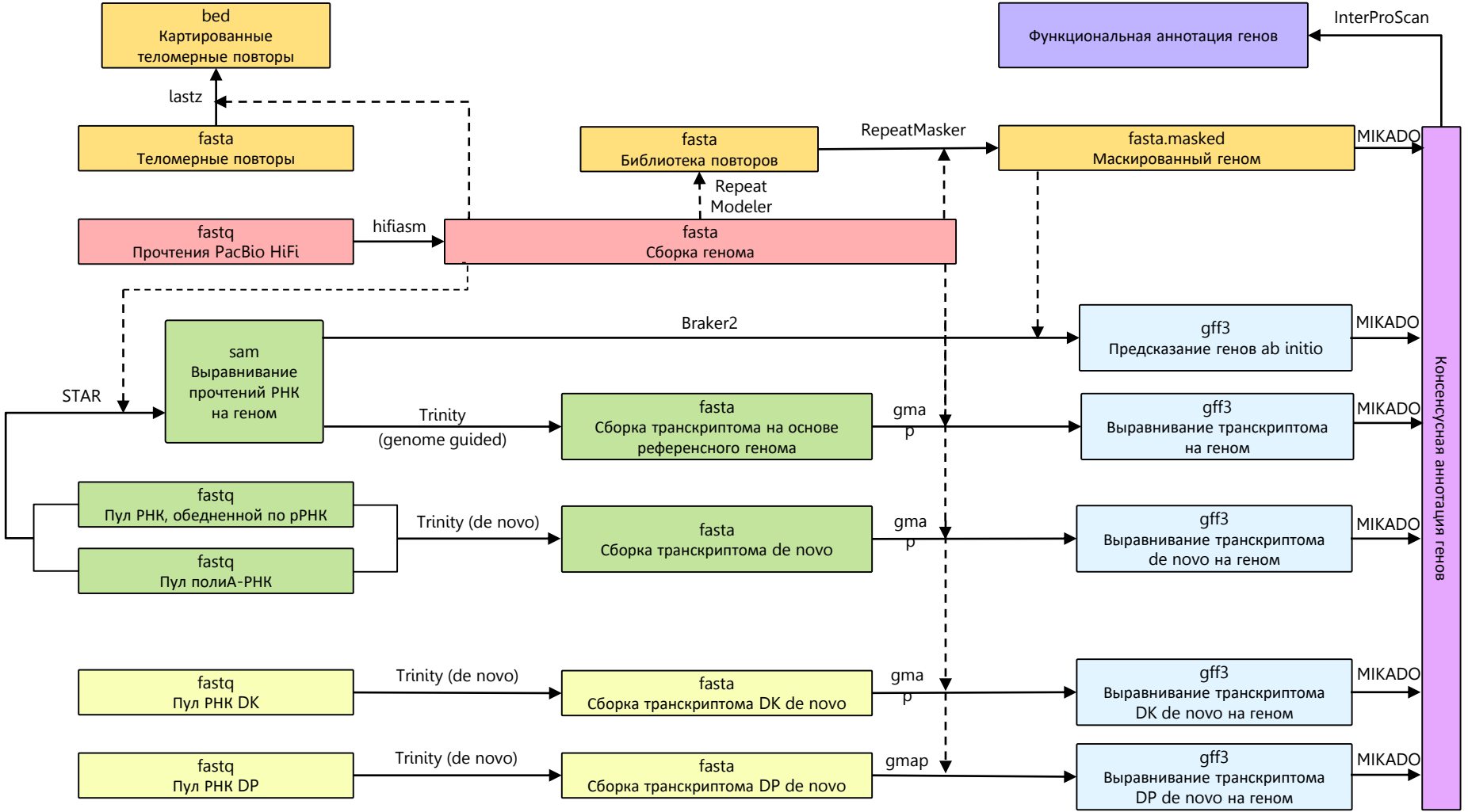


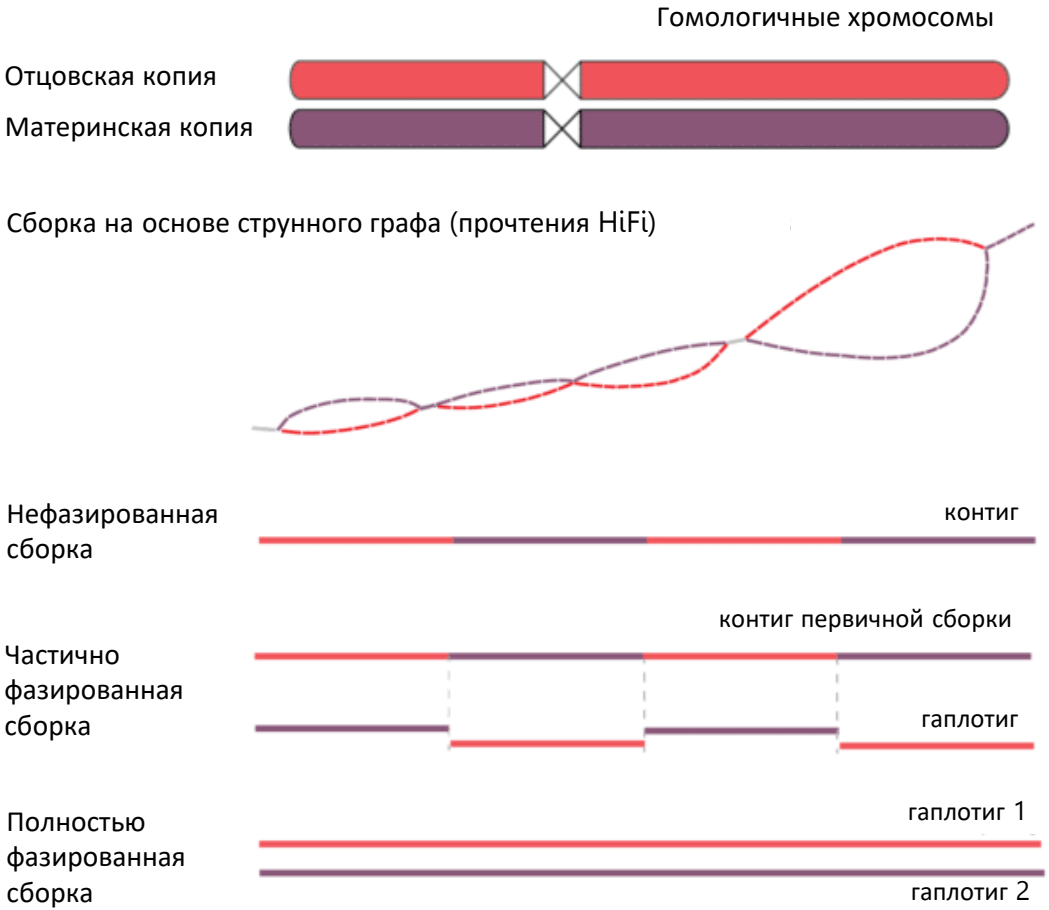
Всего 11478 генов аннотированы хотя бы одной записью InterProScan, по консервативным белковым доменам — 10017.
На картинке по Y — базы данных по доменам, по X — синим аннотированные в базе гены. Чем больше синего, тем больше вклад базы в аннотацию

Лес после вспышки размножения Сибирского шелкопряда превращается в «пороховую бочку»

В Сибирском регионе в отдельные годы площади ущерба исчисляются миллионами гектаров, где гибель деревьев может достигать 50%.

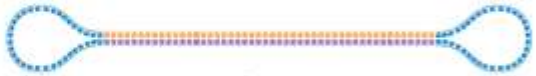
Фото: Natalia Kirichenko, John H. Ghent


bed
Картированные теломерные повторы
lastz
fasta
Теломерные повторы
fastq
Прочтения PacBio HiFi
hifiasm
fasta
Сборка генома
fasta
Библиотека повторов
RepeatMasker
Repeat Modeler
fasta.masked
Маскированный геном
Функциональная аннотация генов
InterProScan
MIKADO
Braker2
gff3
Предсказание генов ab initio
STAR
sam
Выравнивание прочтений РНК на геном
Trinity (genome guided)
fasta
Сборка транскриптома на основе референсного генома
gmap
gff3
Выравнивание транскриптома на геном
fastq
Пул РНК, обедненной по рРНК
fastq
Пул полиА-РНК
Trinity (de novo)
fasta
Сборка транскриптома de novo
gff3
Выравнивание транскриптома de novo на геном
fastq
Пул РНК DK
fasta
Сборка транскриптома DK de novo
gff3
Выравнивание транскриптома DK de novo на геном
fastq
Пул РНК DP
fasta
Сборка транскриптома DP de novo
gff3
Выравнивание транскриптома DP de novo на геном
Консенсусная аннотация генов


Гомологичные хромосомы
Отцовская копия
Материнская копия
Сборка на основе струнного графа (прочтения HiFi)
Нефазированная сборка
контиг
контиг первичной сборки
Частично фазированная сборка
гаплотиг
гаплотиг 1
Полностью фазированная сборка
гаплотиг 2

Двучепочечная высокомолекулярная ДНК

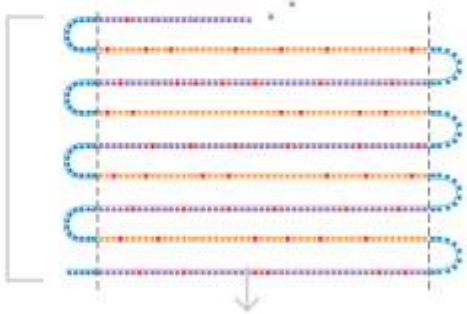
Лигирование SMRTbell адаптеров



Отжиг праймеров и связывание ДНК-полимеразы

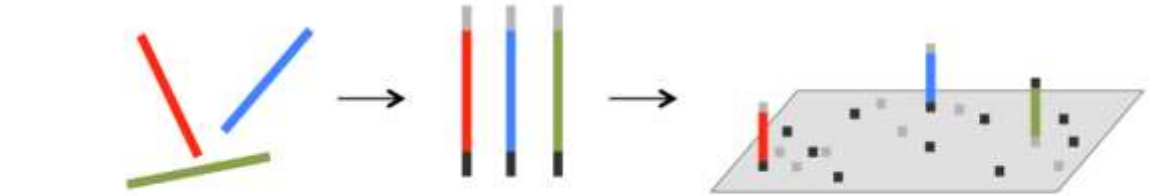
Кольцевая ДНК реплицируется много раз

Удаляются адаптеры и получаются субпрочтения



Из субпрочтений составляется консенсусная последовательность

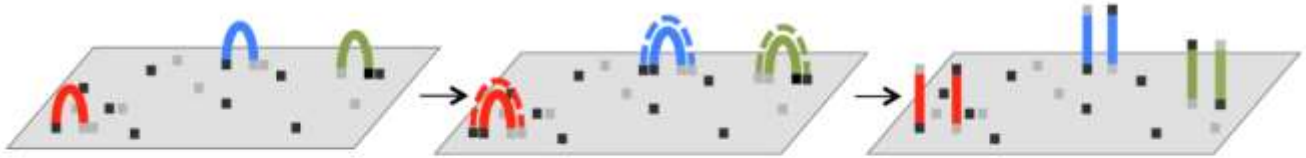
Прочтение HiFi (>99,9% точность)



Фрагменты ДНК

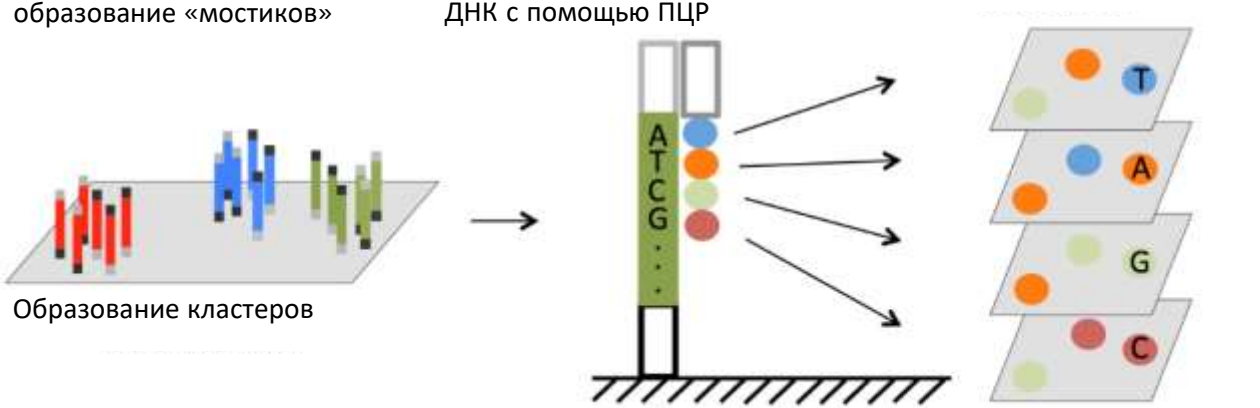
Лигирование адаптеров

Прикрепление к ячейке (flowcell)



Отжиг праймеров и образование «мостиков»

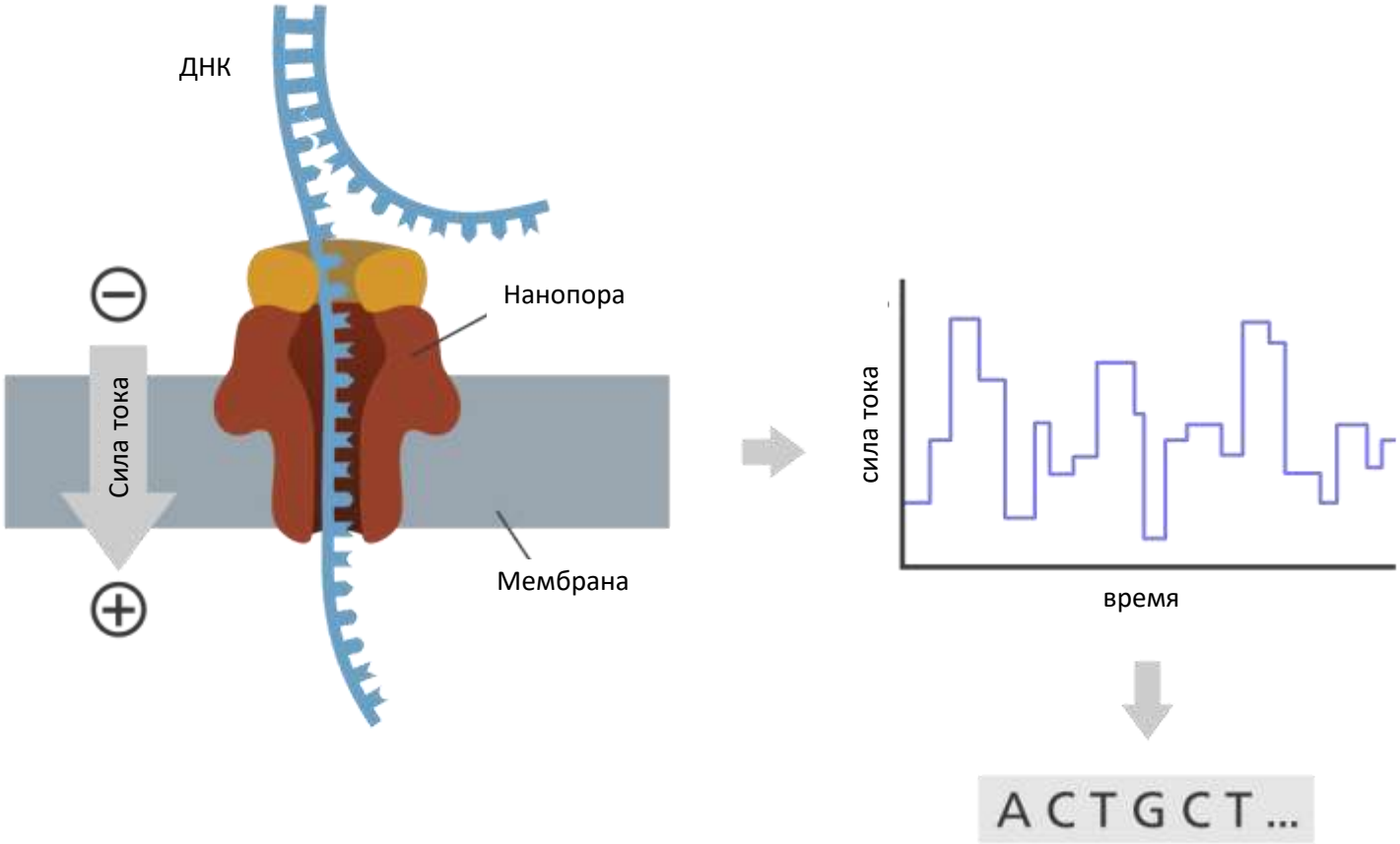
Наработка фрагментов ДНК с помощью ПЦР



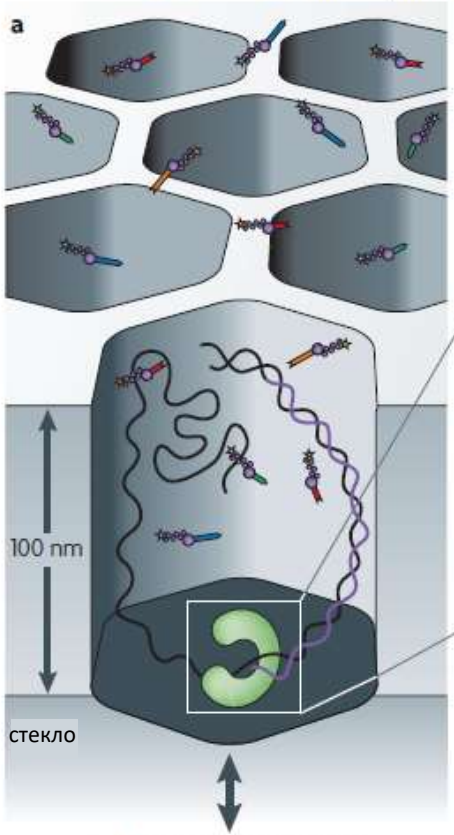
Образование кластеров

Секвенирование

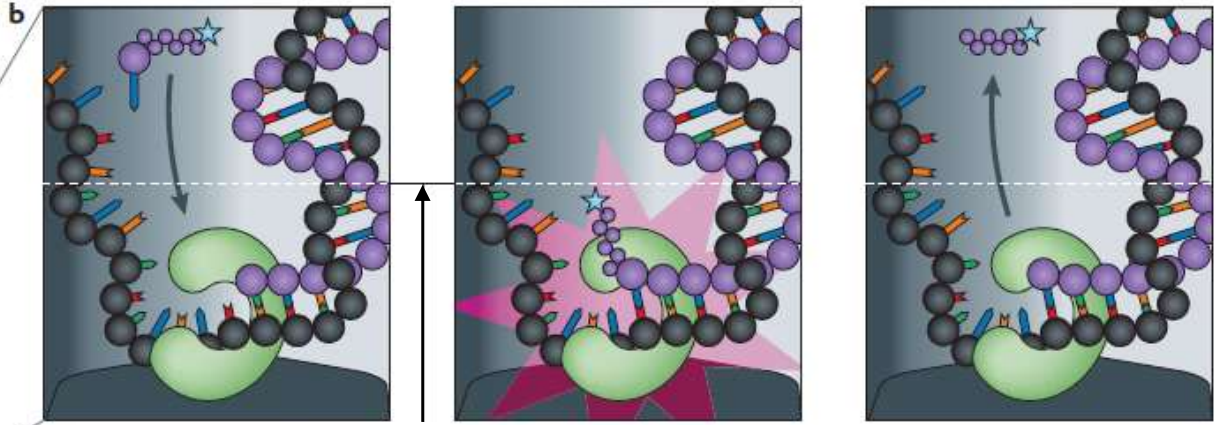
Сканирование сигнала


ДНК
Нанопора
Сила тока
Мембрана
сила тока
время
ACTGCT...

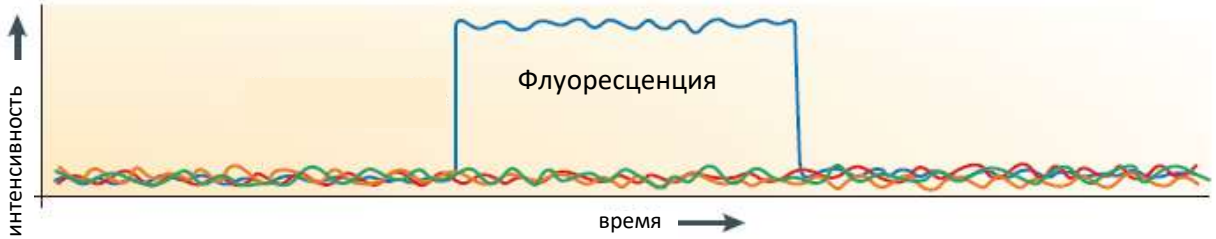
Pacific Biosciences — Real-time sequencing

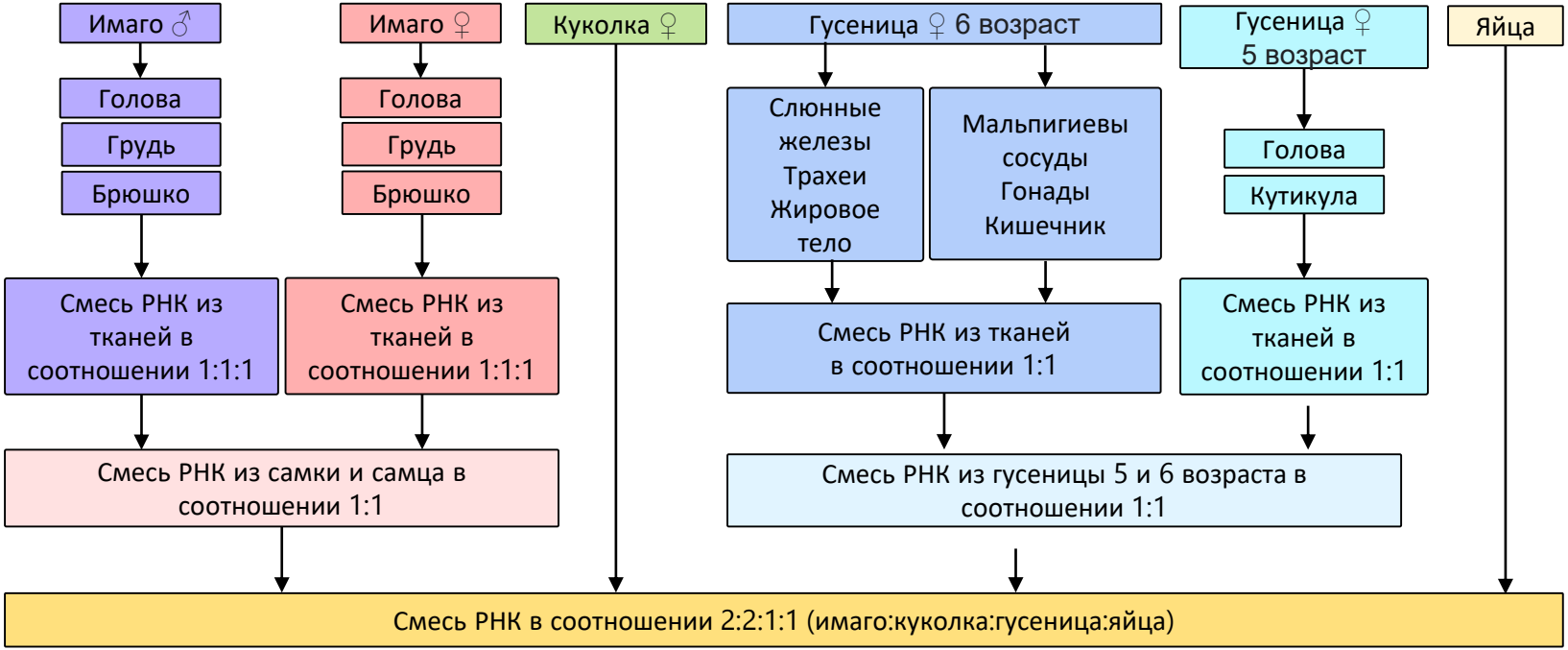


Обнаружение
флуоресцентного сигнала

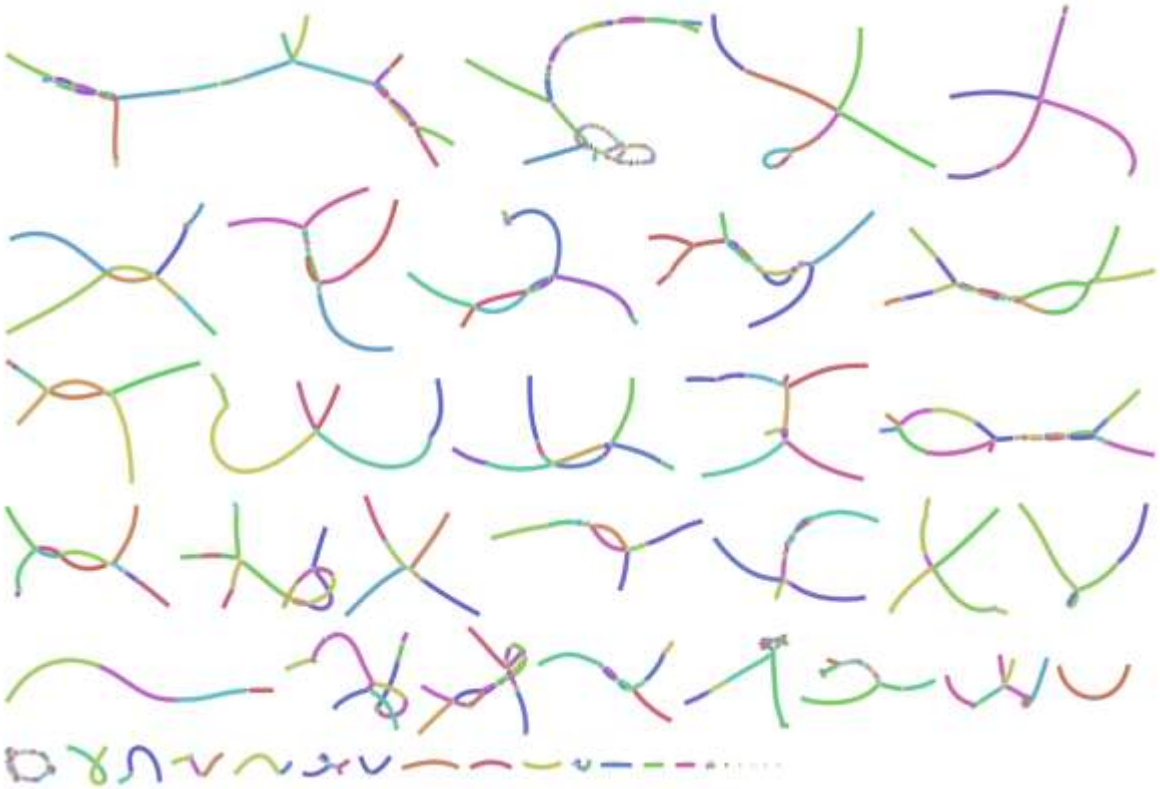


Граница зоны обнаружения флуоресценции




Имаго ♂
Голова
Грудь
Брюшко
Смесь РНК из тканей в соотношении 1:1:1
Имаго ♀
Голова
Грудь
Брюшко
Смесь РНК из тканей в соотношении 1:1:1
Смесь РНК из самки и самца в соотношении 1:1
Куколка ♀
Гусеница ♀ 6 возраст
Слюнные железы
Трахеи
Жировое тело
Мальпигиевы сосуды
Гонады
Кишечник
Смесь РНК из тканей в соотношении 1:1
Гусеница ♀ 5 возраст
Голова
Кутикула
Смесь РНК из тканей в соотношении 1:1
Смесь РНК из гусеницы 5 и 6 возраста в соотношении 1:1
Яйца
Смесь РНК в соотношении 2:2:1:1 (имаго:куколка:гусеница:яйца)

а



б

Кластер генов гистонов

в

Кластер генов рРНК

DBG метод

Разбиение прочтений на k-меры

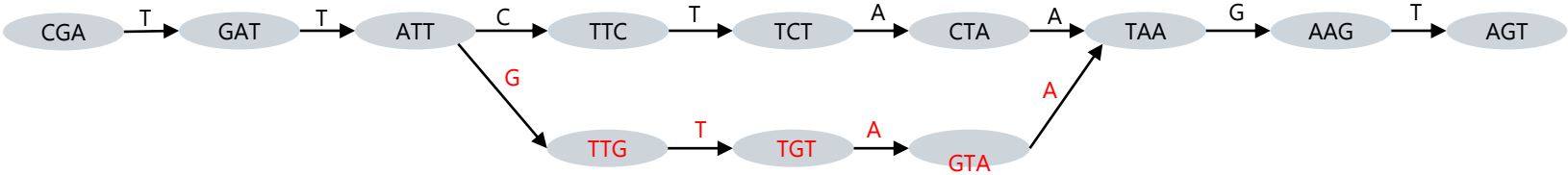
Построение графа

Обход графа и построение консенсусной последовательности

Прочтение 1: TTCTAAGT
k-меры: TTC TCT CTA TAA AAG AGT

Прочтение 2: CGATTCTA
k-меры: CGA GAT ATT TTC TCT CTA

Прочтение 3: GATTGTAA
k-меры: GAT ATT TTG TGT GTA TAA



CGATTCTAAGT

OLC метод

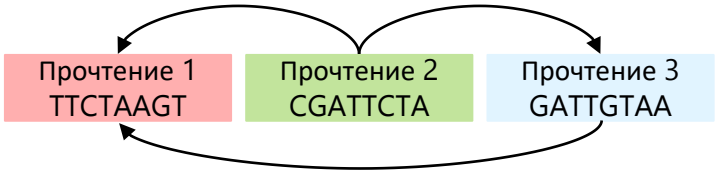
Overlap
(обнаружение перекрытий среди всех прочтений)

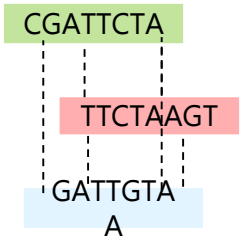
↓

Layout
(компоновка прочтений)

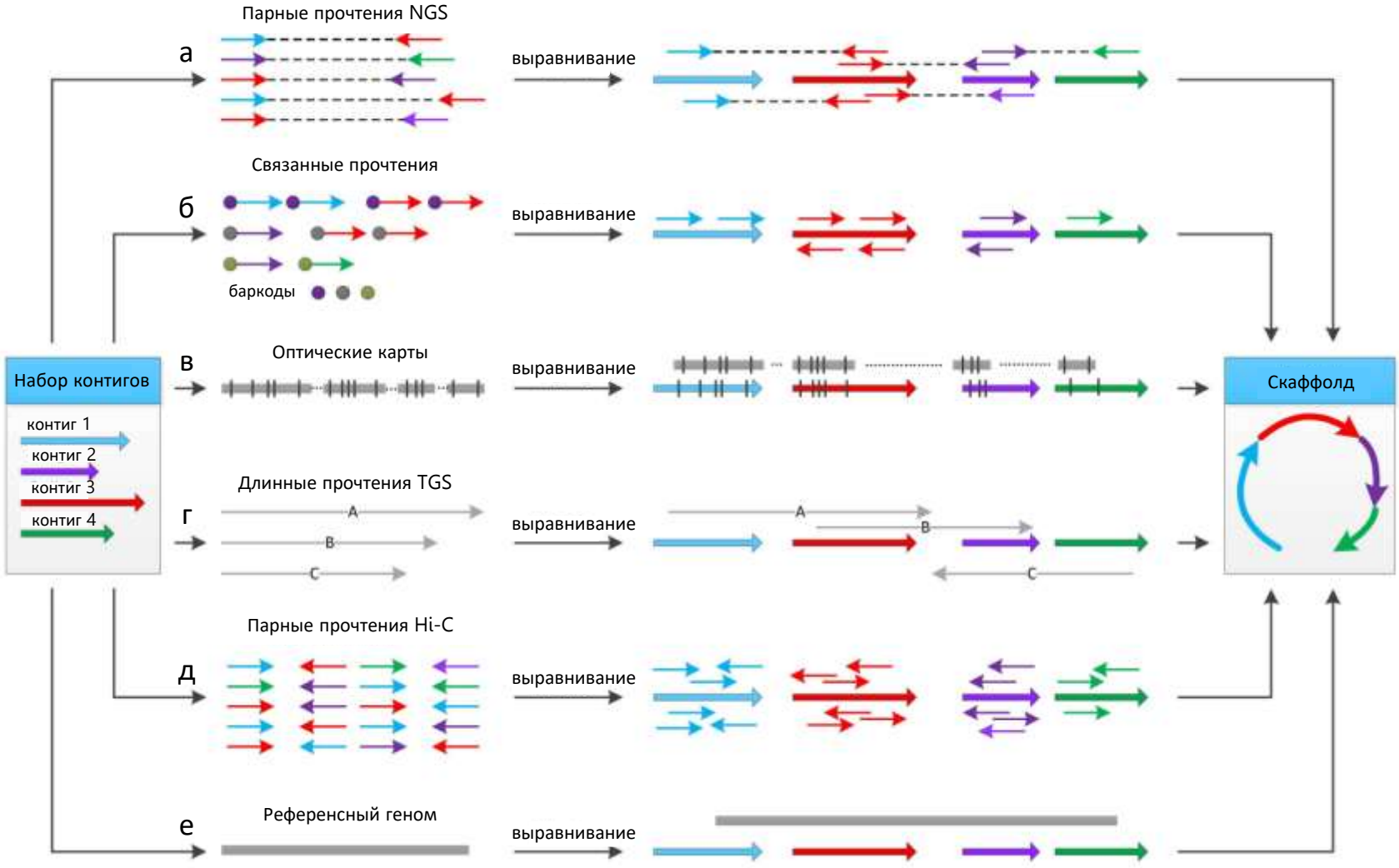
↓

Consensus
(составление консенсусной последовательности)

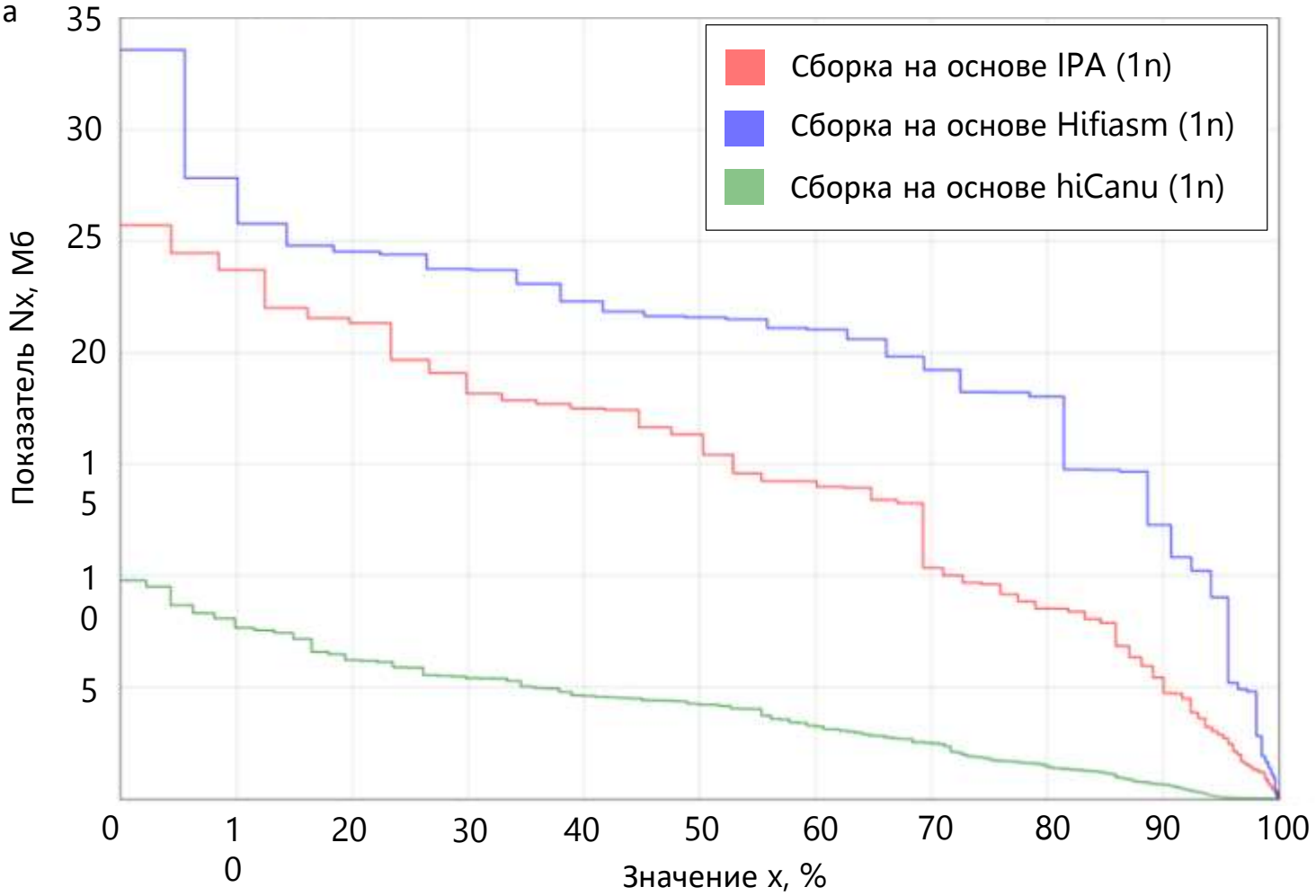




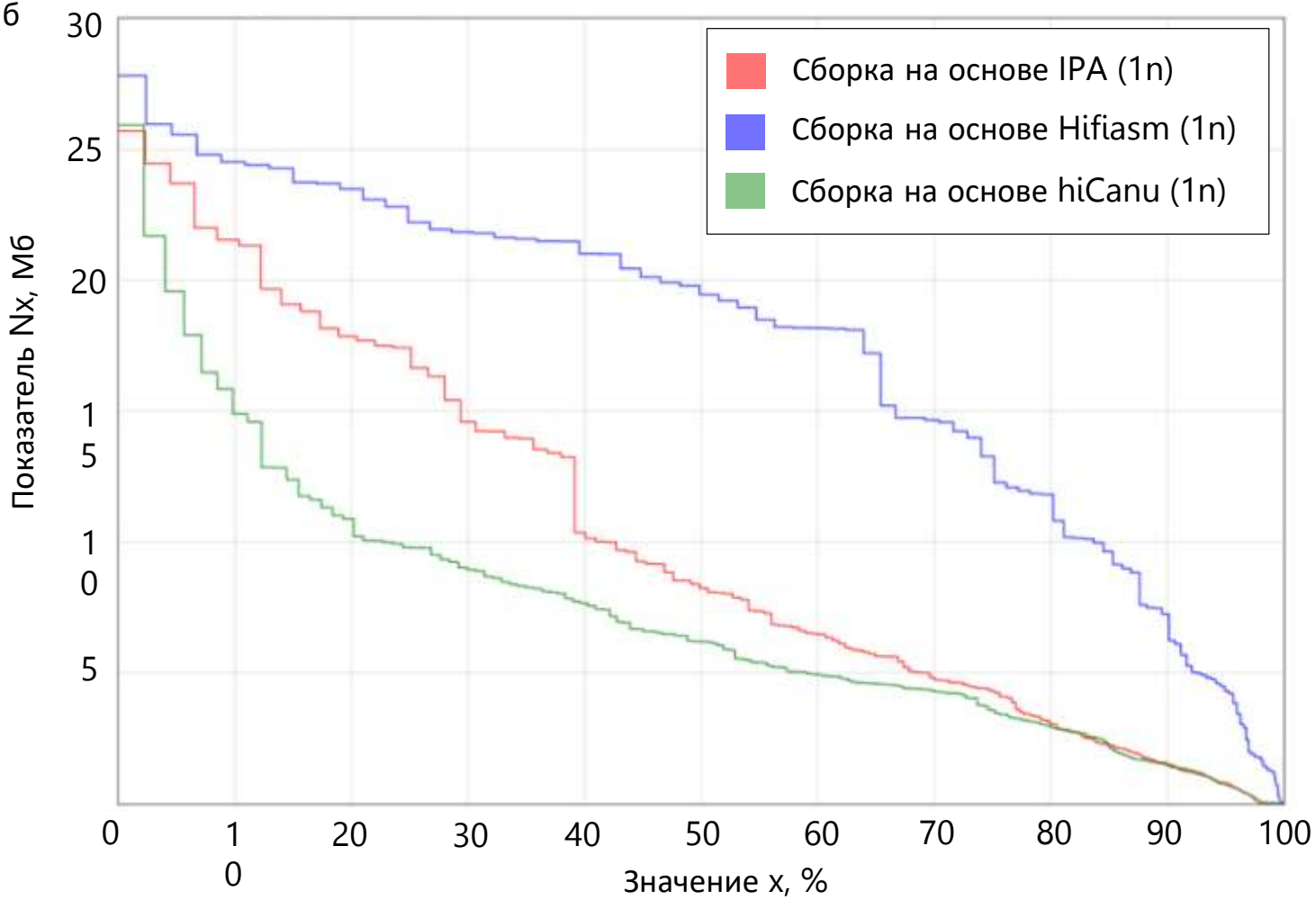

```
CGATTCTA
    TTCTAAGT
 GATTGTAA
 ____________
CGATTCTAAGT
```


Парные прочтения NGS
а
выравнивание
Связанные прочтения
б
баркоды
выравнивание
Набор контигов
контиг 1
контиг 2
контиг 3
контиг 4
в
Оптические карты
выравнивание
Скаффолд
Длинные прочтения TGS
г
A
B
C
выравнивание
Парные прочтения Hi-C
д
выравнивание
Референсный геном
е
выравнивание

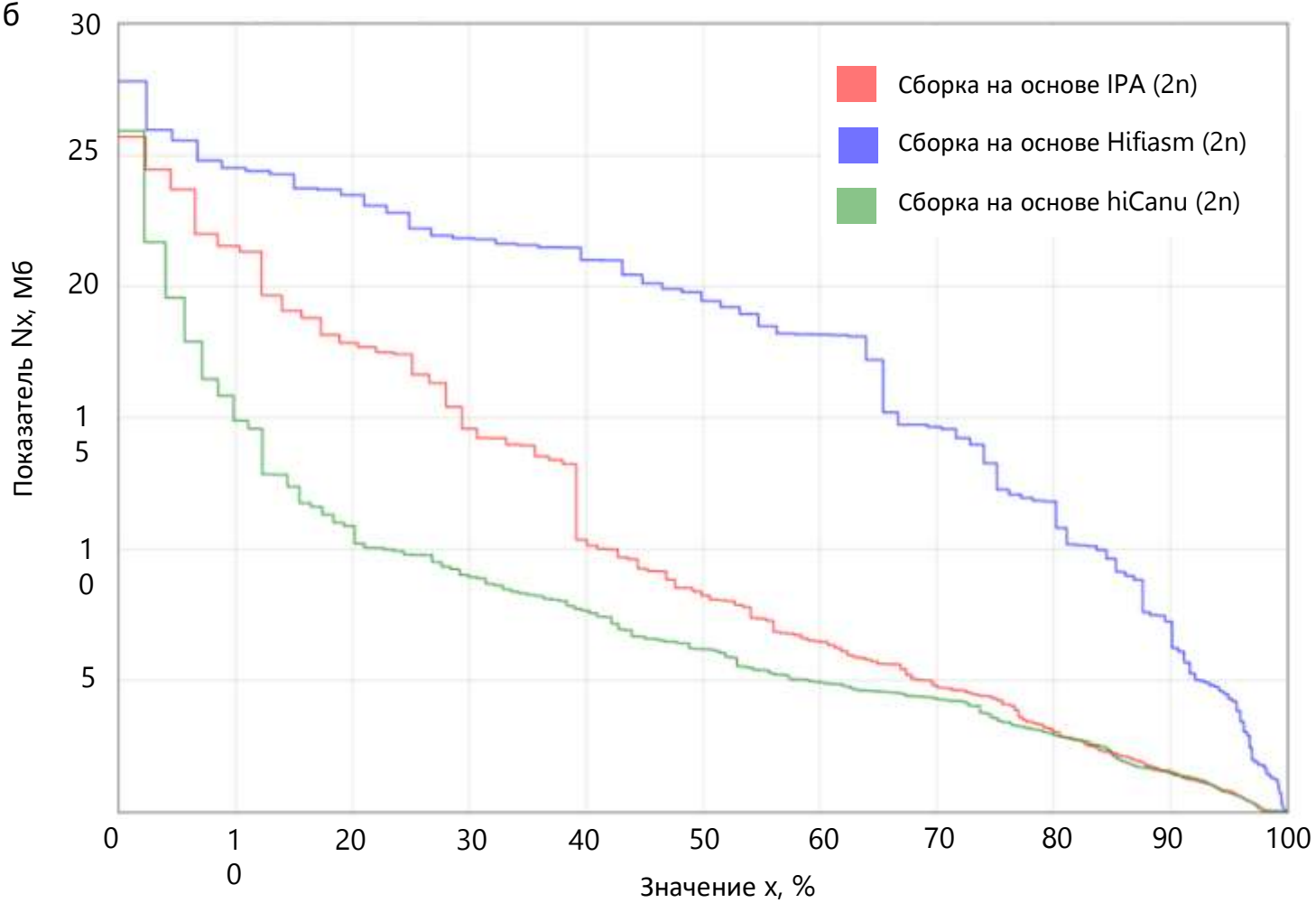
а

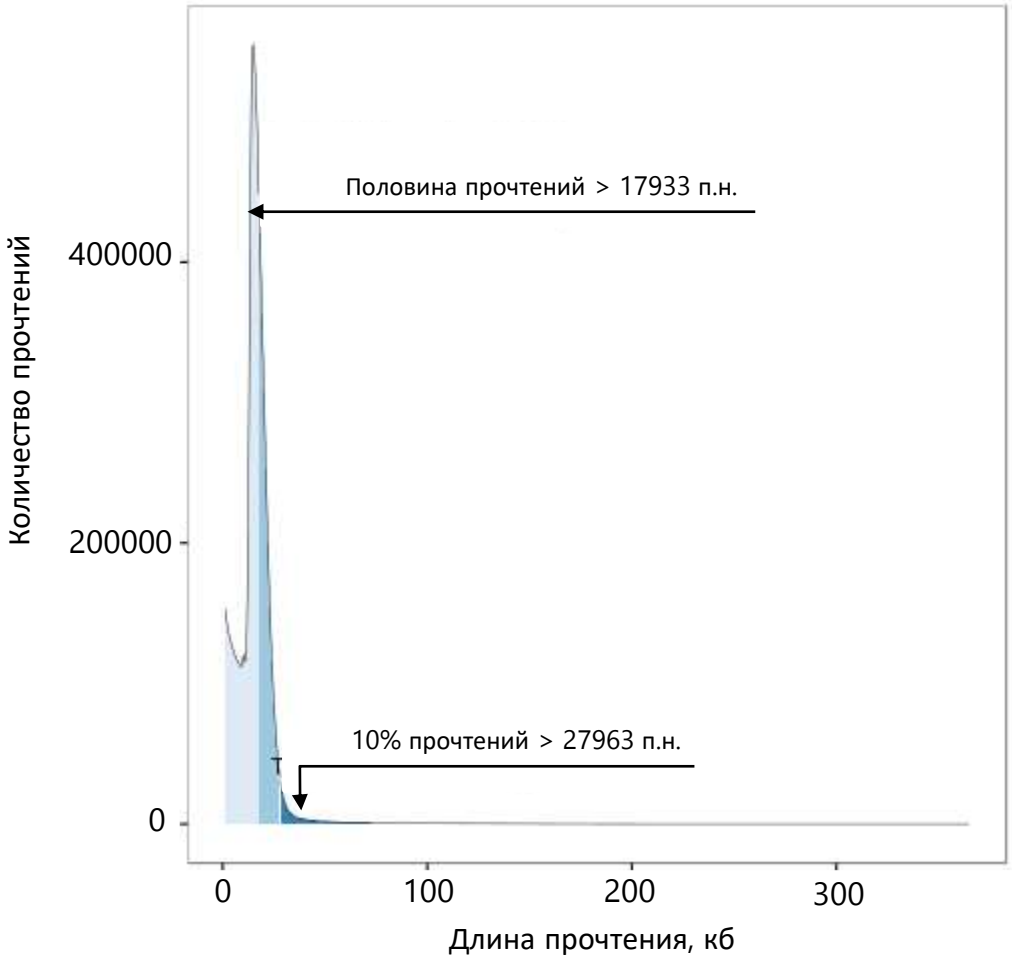

Сборка на основе IPA (1n)
Сборка на основе Hifiasm (1n)
Сборка на основе hiCanu (1n)
Показатель Nx, Мб
35
30
25
20
15
10
5
0
10
20
30
40
50
60
70
80
90
100
Значение x, %

б

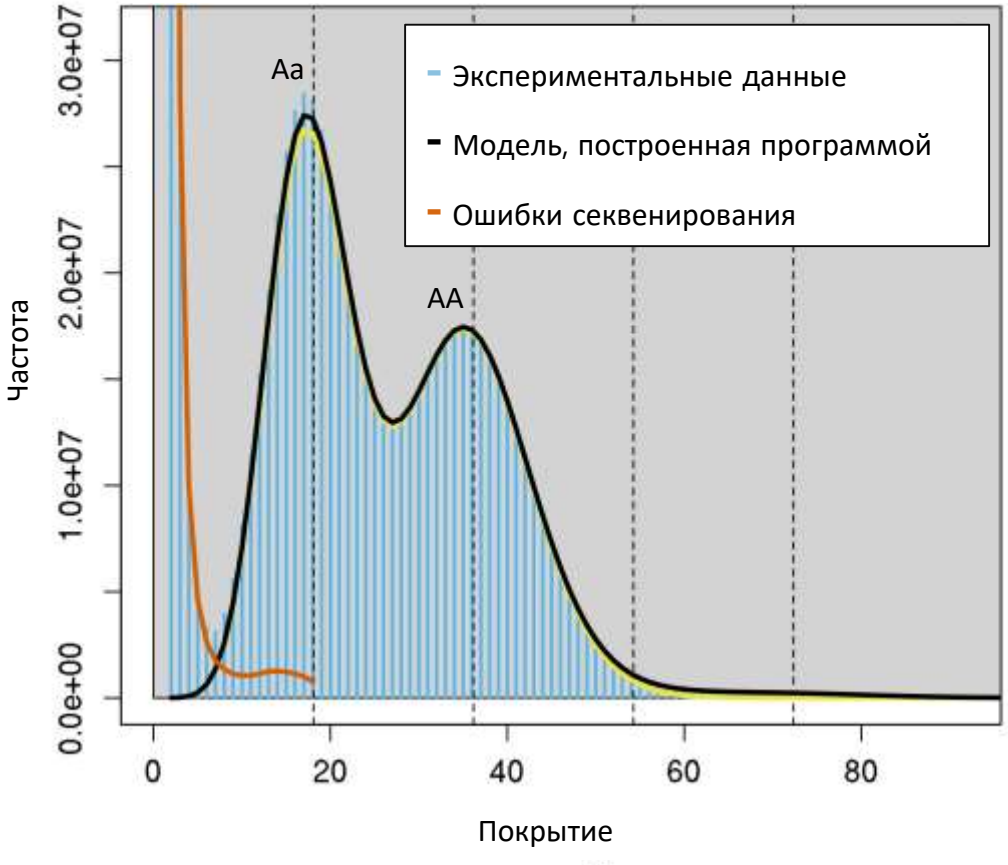

Сборка на основе IPA (1n)
Сборка на основе Hifiasm (1n)
Сборка на основе hiCanu (1n)
Показатель Nx, Мб
30
25
20
15
10
5
0
10
20
30
40
50
60
70
80
90
100
Значение x, %

б

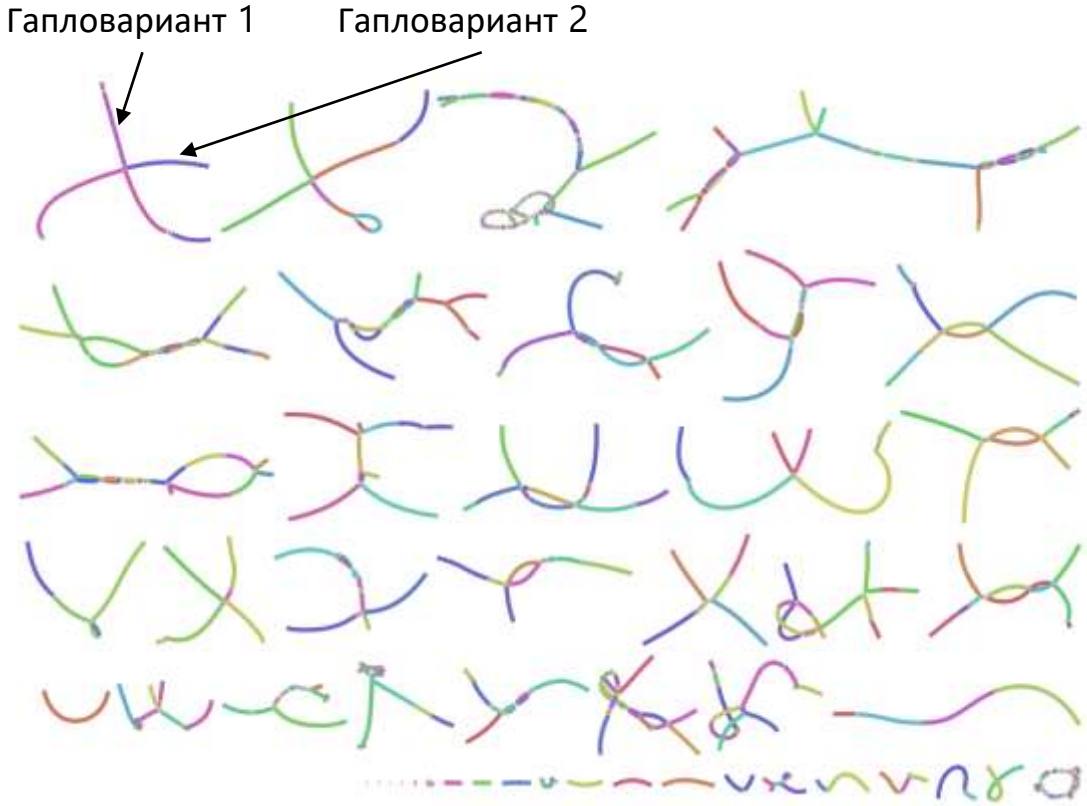

Сборка на основе IPA (2n)
Сборка на основе Hifiasm (2n)
Сборка на основе hiCanu (2n)
Показатель Nx, Мб
30
25
20
15
10
5
0
10
20
30
40
50
60
70
80
90
100
Значение x, %

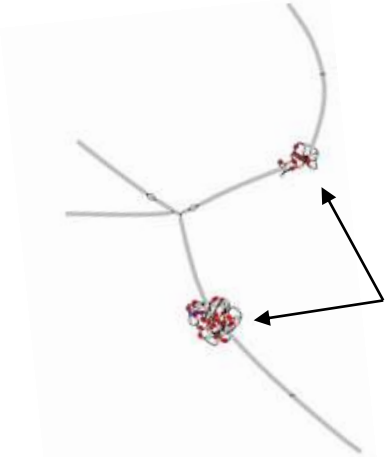

Половина прочтений > 17933 п.н.
10% прочтений > 27963 п.н.
Количество прочтений
400000
200000
0
0
100
200
300
Длина прочтения, кб

Длина: 507831422 п.н., уникальные последовательности: 90,6%,
гомозиготность: 99%, гетерозиготность 0,962%,
покрытие k-мерами: 18,1, ошибки: 0,0694%,
размер k-мера: 45, плоидность: 2

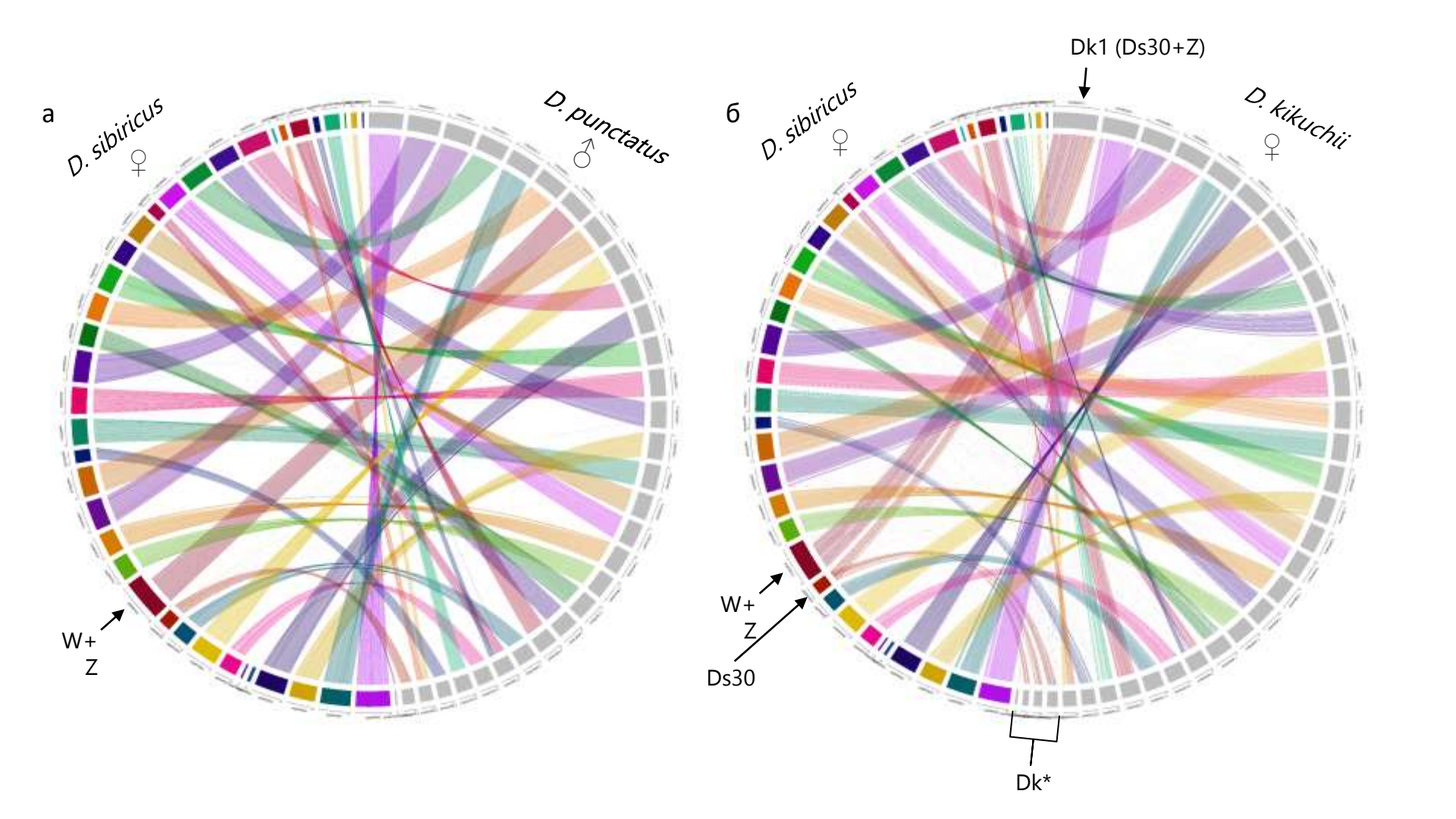


Визуализация полученного графа сборки





Визуализация сборки генома в графах юнитигов (обозначены случайным цветом).
Большие графы соответствуют хромосомам


а
D. sibiricus
♀
D. punctatus
♂
W+
Z
б
Dk1 (Ds30+Z)
D. sibiricus
♀
D. kikuchii
♀
W+
Z
Ds30
Dk*