

# *De novo* сборка генома

# В идеальном мире



ДНК



Портативный USB3  
секвенатор



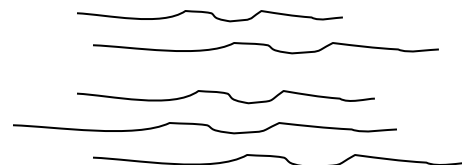
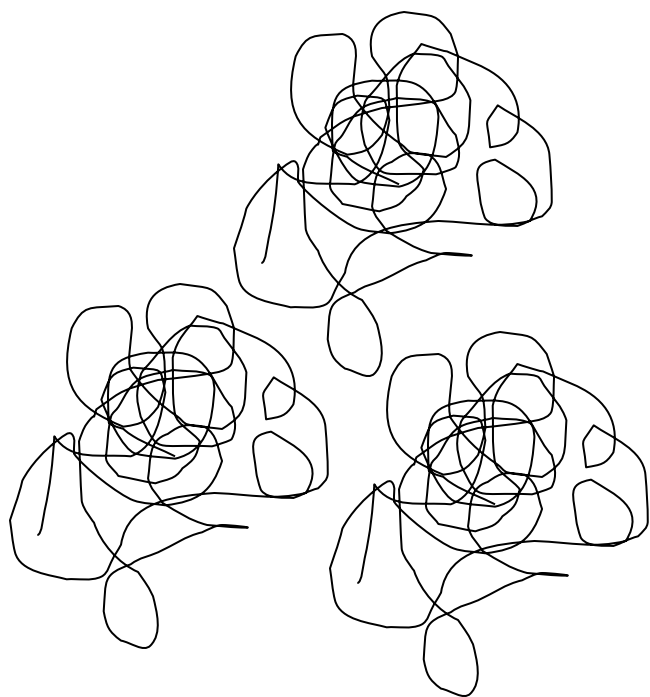
AGTCTAGGATTCGCTA  
TAGATTCAGGCTCTGA  
TATATTCGCGGGATT  
AGCTAGATCGCTATGC  
TATGATCTAGATCTCG  
AGATTCGTATAAGTCT  
AGGATTCGCTATAGAT  
TCAGGCTCTGATATAT  
TTCGCGGGATTAGCTA

Полные  
последовательности  
хромосом

# WGS секвенирование

Несколько копий ДНК молекул

**Фрагменты длиной 200 -  
200,000 п.н.**



Не остается  
информации из какой  
части генома взят тот  
или иной фрагмент

# Технологии секвенирования

- 1-е поколение

Sanger sequencing



- 2-е поколение



illumina



ion torrent



AB applied biosystems



- 3-е поколение

Helicos  
BioSciences Corporation



PACIFIC  
BIOSCIENCES



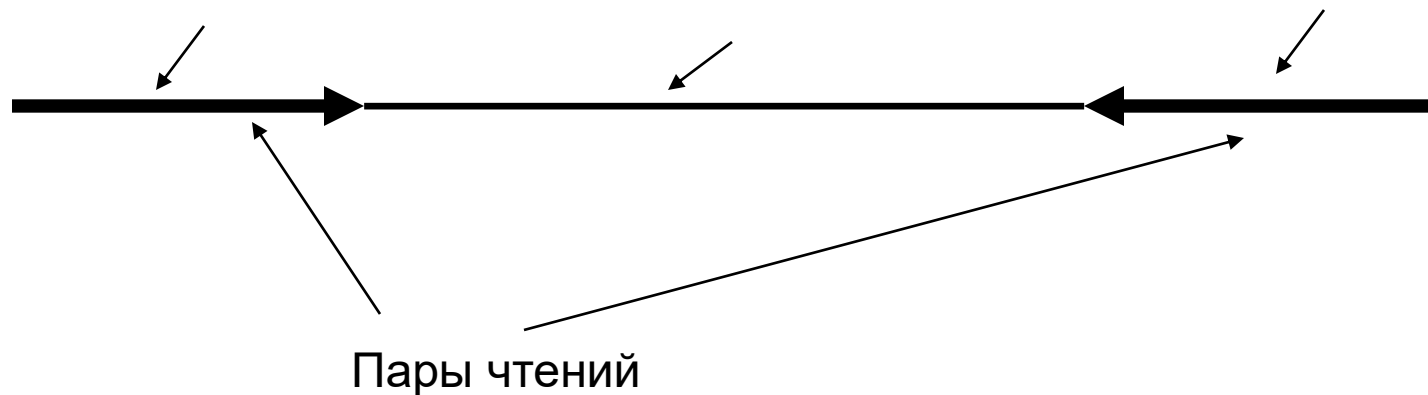
Oxford  
NANOPORE  
Technologies



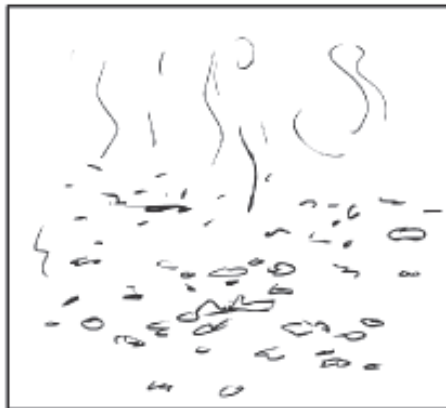
# WGS секвенирование: фрагменты

Секвенатор считывает по **100-1000 п.н.** с конца/концов фрагмента. Размеры фрагментов известны с точностью  **$\pm 10-20\%$** .

CAAGCTGAT... Неизвестная последовательность...GTTTGGAAC



# Сборка генома



# Пушкиномика

У лукоморья  
дуб зеленый;  
Златая цепь  
на дубе том:



# Пушкиномика

## Чтения:

У лукоморья дуб  
морья дуб зеле  
зуб зеленый; З

лenny; Златая  
Златая цепь н  
я цепь на дубе  
пъ на дубе том





# Пушкиномика

## Чтения:

У лукоморья дуб  
морья дуб зеле  
зуб зеленый; З  
лenny; Златая  
Златая цепь н  
я цепь на дубе  
пь на дубе том

## Перекрытия:

У лукоморья дуб  
морья дуб зеле  
зуб зеленый; З  
лenny; Златая  
Златая цепь н  
я цепь на дубе  
пь на дубе том



# Пушкиномика

## Перекрытия:

У лукоморья **ду**

морья **дуб** зеле

**з**уб зеленый; З

лenny; Златая

Златая цепь н

я цепь на дубе

пъ на дубе том



# Пушкиномика

## Перекрытия:

У лукоморья **ду**

морья **дуб** зеле

**зуб** зеленый; З

лenny; Златая

Златая цепь н

я цепь на дубе

пъ на дубе том



## Консенсус:

У лукоморья **дуб** зеленый; Златая цепь на дубе том

# Overlap graph

AGCTACAGTATGCT

TACAGTATGCTTAT

GTATGCTTATCTGA

TGATACCTTAGCCA

TGCTTATCTGATAC

# Overlap graph

TACAGTATGCTTAT

AGCTACAGTATGCT

TACAGTATGCTTAT

TACAGTATGCTTAT

GTATGCTTATCTGA

TACAGTATGCTTAT

TGCTTATCTGATAC

# Overlap graph

AGCTACAGTATGCT



AGCTACAGTATGCT

TACAGTATGCTTAT



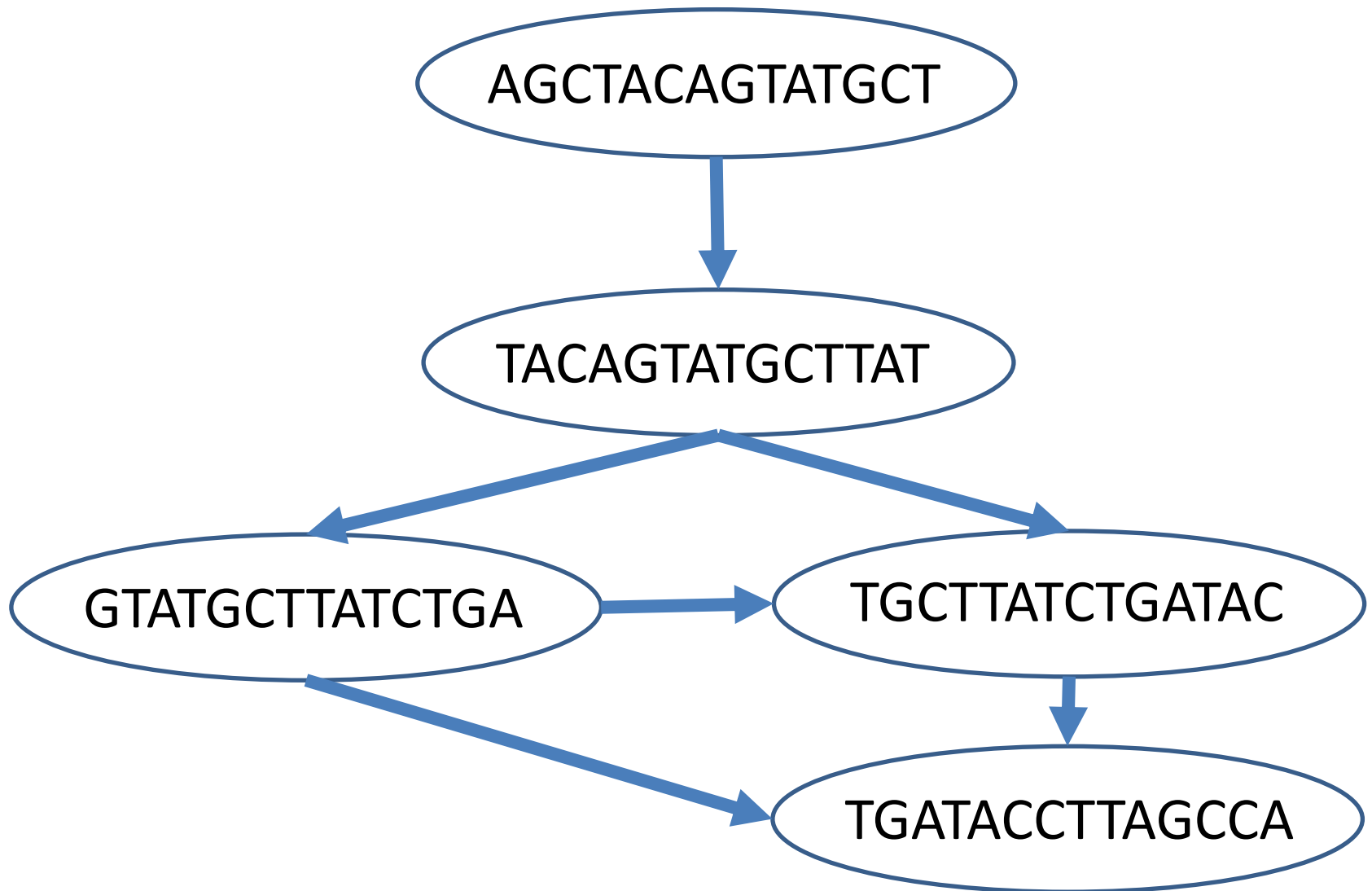
TACAGTATGCTTAT

# Overlap graph

AGCTACAGTATGCT  
TACAGTATGCTTAT

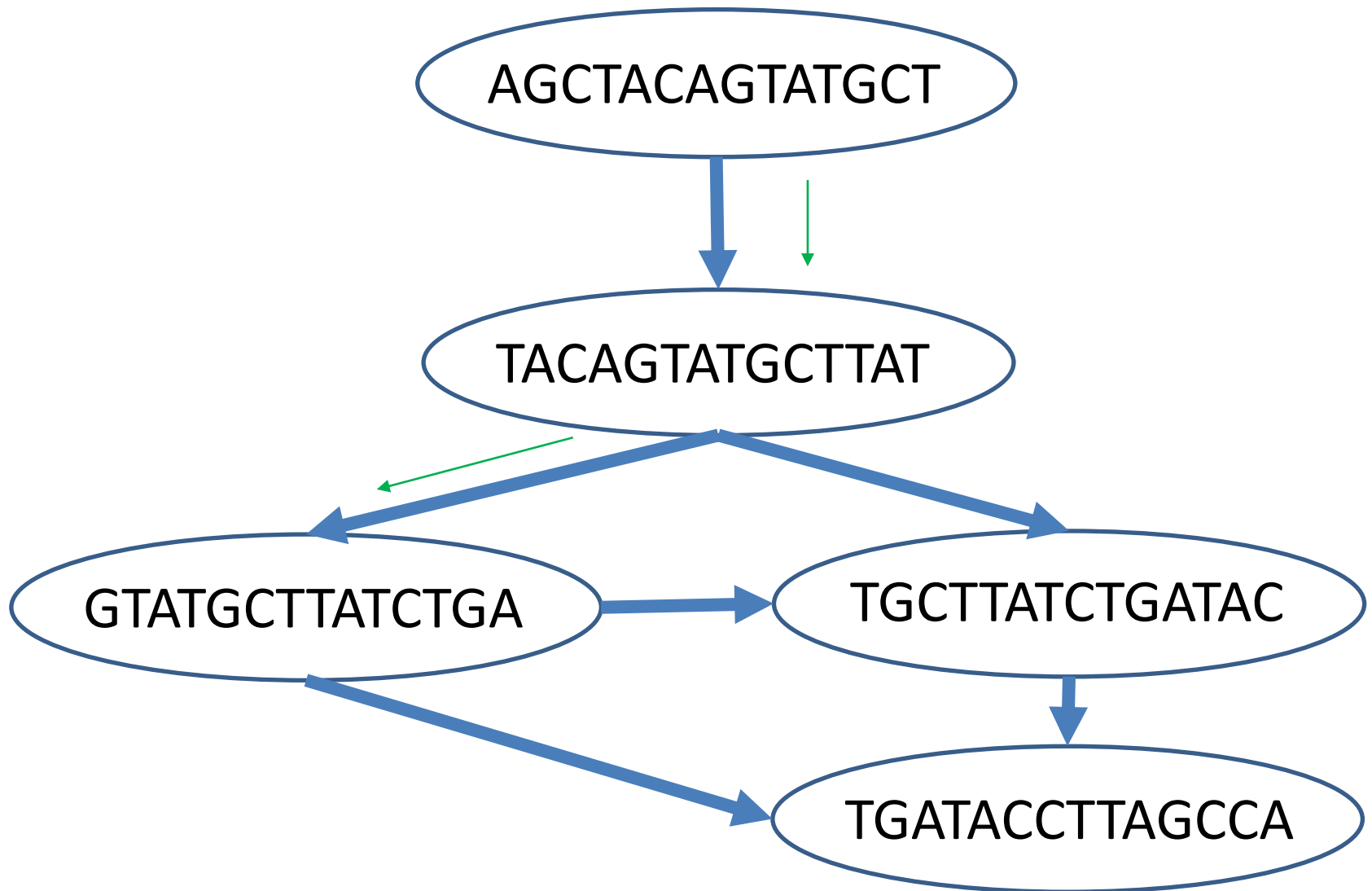


# Overlap graph

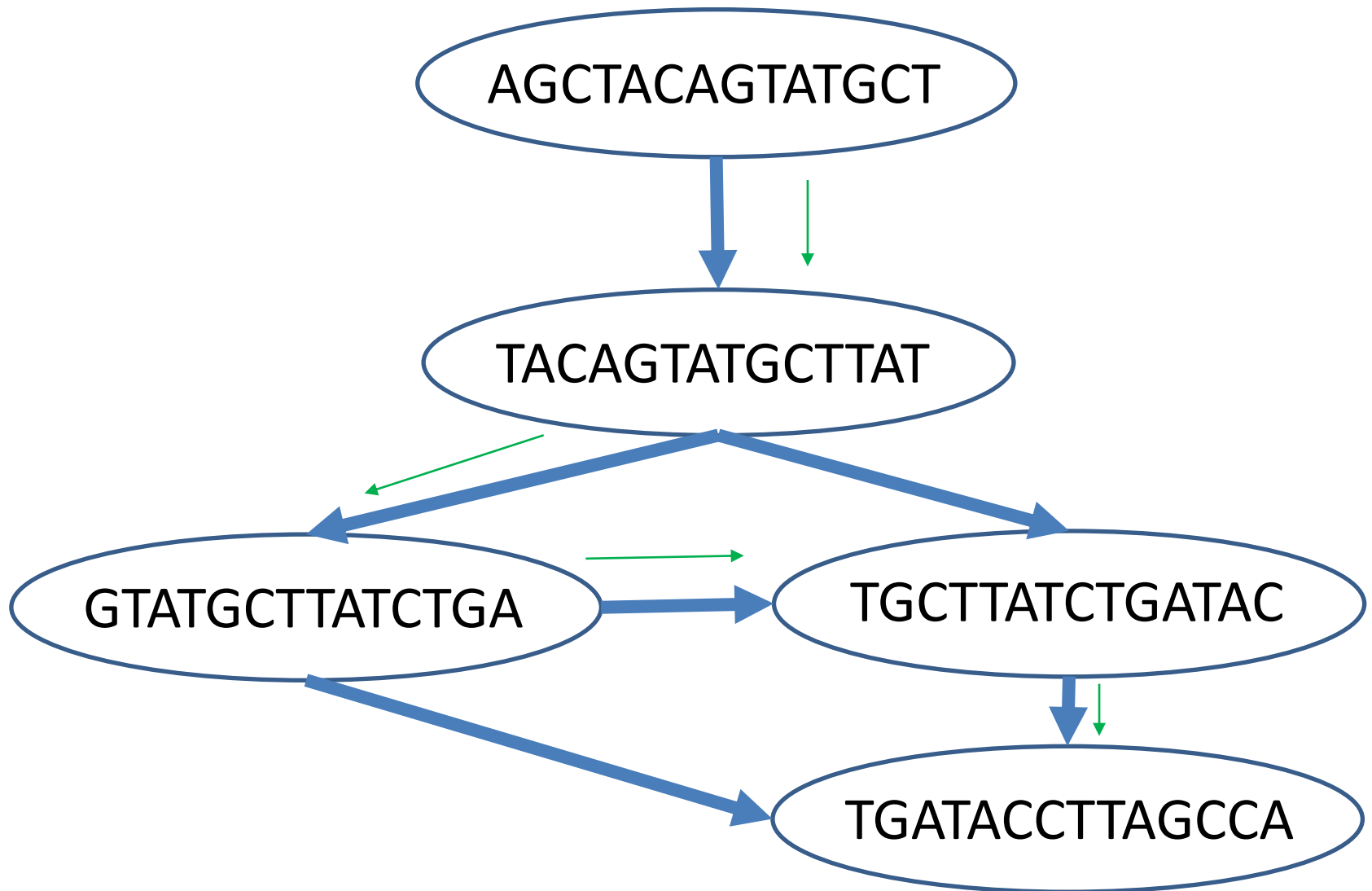




# Overlap graph



# Overlap graph



# Overlap graph

AGCTACAGTATGCT

TACAGTATGCTTAT

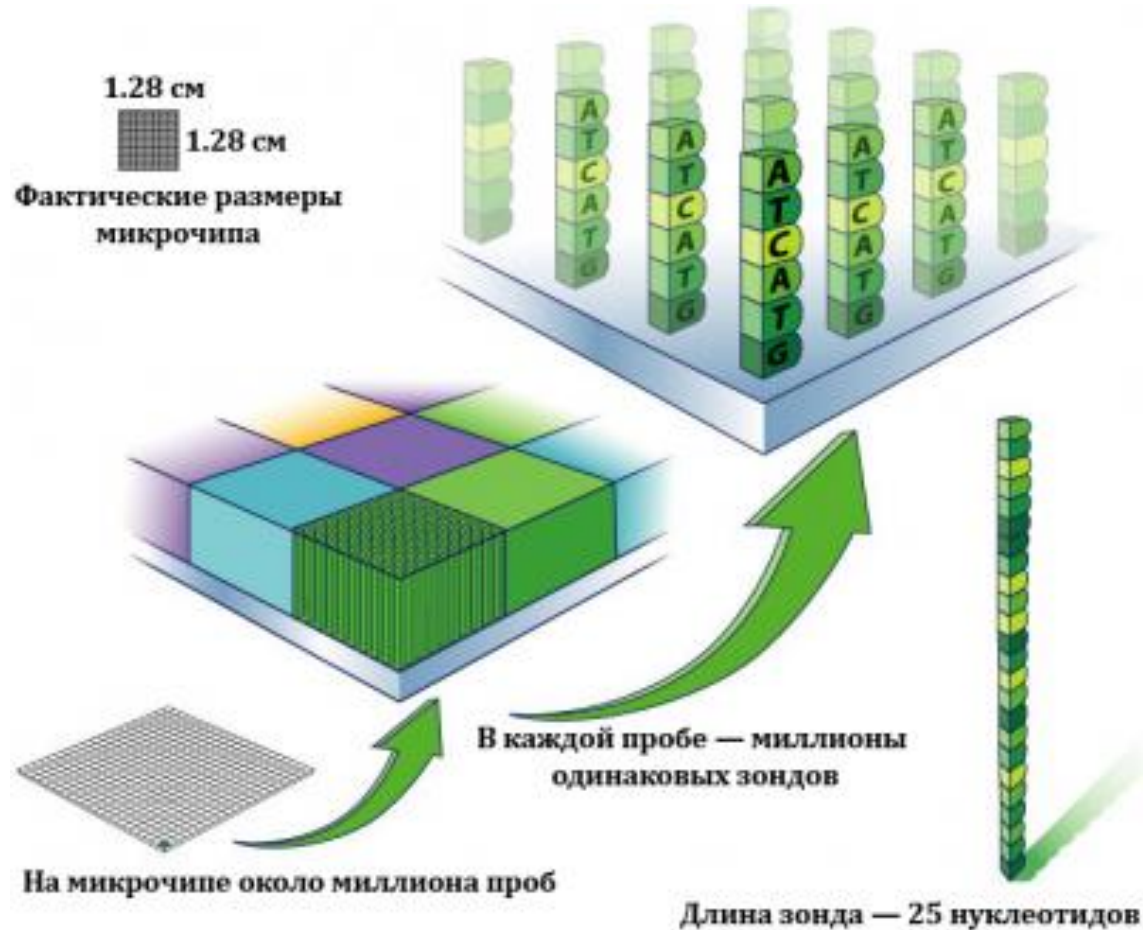
GTATGCTTATCTGA

TGCTTATCTGATAC

TGATACCTTAGCCA

AGCTACAGTATGCTATCTGATACCTTAGCCA

# Секвенирование с помощью гибридизации



# Секвенирование с помощью гибридизации

- ДНК чип с  $4^3$  пробами  
Целевая ДНК: **AAATGCG**

AAA↵	AAC↵	AAG↵	AAT↵	ACA↵	ACC↵	ACG↵	ACT↵	↵
ATT↵	ATG↵	ATC↵	ATA↵	AGG↵	AGT↵	AGC↵	AGA↵	↵
CCC↵	CCA↵	CCG↵	CCT↵	CAA↵	CAC↵	CAG↵	CAT↵	↵
CTC↵	CTG↵	CTA↵	CTT↵	CGA↵	CGC↵	CGG↵	CGT↵	↵
GGA↵	GGC↵	GGT↵	GGG↵	GAA↵	GAT↵	GAC↵	GAG↵	↵
GTT↵	GTG↵	GTC↵	GTA↵	GCG↵	GCT↵	GCC↵	GCA↵	↵
TTA↵	TTC↵	TTG↵	TTT↵	TAA↵	TAC↵	TAG↵	TAT↵	↵
TGT↵	TGG↵	TGC↵	TGA↵	TCC↵	TCA↵	TCG↵	TCT↵	↵

# K-меры. De Bruijn граф.

K=5

AGCTACAGTATGC

AGCTA

# К-меры. De Bruijn граф.

K=5

AGCTACAGTATGC



AGCTA    GCTAC

# К-меры. De Bruijn граф.

K=5

AGCTACAGTATGC



AGCTA    GCTAC    CTACA



# К-меры. De Bruijn граф.

K=5

AGCTACAGTATGC



AGCTA    GCTAC    CTACA

TACAG

# К-меры. De Bruijn граф.

K=5

AGCTACAGTATGC



AGCTA    GCTAC    CTACA

TACAG    ACAGT

# К-меры. De Bruijn граф.

K=5

AGCTACAGTATGC



AGCTA    GCTAC    CTACA

TACAG    ACAGT    CAGTA

# К-меры. De Bruijn граф.

K=5

AGCTACAGTATGC



AGCTA    GCTAC    CTACA

TACAG    ACAGT    CAGTA

AGTAT

# К-меры. De Bruijn граф.

K=5

AGCTACAGTATGC



AGCTA    GCTAC    CTACA

TACAG    ACAGT    CAGTA

AGTAT    GTATG

# К-меры. De Bruijn граф.

K=5

AGCTACAGTATGC

AGCTA    GCTAC    CTACA

TACAG    ACAGT    CAGTA

AGTAT    GTATG    TATGC

К-меры. De Bruijn граф.

AGCTACAGTATGC

TATGCTTATCTGA

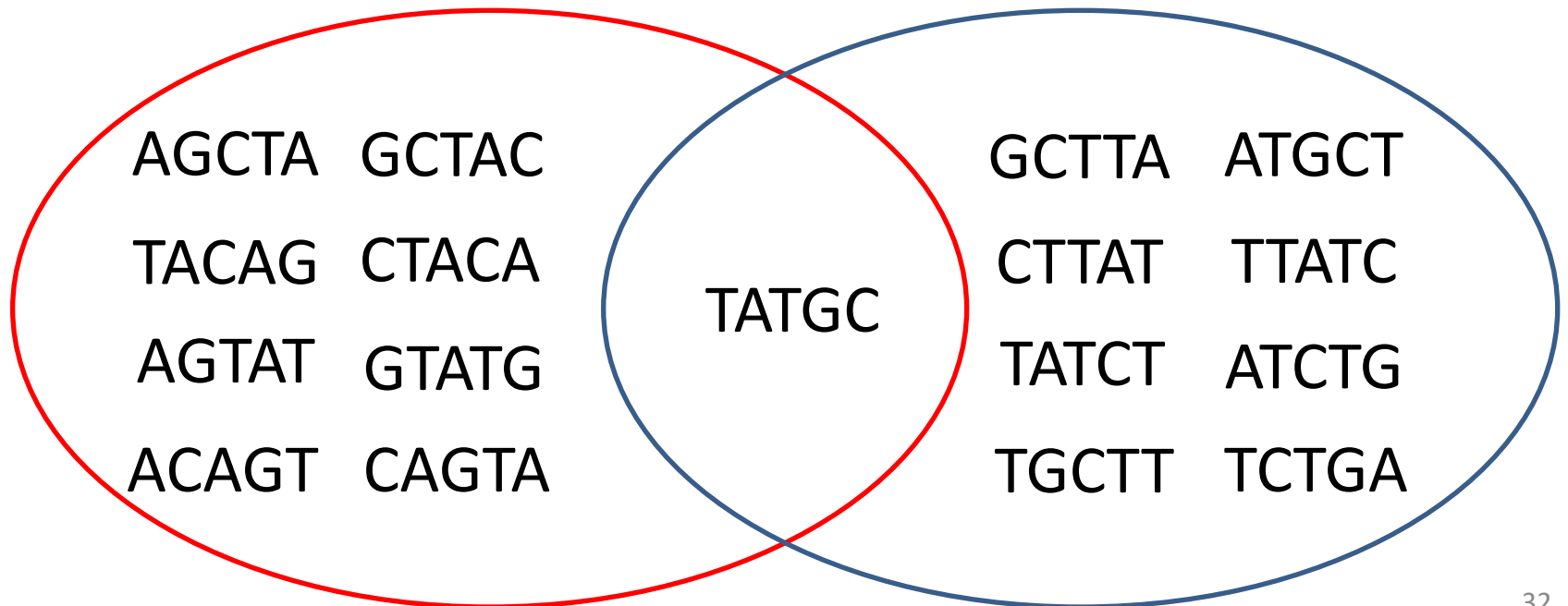
# К-меры. De Bruijn граф.

AGCTACAGTATGC

TATGCTTATCTGA

AGCTACAGTATGC

TATGCTTATCTGA





# К-меры. De Bruijn граф.

$$K+1=6$$

AGCTACAGTATGC

AGCTA    GCTAC    CTACA

AGCTAC

TACAG    ACAGT    CAGTA

AGTAT    GTATG    TATGC

# К-меры. De Bruijn граф.

$$K+1=6$$

AGCTACAGTATGC

AGCTA GCTAC CTACA

AGCTAC GCTACA

TACAG ACAGT CAGTA

AGTAT GTATG TATGC

# К-меры. De Bruijn граф.

$$K+1=6$$

AGCTACAGTATGC

AGCTA GCTAC CTACA

AGCTAC GCTACA CTACAG

TACAG ACAGT CAGTA

TACAGT ACAGTA CAGTAT

AGTAT GTATG TATGC

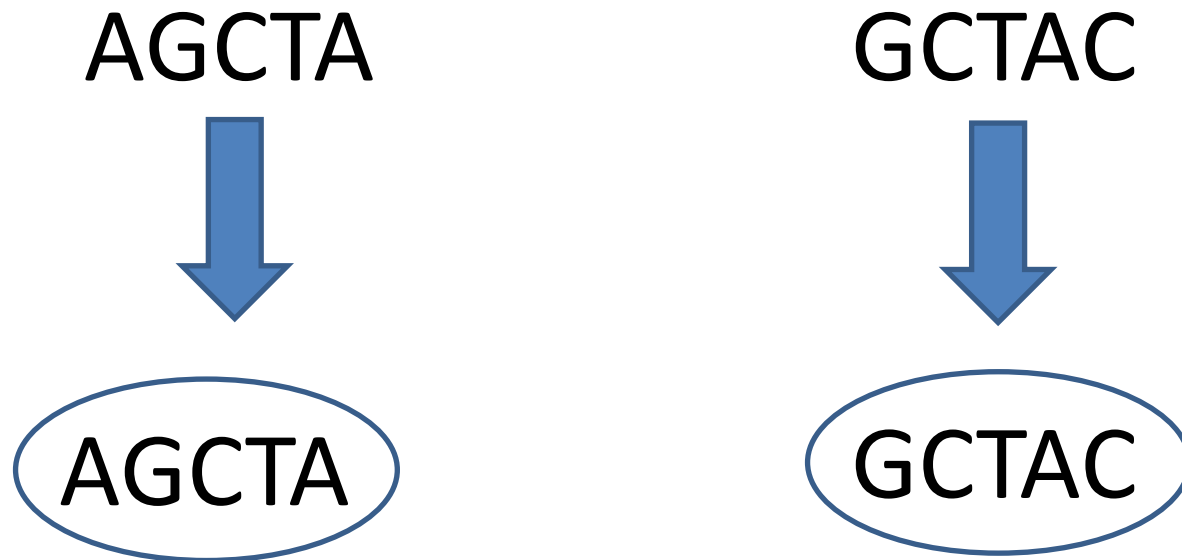
AGAGTG GTATGC

# К-меры. De Bruijn граф.

AGCTA

GCTAC

# К-меры. De Bruijn граф.



# К-меры. De Bruijn граф.

AGCTA  
AGCTAC  
GCTAC

AGCTA

GCTAC

# К-меры. De Bruijn граф.

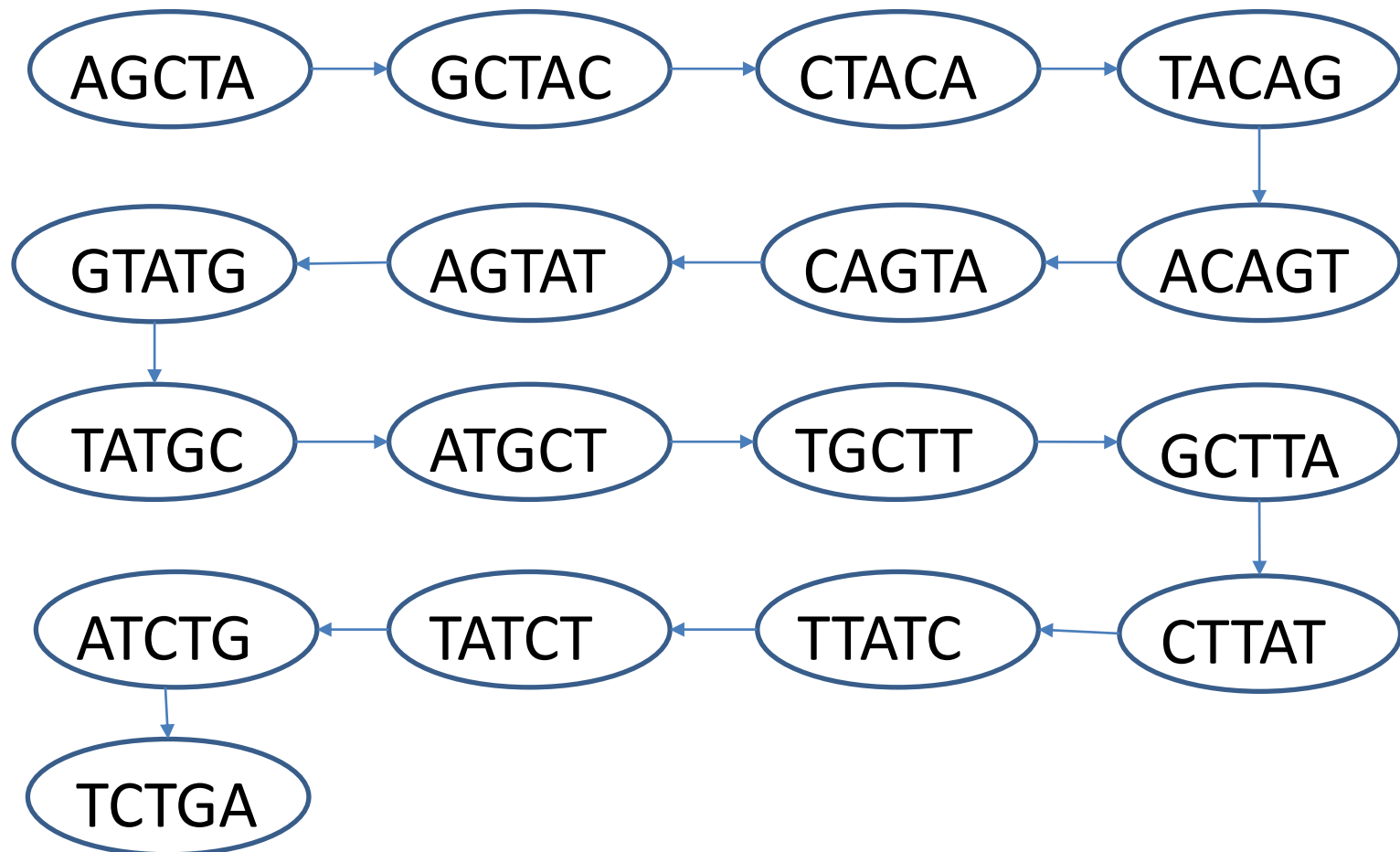
AGCTA  
AGCTAC  
GCTAC



# К-меры. De Bruijn граф.

K=5

AGCTACAGTATGC  
TATGCTTATCTGA

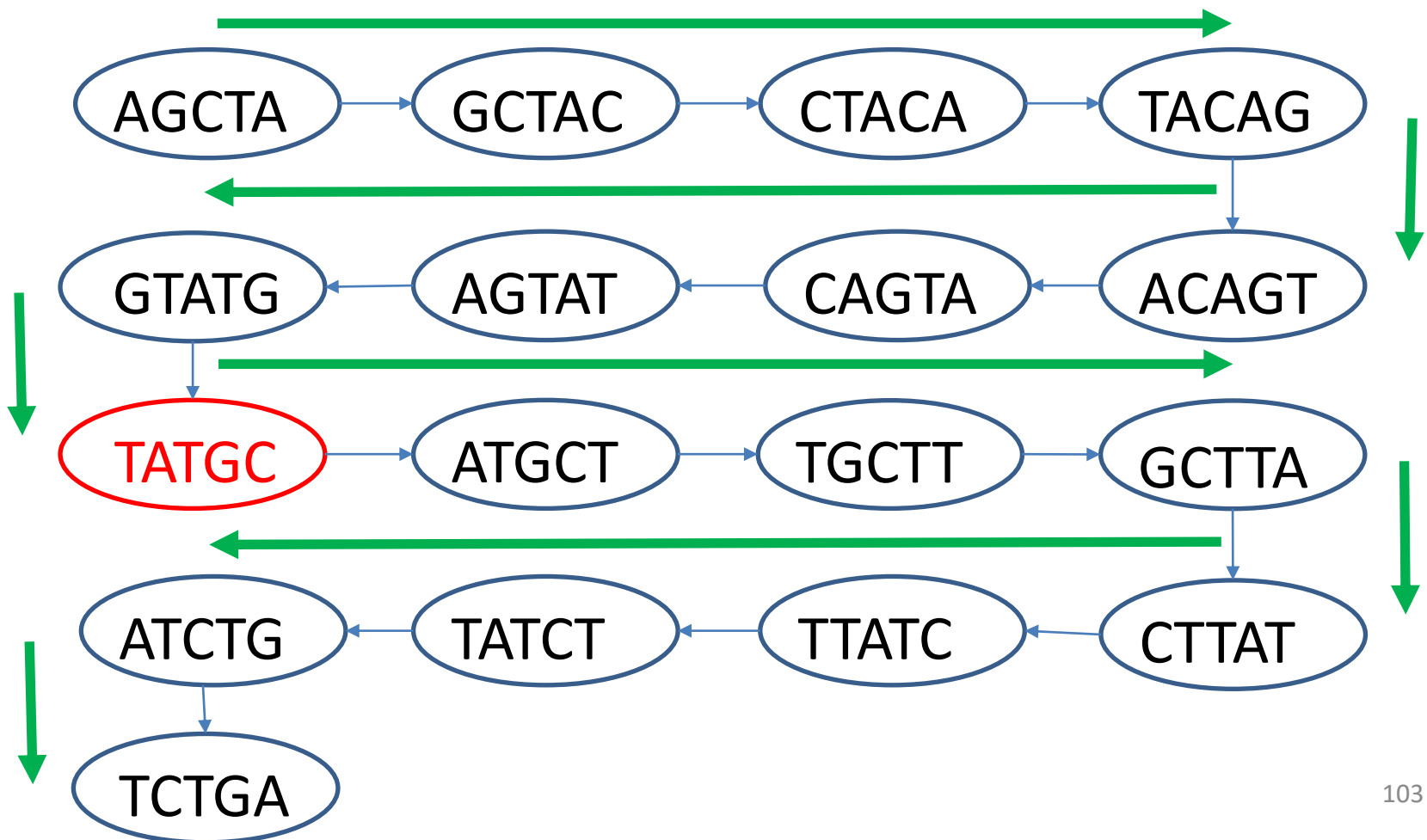




# K-меры. De Bruijn граф.

K=5

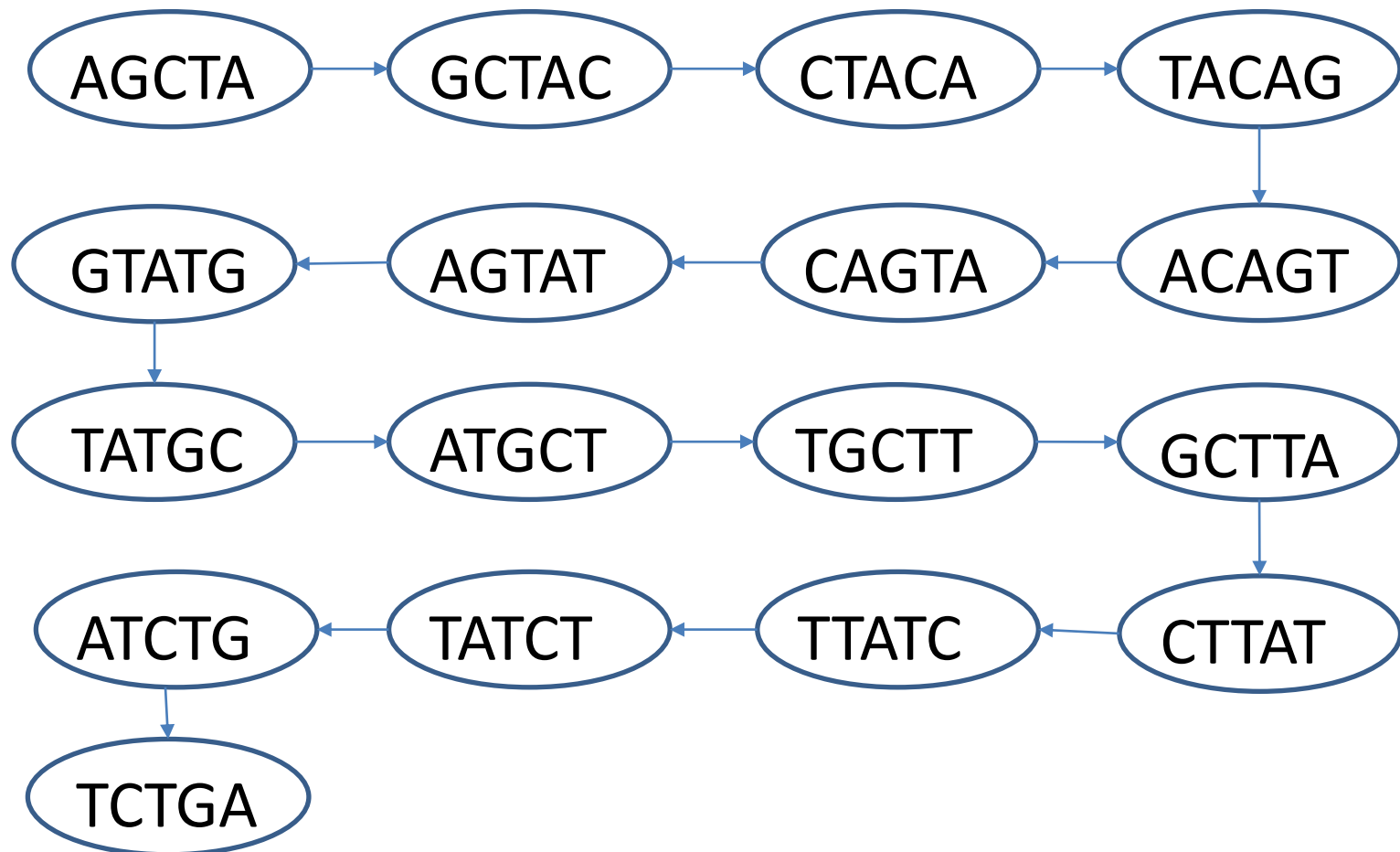
AGCTACAGTATGC  
TATGCTTATCTGA



# K-меры. De Bruijn граф.

K=5

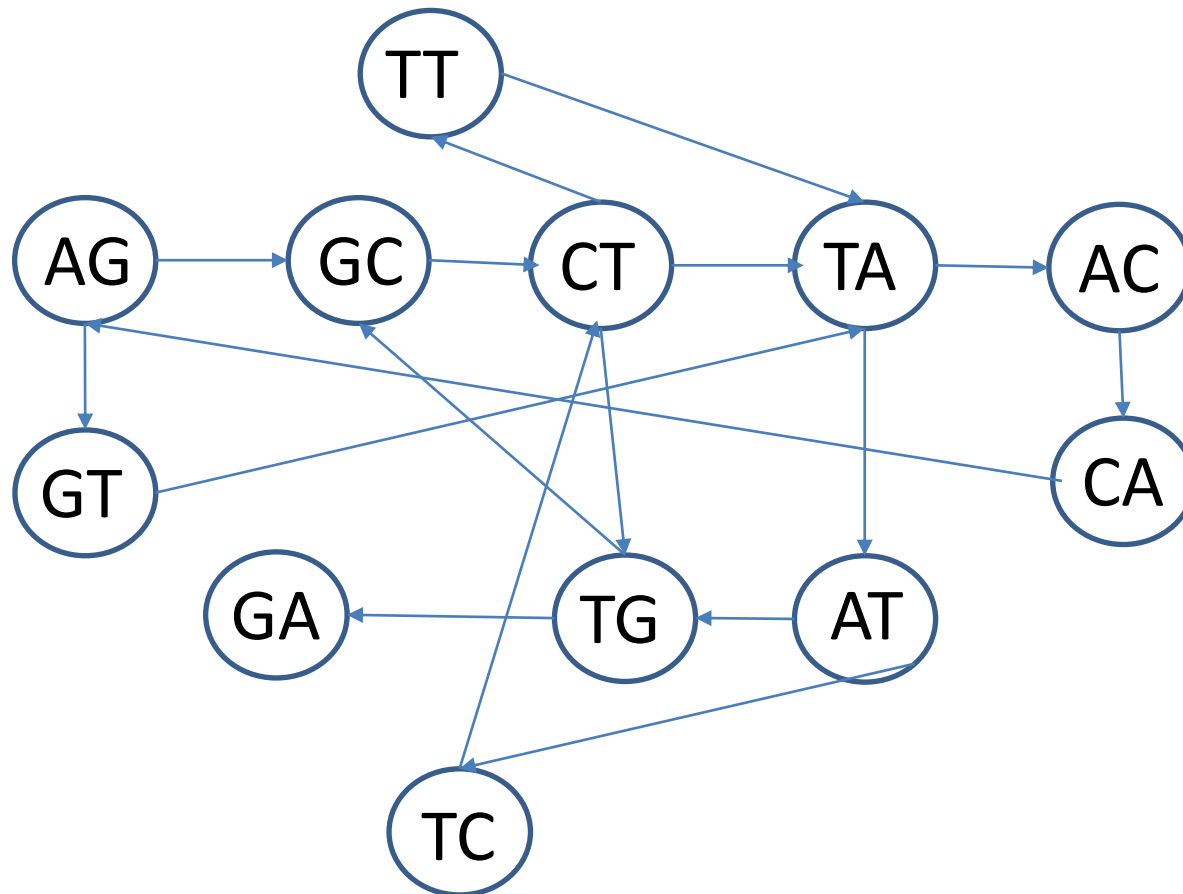
AGCTACAGTATGC  
TATGCTTATCTGA



# К-меры. De Bruijn граф.

K=2

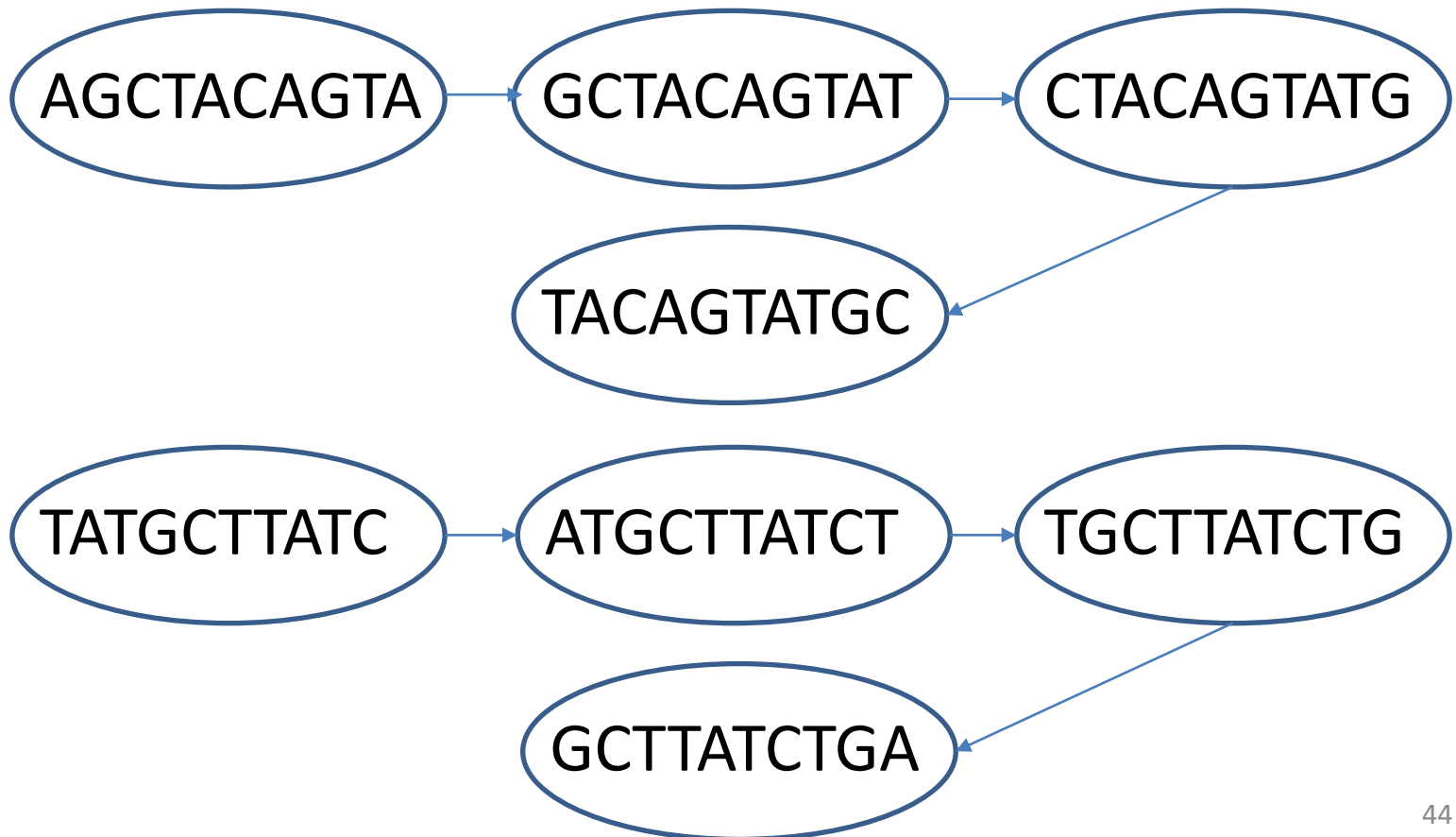
AGCTACAGTATGC  
TATGCTTATCTGA



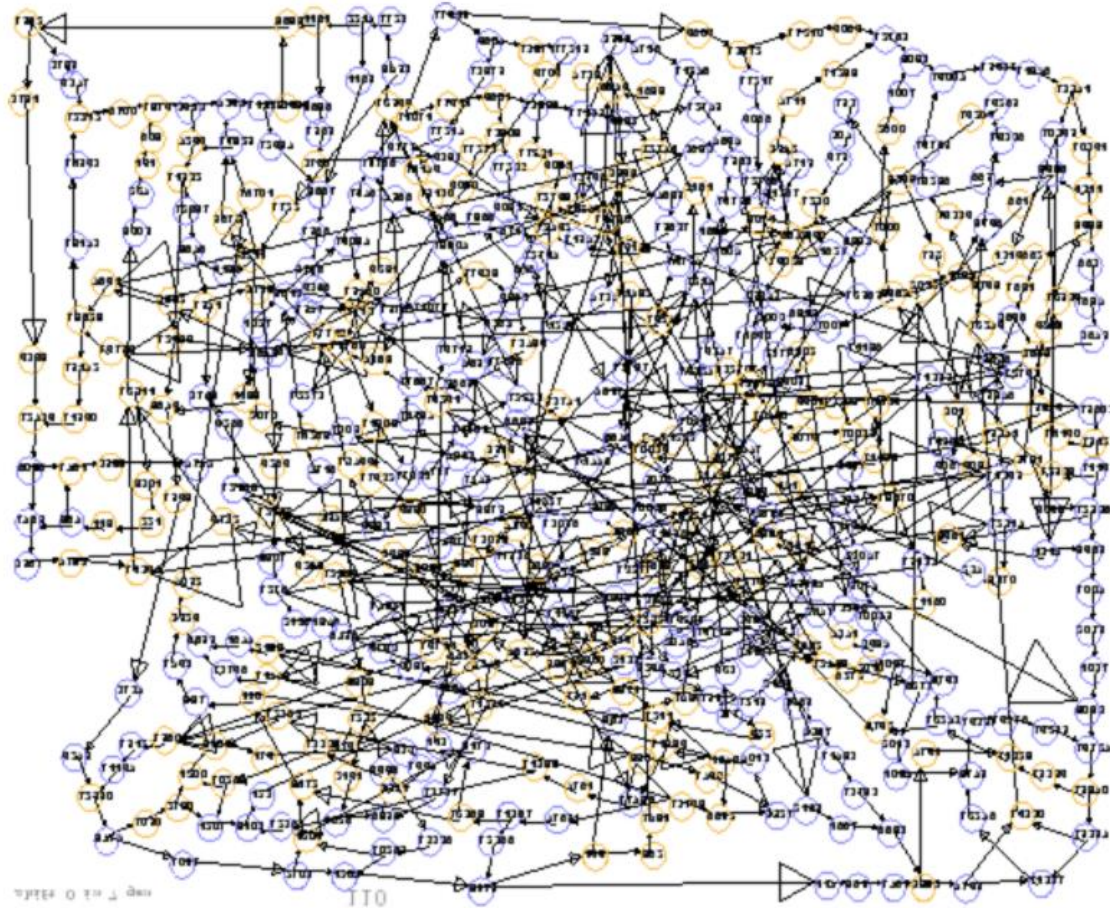
# K-мер. De Bruijn граф.

K=10

AGCTACAGTATGC  
TATGCTTATCTGA

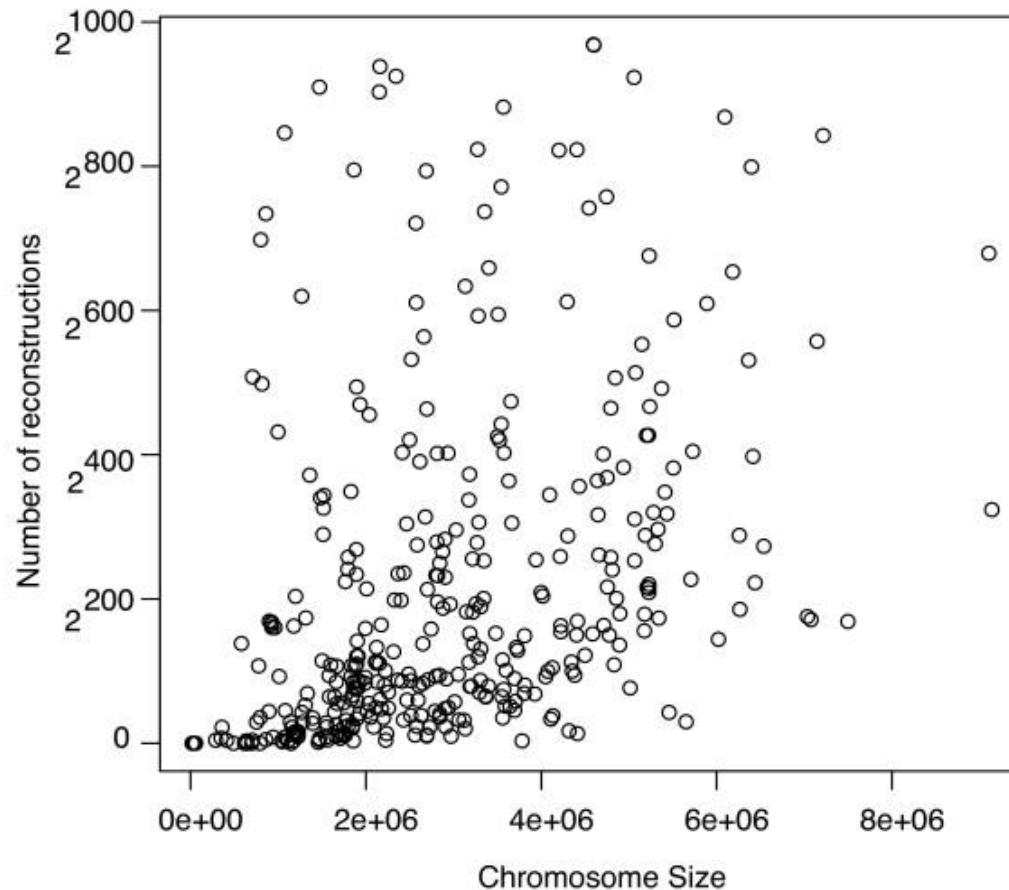


# Более реалистичный пример графа



[[http://bioinformatics.org.au/ws13/wp-content/uploads/ws13/sites/3/FullPresentations/Torsten-Seemann\\_2013-Winter-School-presentation.pdf](http://bioinformatics.org.au/ws13/wp-content/uploads/ws13/sites/3/FullPresentations/Torsten-Seemann_2013-Winter-School-presentation.pdf)]

# Число возможных реконструкций генома



[Kingsford C. et al. Assembly complexity of prokaryotic genomes using short reads. BMC Bioinformatics. 2010 Jan 12;11:21]

# Что усложняет графы



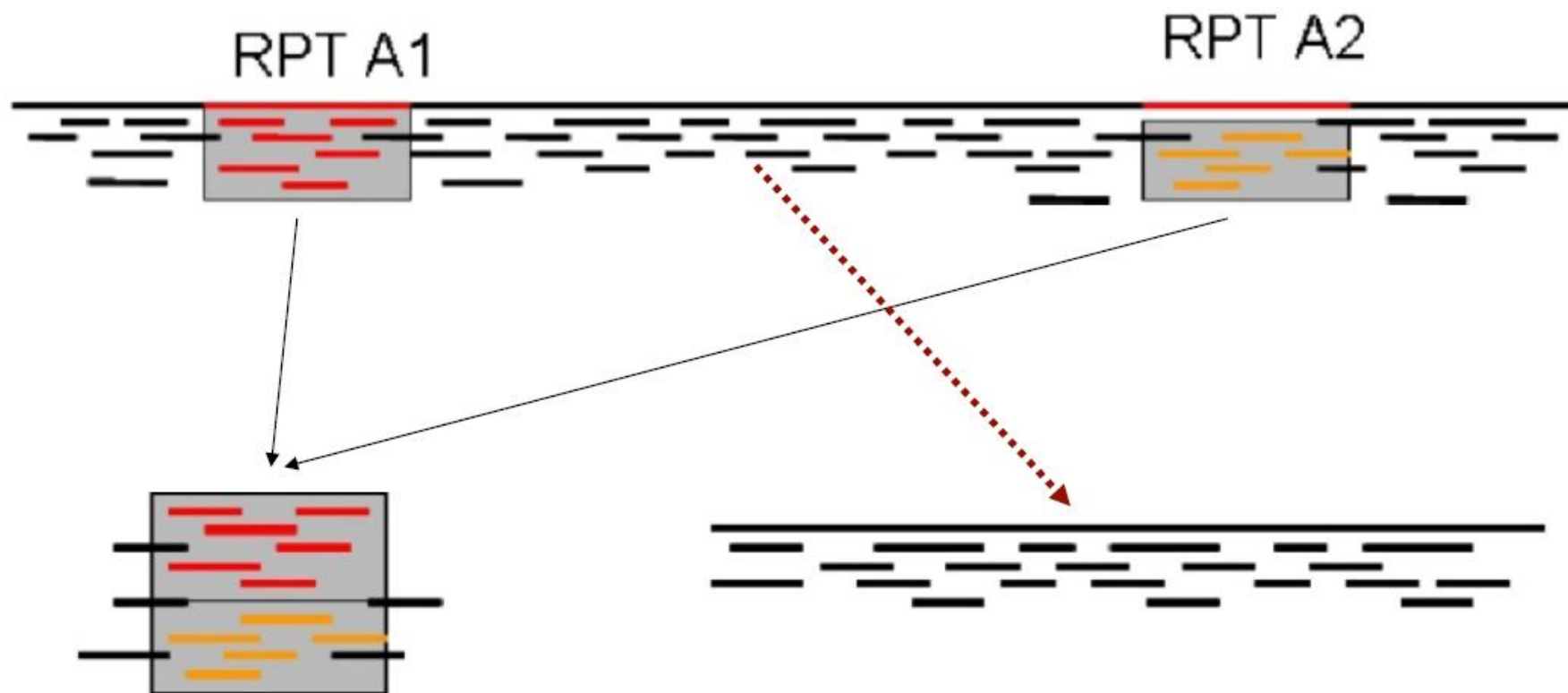
- **Ошибки в чтениях**
  - Приводят к появлению в графе ошибочных ребер и вершин.
- **Диплоидные и полиплоидные организмы**
  - Приводит к появлению дополнительных путей в графе
- **Повторы**

# Что такое повтор?

- **Участок ДНК, который встречается более одного раза в геномной последовательности.**
- Наиболее частые
  - Транспозоны
  - Сателлитные повторы
  - Дублицированные гены(паралоги)



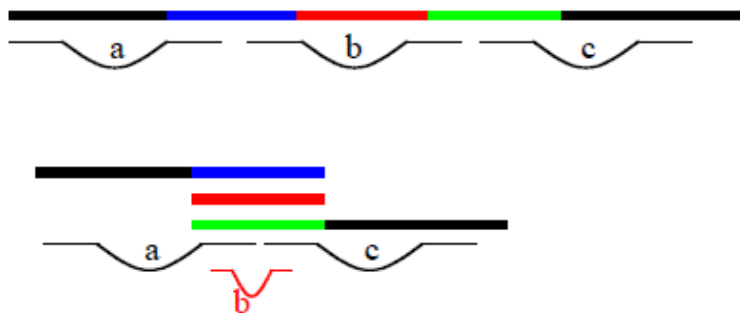
# Эффект оказываемый на сборку



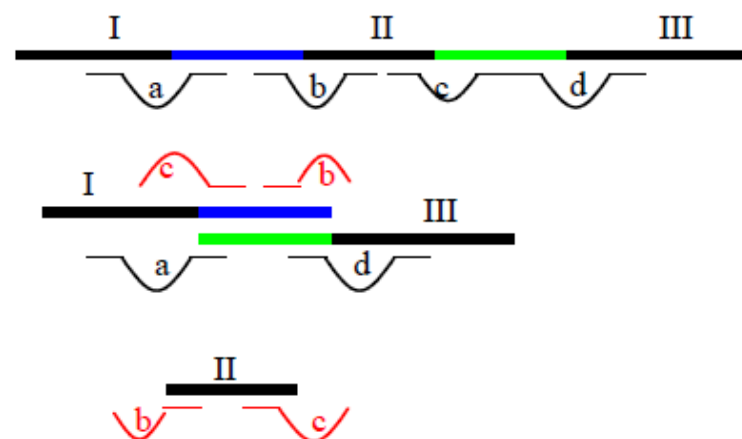
Повторяющиеся элементы сливаются в один контиг.

# Эффект оказываемый на сборку

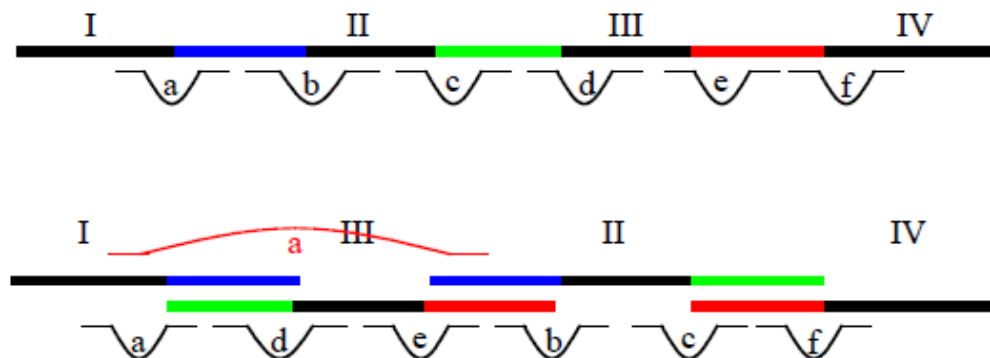
Слияние тандемных повторов



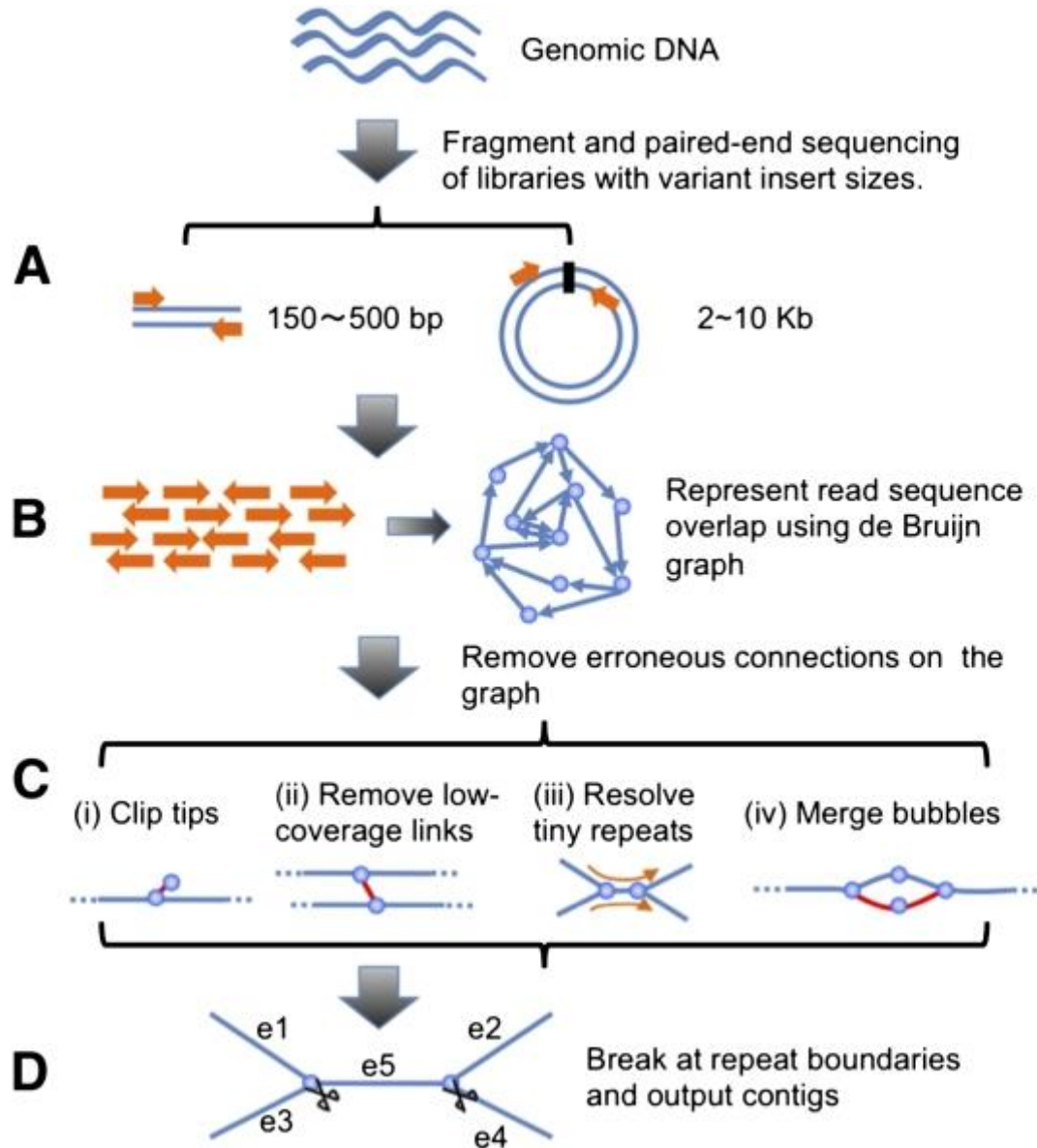
Исключение участка между повторами



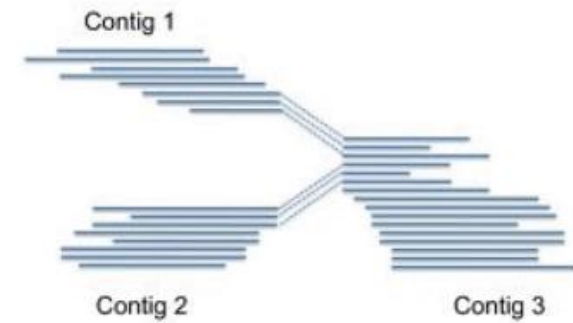
Перестановка



# SOAPdenovo



# КОНТИГИ



- **Непрерывные, однозначные фрагменты, собираемой ДНК последовательности.**
- **Концы контигов соответствуют**
  - **Настоящим концам**(для линейных ДНК молекул)
  - **Dead ends**(провалы покрытия)
  - **Точкам принятия решений**(узлам в графе в которые входит и/или выходит больше одного ребра)

[[http://bioinformatics.org.au/ws13/wp-content/uploads/ws13/sites/3/FullPresentations/Torsten-Seemann\\_2013-Winter-School-presentation.pdf](http://bioinformatics.org.au/ws13/wp-content/uploads/ws13/sites/3/FullPresentations/Torsten-Seemann_2013-Winter-School-presentation.pdf)]

# Скаффолдинг



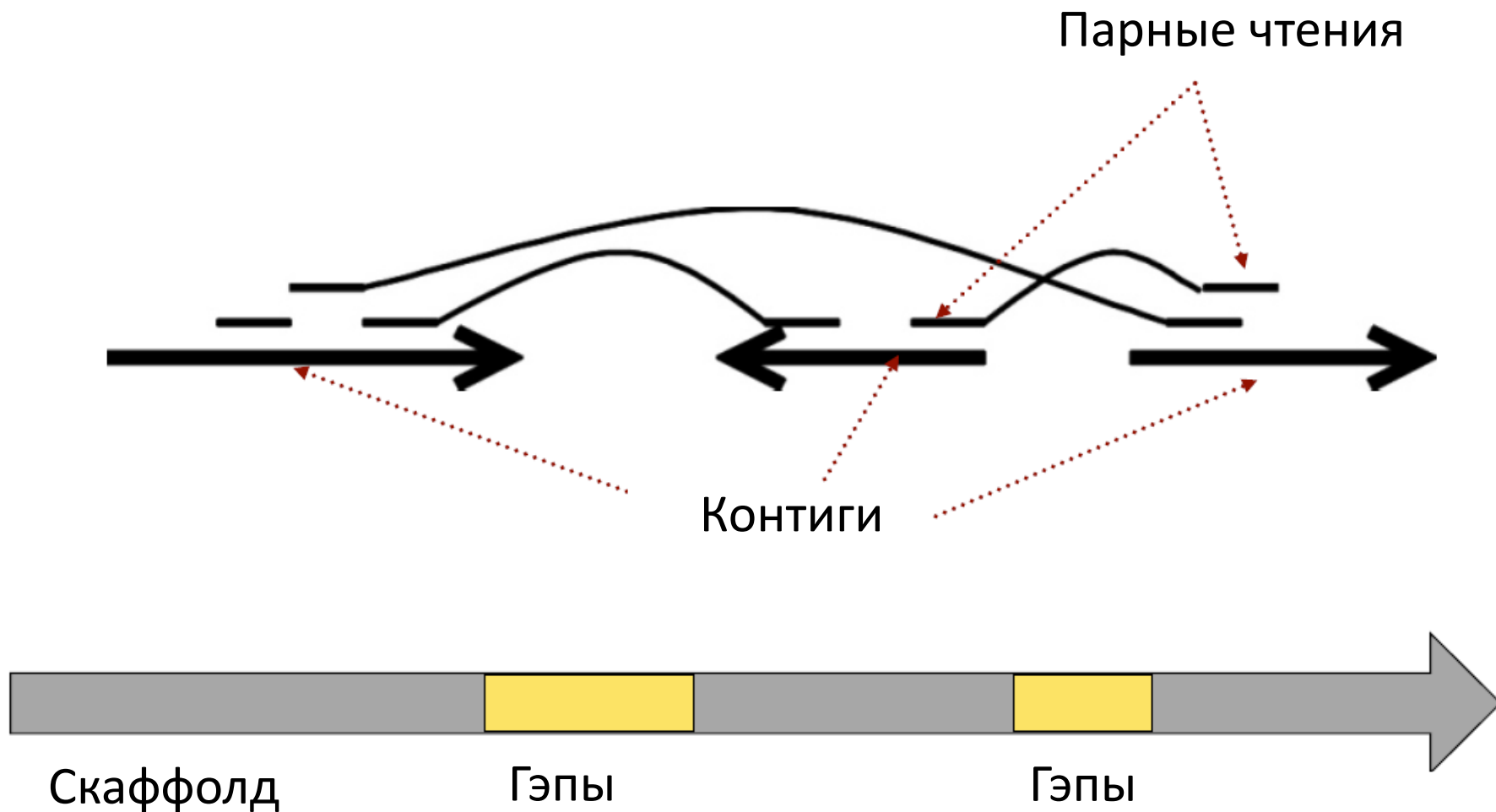
# Типы чтений

- Пример фрагмента
  - atcgtatgatcttgagattctctcttcccttatagctgctata
- Одноконцевое чтение
  - atcgtatgatcttgagattctctcttcccttatagctgctata
  - Последовательность с одного из концов
- Парноконцевое чтение
  - atcgtatgatcttgagattctctcttcccttatagctgctata
  - Последовательность с обоих концов
  - эту информацию можно использовать!

# Что такое длина вставки?



# От контигов к скаффолдам

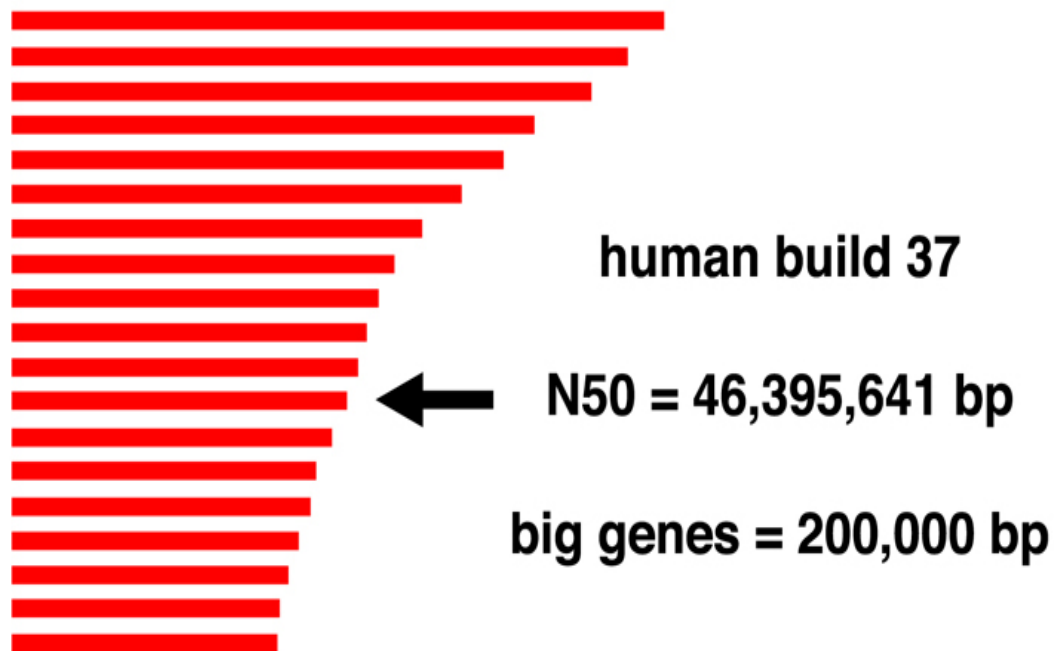




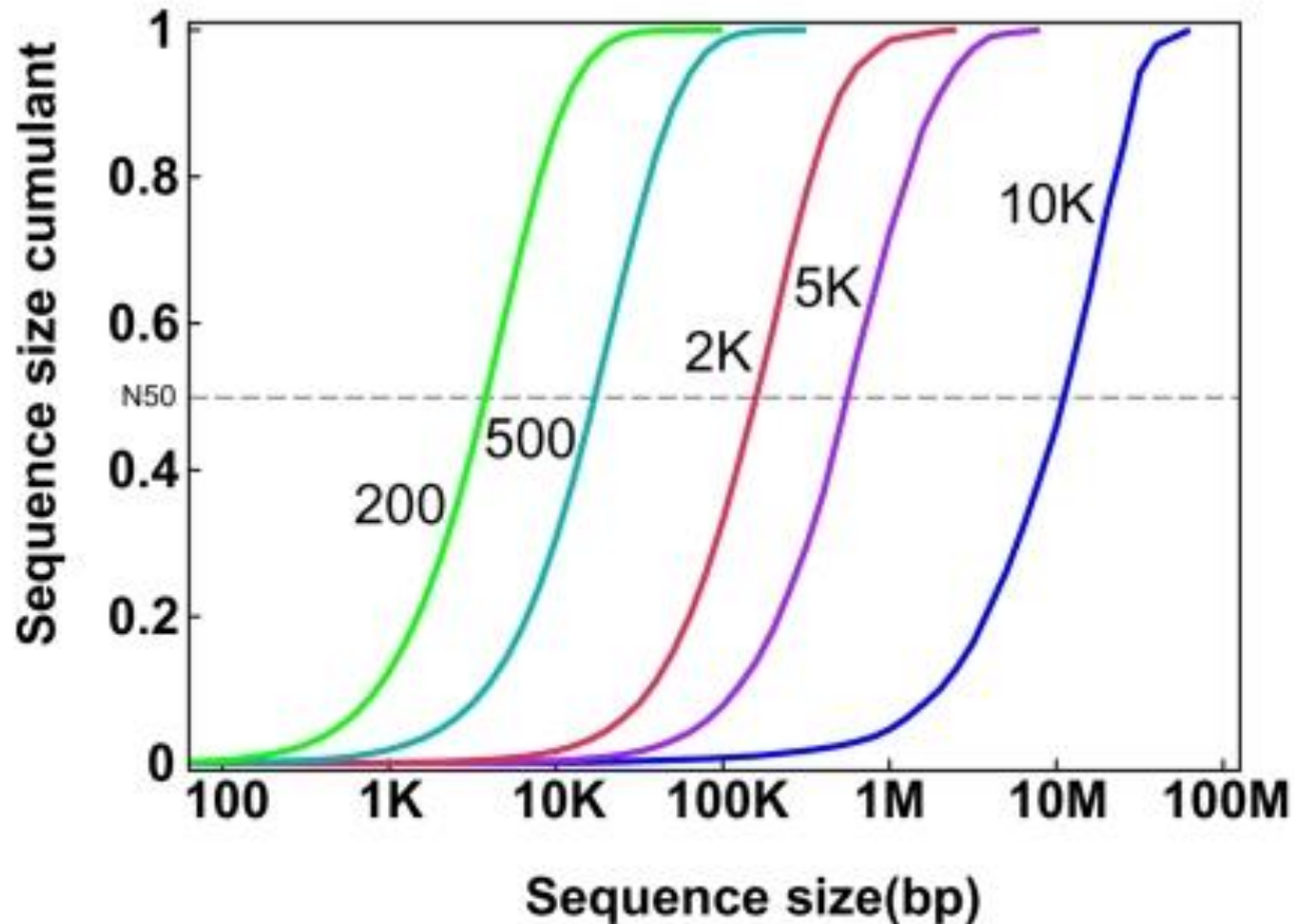
# Что такое N50?

N50 показывает качество сборки

N50 – это такая минимальная длина контига (скаффолда), что контиги (скаффолды) с длинами большими либо равными ей покрывают 50% генома.



# Влияние длины вставки библиотеки на N50



# Что нужно знать о данных из которых вы собираетесь делать de novo сборку



- **Технология** секвенирования.
- **Длина** чтения.
- Тип библиотеки. SE, PE. **Длина вставки.**
- **Покрытие.**
- Имеется ли **загрязнение** образца?

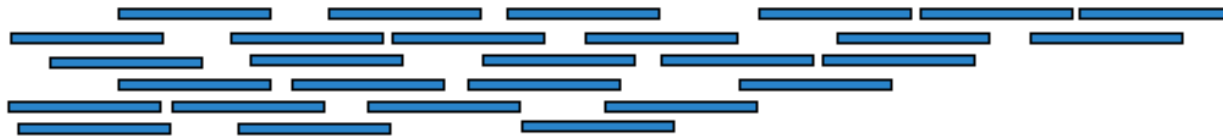
# Что такое покрытие?

Это сколько раз в среднем покрыт ридом нуклеотид генома?

**Multiple Copies of a Genome**



**Reads**



**High Coverage**

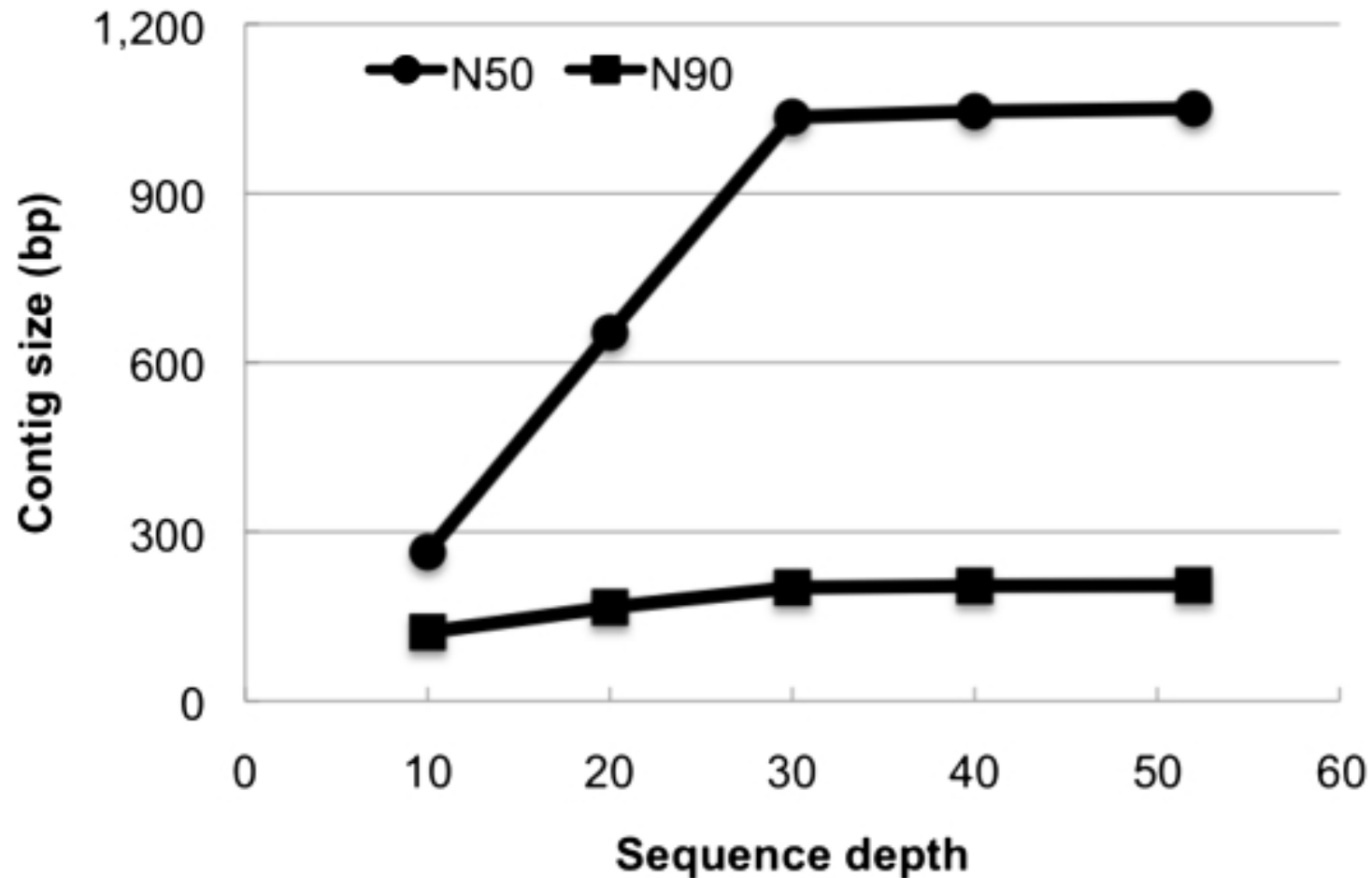
**Low Coverage**



**Consensus Sequence**



# Влияние покрытия на N50



# Если вы хотите собрать большой геном

Требуйте библиотеки с разными длинами вставок

Геном *Ficedula flycatchers* - 1.1 Gb



Библиотеки:

1)~200

4)~500

7)~5100

2)~300

5)~2400

8)~18000

3)~400

6)~4100

9)~21000

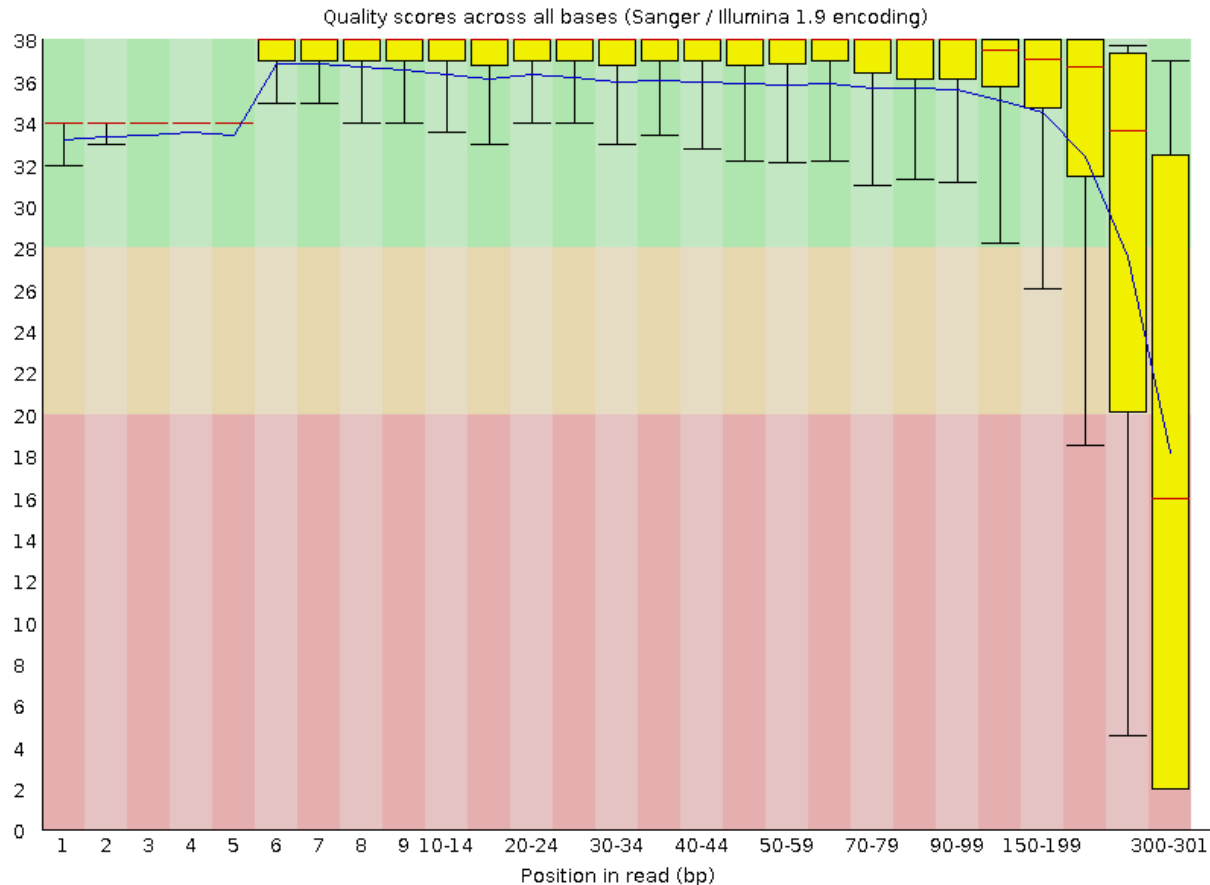
[Ellegren H et al. , The genomic landscape of species divergence in *Ficedula* flycatchers. Nature 2012, 491.]

# Подготовка чтений



- Удаление адаптеров и тримминг

Trimmomatic

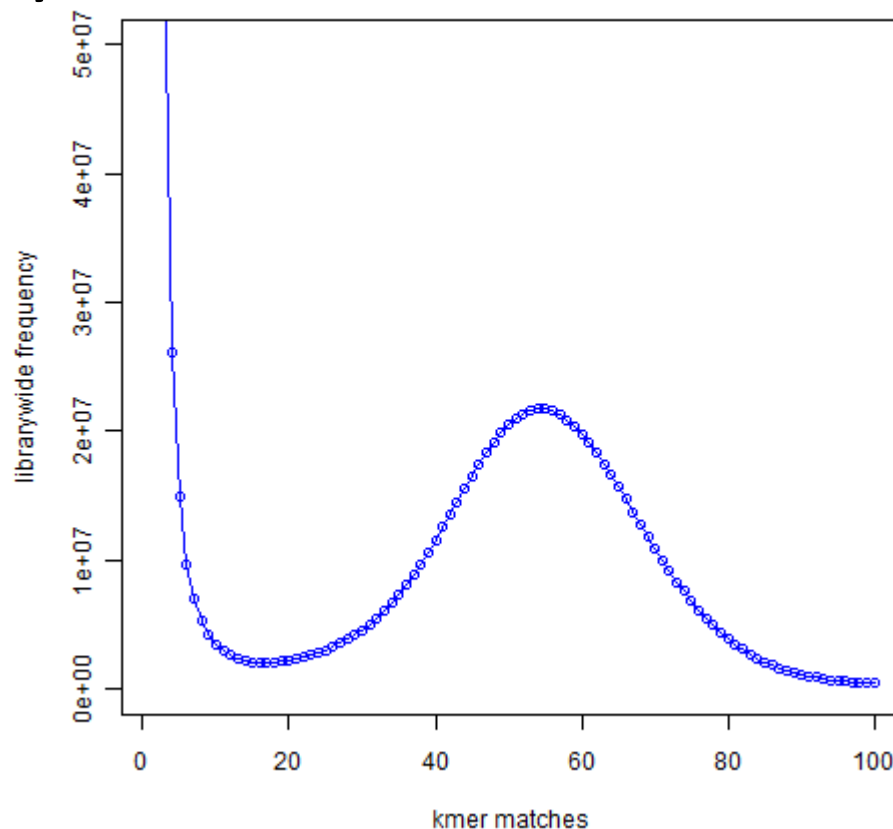


# Подготовка чтений



- **Фильтрация по содержанию к-меров**

Quake, BayesHammer, ...



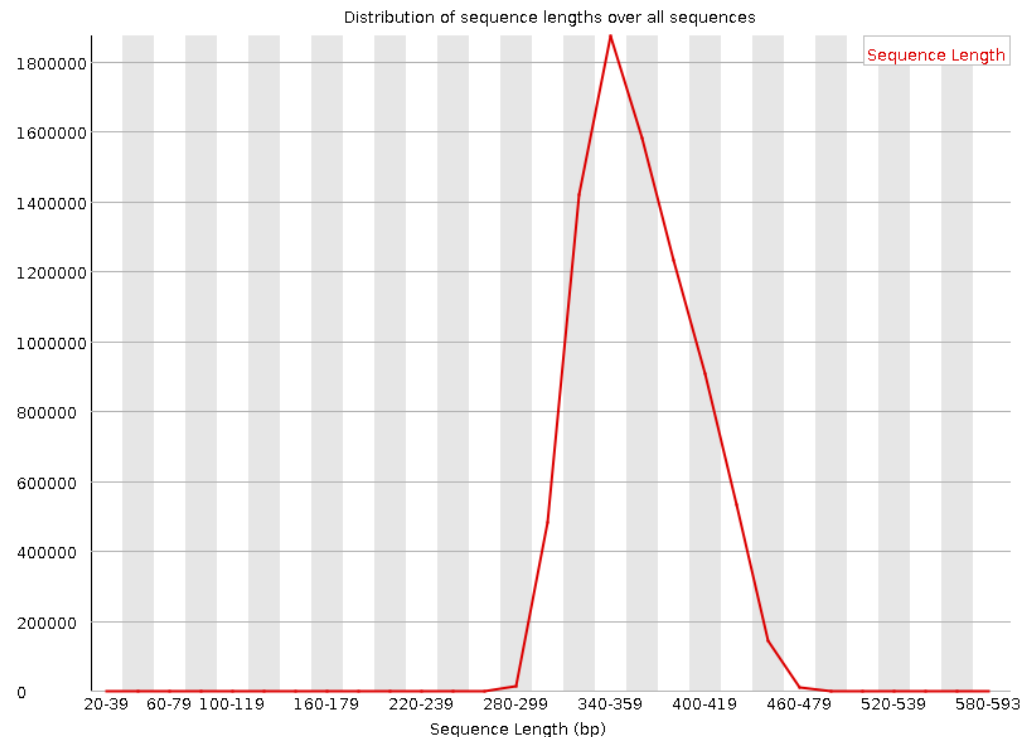
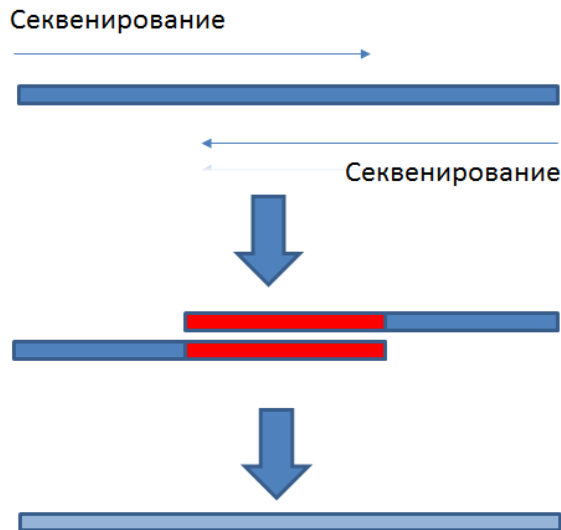
[<http://www.homolog.us/blogs/blog/2011/09/20/maximizing-utility-of-available-rams-in-k-mer-world/>]



# Подготовка чтений



- Парноконцевые чтения с перекрывающейся вставкой



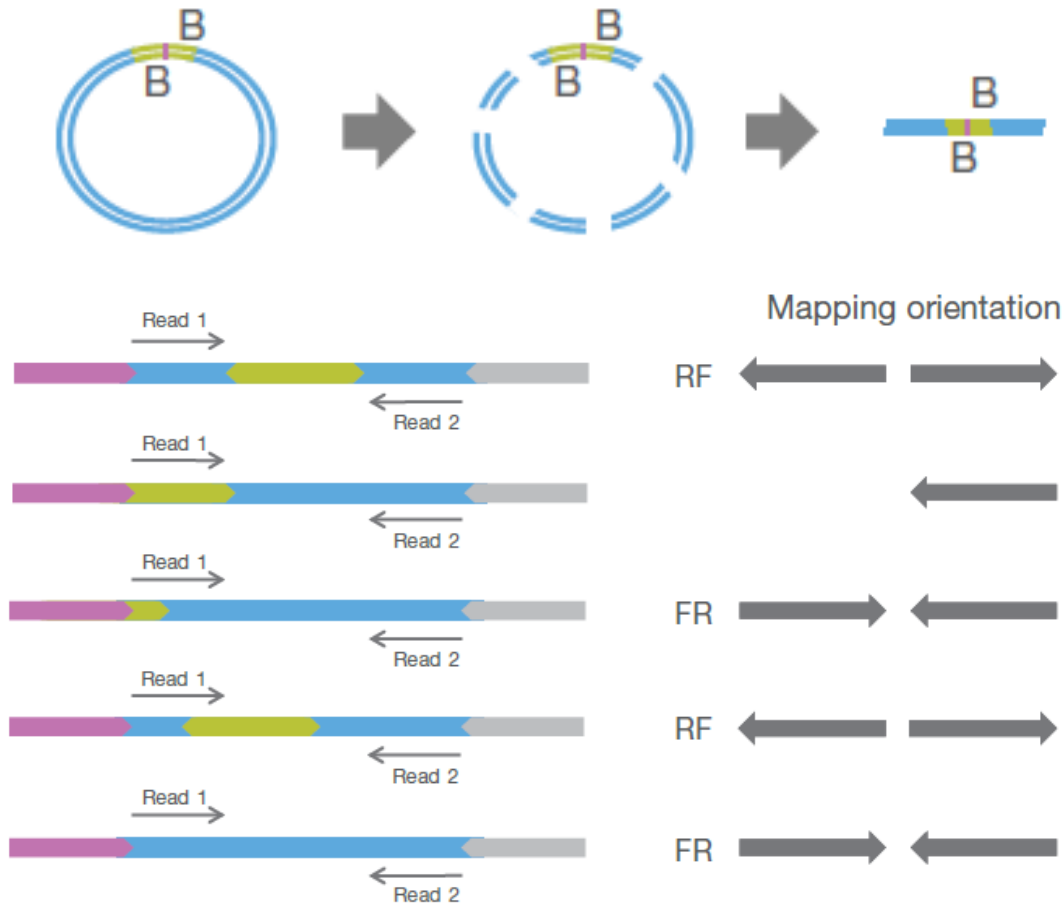
PEAR. FLASH ...

<http://thegenomefactory.blogspot.ru/2012/11/tools-to-merge-overlapping-paired-end.html>

# Подготовка чтений



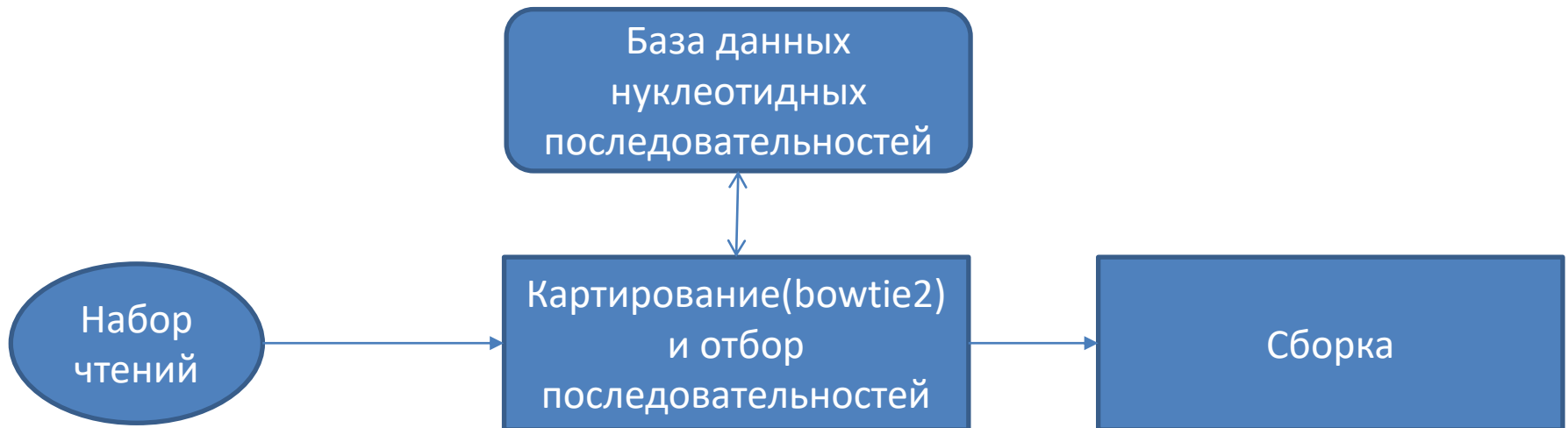
- **Mate pairs**(Nextera MP reads - NxTrim, NextClip)



[[http://www.illumina.com/documents/products/technotes/technote\\_nextera\\_matepair\\_data\\_processing.pdf](http://www.illumina.com/documents/products/technotes/technote_nextera_matepair_data_processing.pdf)]

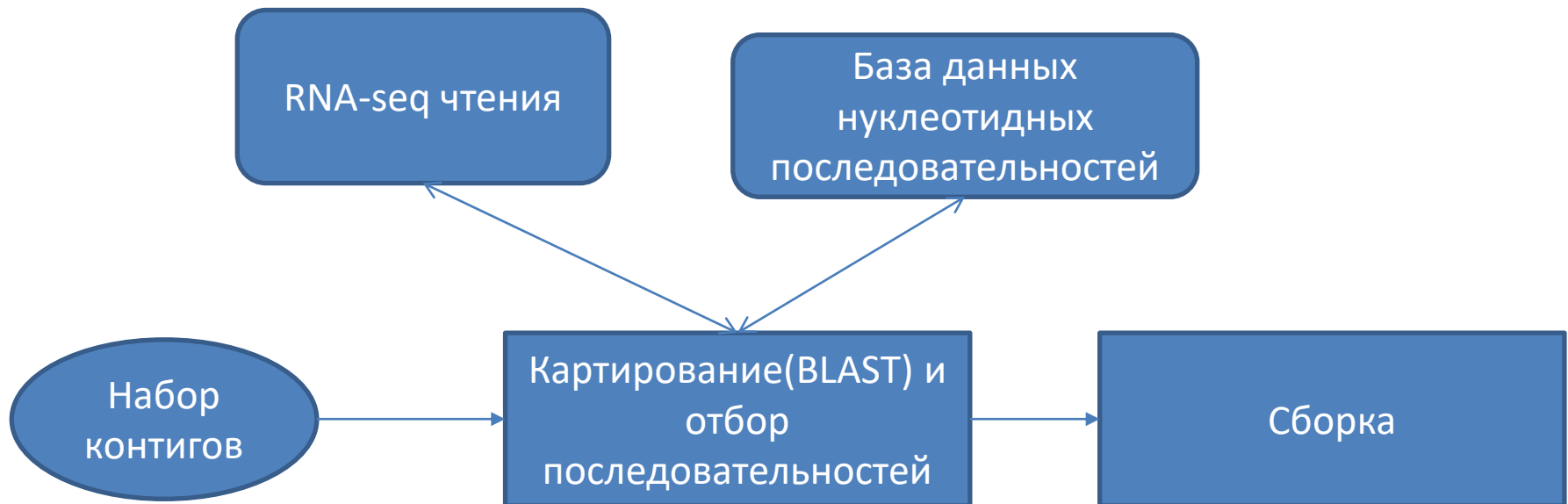
# Что делать с загрязнением?

- «Очистка» чтений.



# Что делать с загрязнением?

- «Очистка» КОНТИГОВ.



Evgeny V Leushkin et al. The miniature genome of a carnivorous plant *Genlisea aurea* contains a low number of genes and short non-coding sequences. BMC Genomics 2013, 14:476

# Чем собирать?

- **Геном размером до 100-200 млн. п.н.**

Spades, Ray, IDBA, Abyss....

- **Большие геномы.**

– *Риды до 200 п.н.*

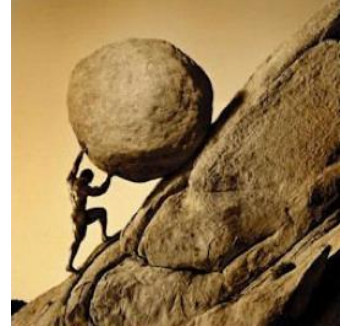
SOAP, MaSuRCA, Meraculous, Platanus, ALLPaths-LG, IDBA, Ray, Abyss, Minia.....

– *Риды длиннее 200 п.н.*

Newbler, Celera assembler, MIRA, ARACHNE, SGA, HGAP, Falcon, MHAP, SSPACE, CANU, PBcR, Sprai....

- **Геномы видов с высокой гетерозиготностью**

dipSpades, Platanus, newbler





# Оценка качества сборки

- **Число контигов**
  - чем меньше тем лучше.
- **N50**
  - чем больше тем лучше
- **Total consensus**
  - должен быть близок к ожидаемой длине генома
- **Число “N”**
  - чем меньше тем лучше

# Валидация сборки



- **Самосогласованность**
  - Картирование чтений обратно на контиги
  - Тест на наличие ошибок или **несогласованных парноконцевых чтений**
- **Второе мнение**
  - Использование **двух** друг друга дополняющих **методов секвенирования**
  - Проверка «**подозрительных**» регионов с использованием **ПЦР**
  - Использование полногеномных «**рестрикционных карт**»

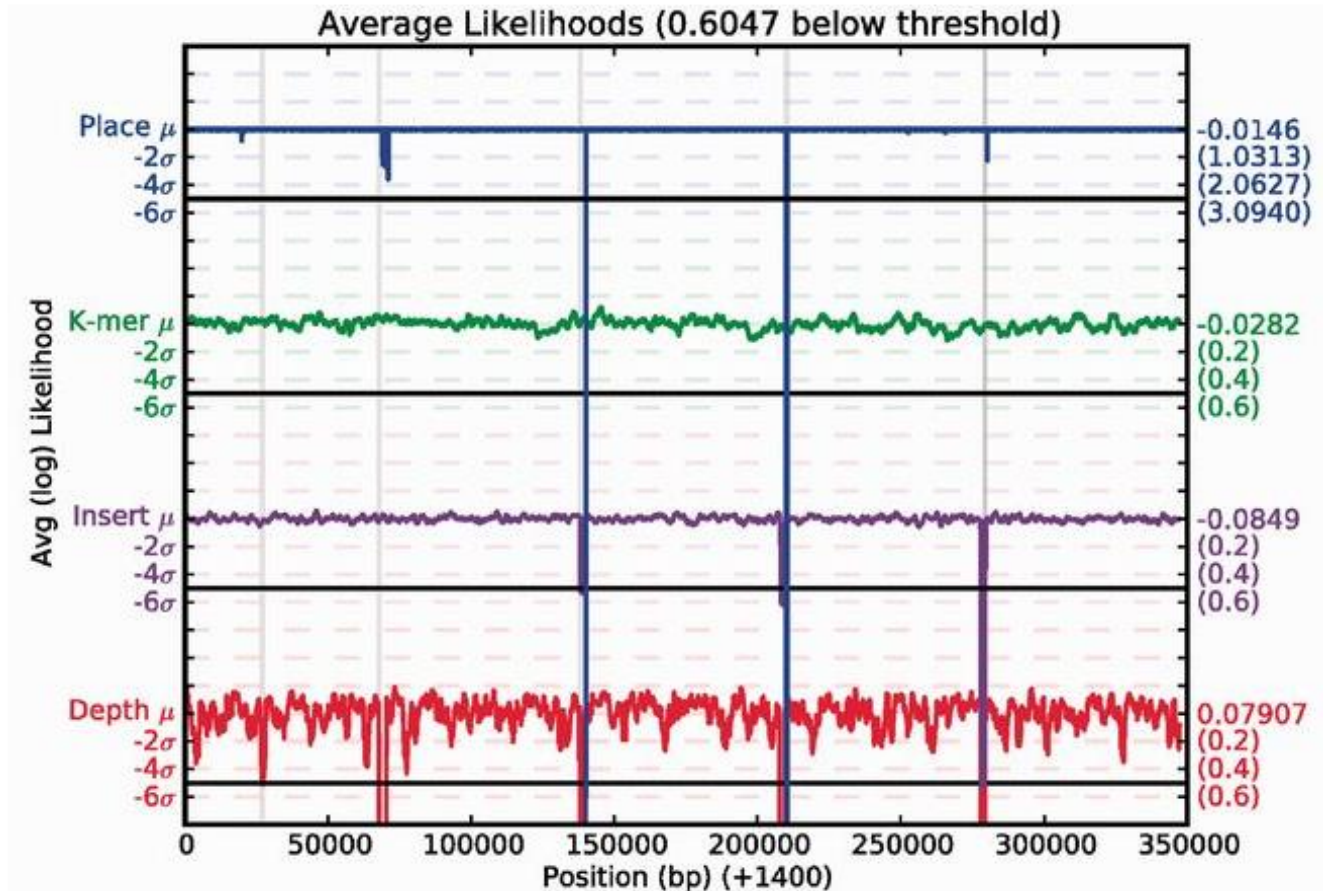
# Программы для оценки качества и валидации сборки



- **Оценка качества. QUAST**
- **Оценка числа реконструированных генов – BUSCO**
- **Валидация сборки** путем картирования чтений обратно на **сборку**.  
REAPR, ALE...

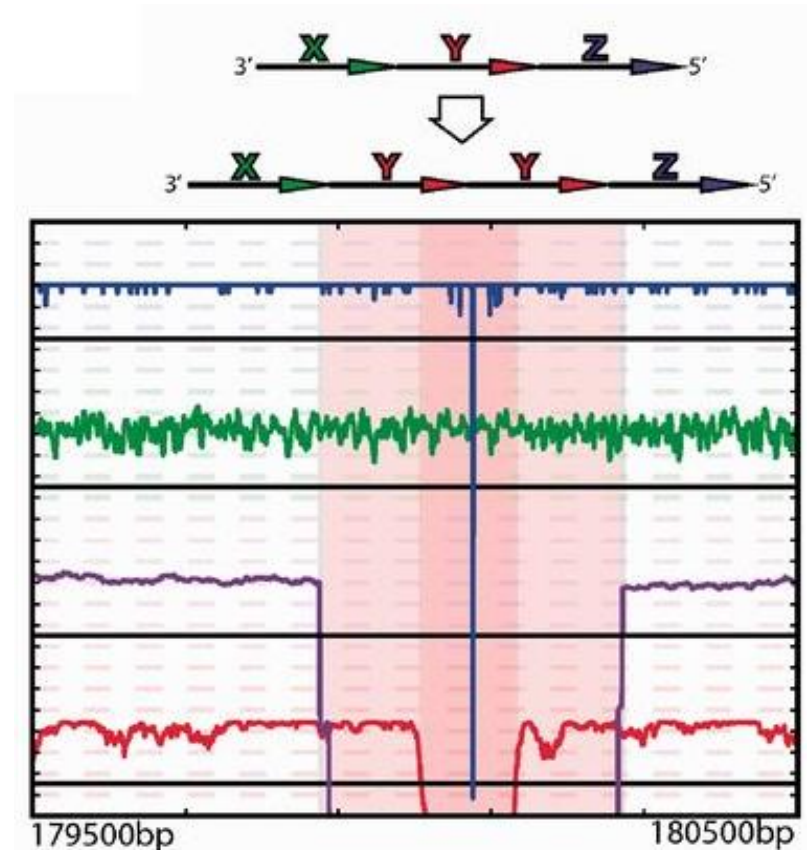
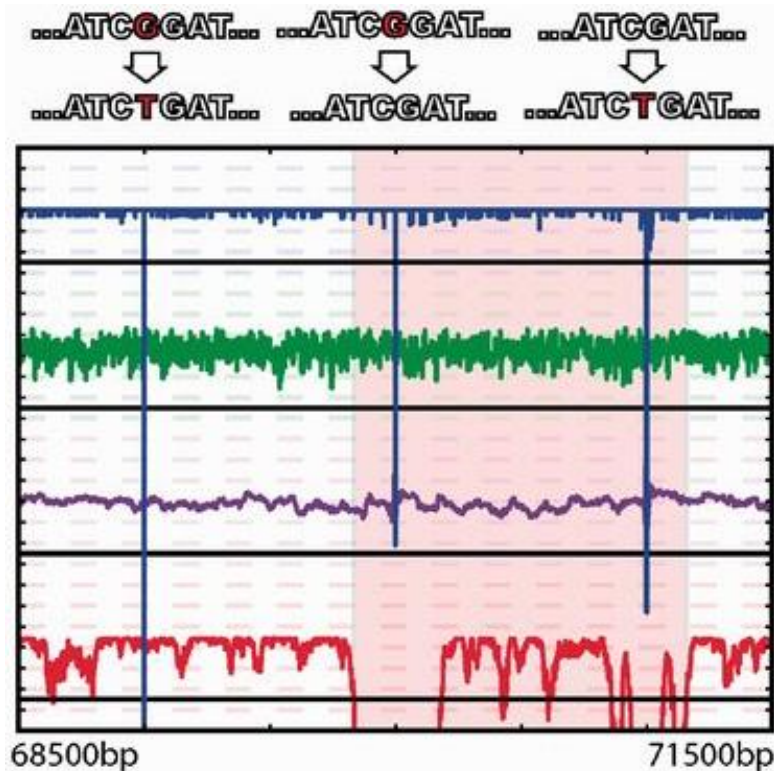


# ALE



[Scott C. Clark et al. ALE: a generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies. *Bioinformatics* (2013) 29 (4): 435-443.]

# ALE



[Scott C. Clark et al. ALE: a generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies. *Bioinformatics* (2013) 29 (4): 435-443.]

# ALE

## Likelihoods of GAGE assemblies of human chromosome 14

Assembler	Likelihood	Number of reads mapped	Coverage (%)	Scaffold N50 (kb)	Contig N50 (kb)
ABYSS	$-23.44 \times 10^8$	22096466	82.22	2.1	2
ALLPATHS-LG	$-22.77 \times 10^8$	23122569	97.24	81647	36.5
CABOG	$-21.26 \times 10^8$	23433424	98.32	393	45.3
SOAPdenovo	<sup>a</sup>	<sup>a</sup>	98.17	455	14.7
Reference	$-19.04 \times 10^8$	23978017	-	-	-

<sup>a</sup> Likelihood not computed as reads could not be mapped with Bowtie 2.

# Финализация сборки



- **Скаффолдинг по транскриптому.**

L\_RNA\_scaffolder, BESST ...

- **Скаффолдинг «вручную»**

Bandage ....

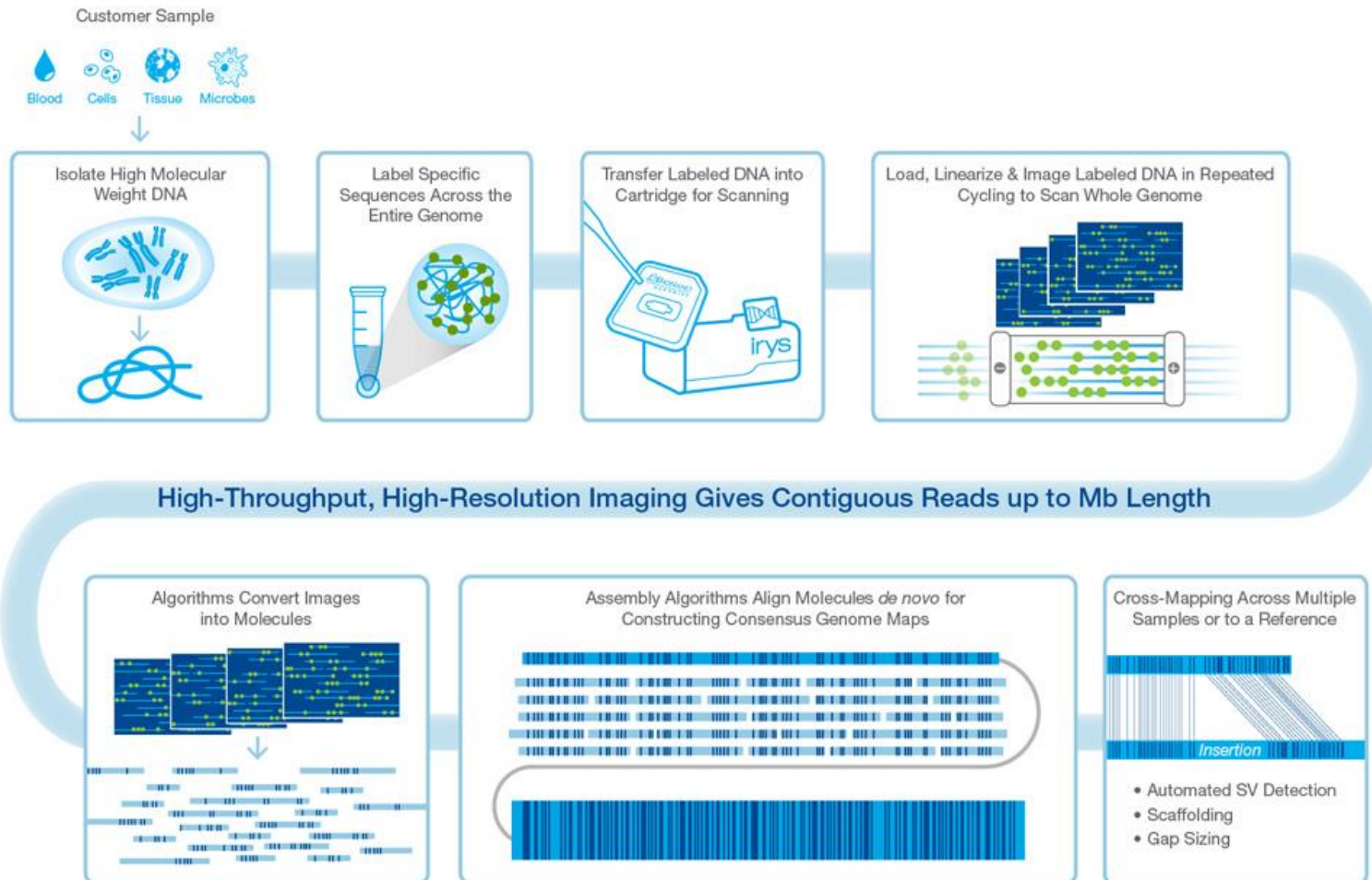
- **Объединение сборок(Assembly reconciliation)**

MIX, Slicemblem ...

- **Соединение концов скаффолдов с использованием ПЦР.**

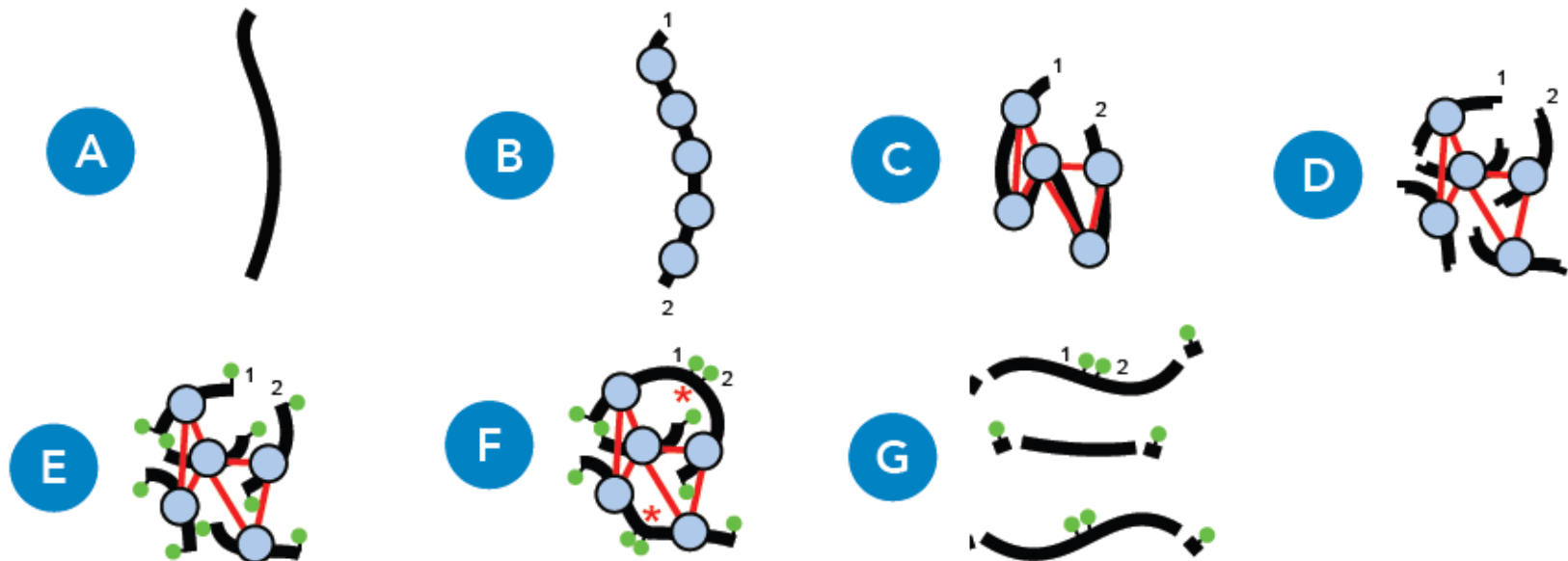
- **Генетическая карта.**

# Irys technology



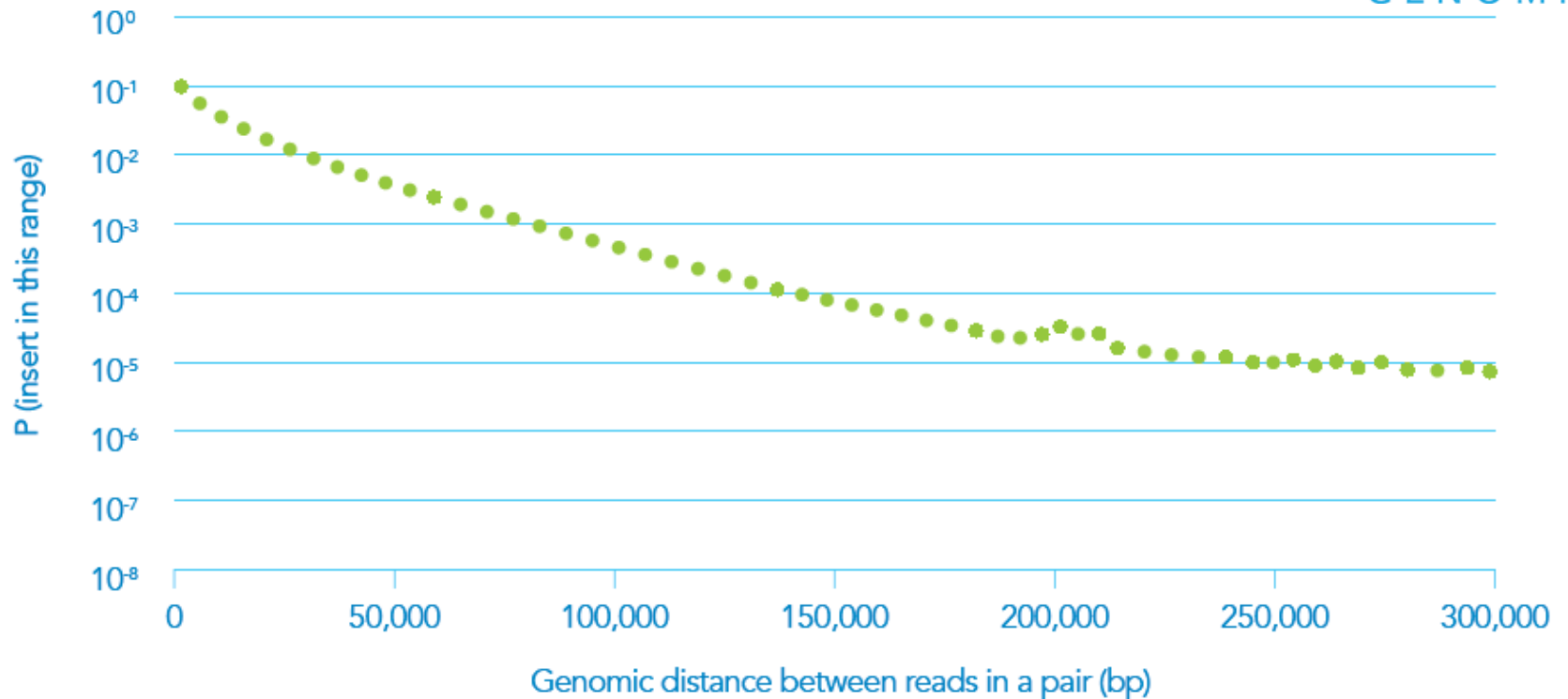


# Chicago method



[<https://dovetailgenomics.com/wp-content/uploads/2016/01/Dovetail-White-Paper.pdf>]

# Chicago method



[<https://dovetailgenomics.com/wp-content/uploads/2016/01/Dovetail-White-Paper.pdf>]

# Как собирали геном морковки?

Какие были данные



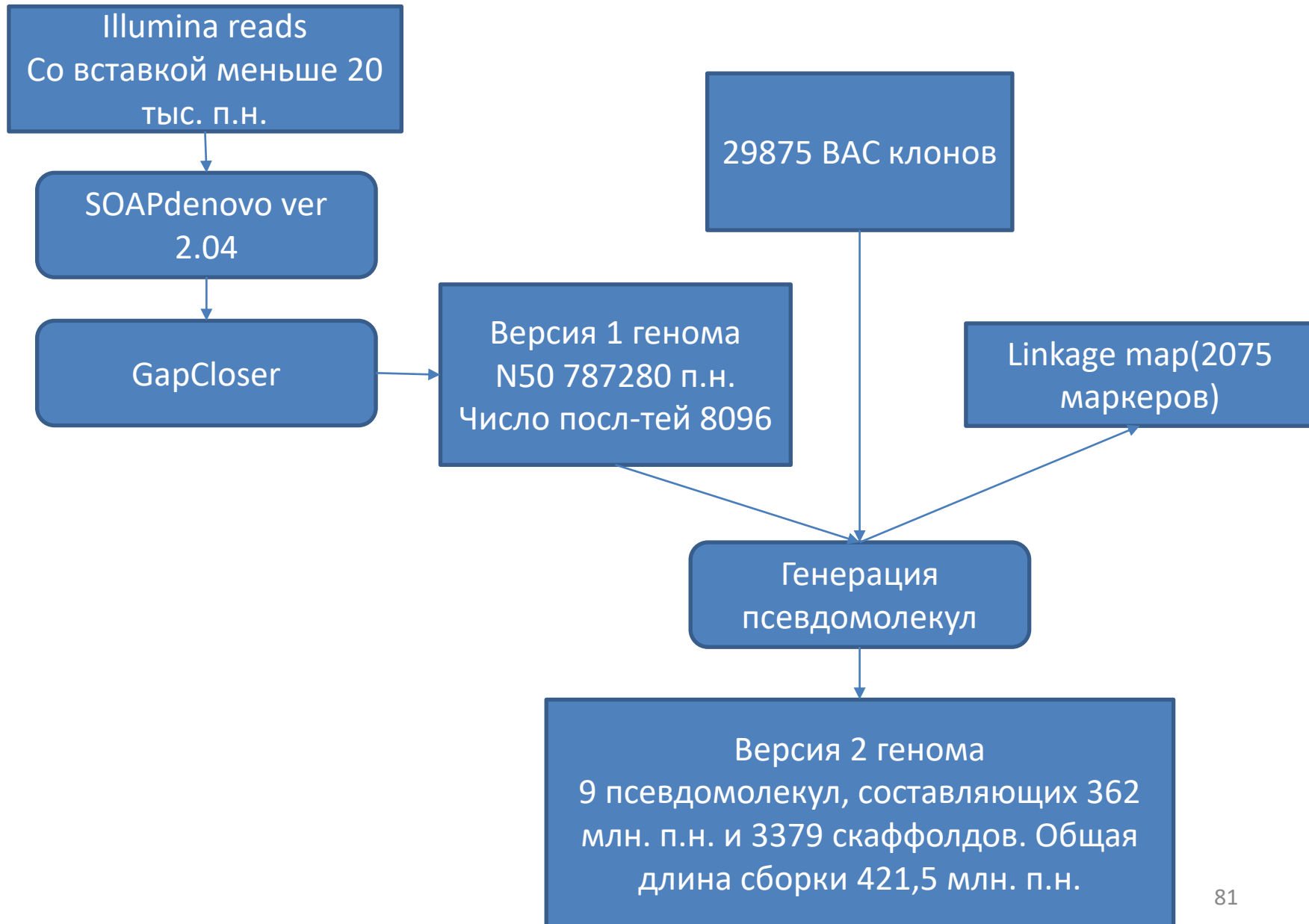
Размер генома  
~470 млн. п.н.

*Daucus carota*

Sequencing method	Insert Size	Read Length (nt)	Total Data (nt×10 <sup>9</sup> )	Sequence Depth (×)
Illumina, Paired-ends	170nt	100	29,2	61,7
	285nt	100	25,1	53,2
	800nt	100	15,5	32,8
	2knt	49	12,9	27,2
	5knt	49	7,1	14,9
	10knt	49	20,5	43,3
	20knt	49	22,4	47,4
	40knt	49	14,4	30,5
Total	—	—	147,2	311,1
Sanger, BAC	148±70knt	566	0,04	0,08

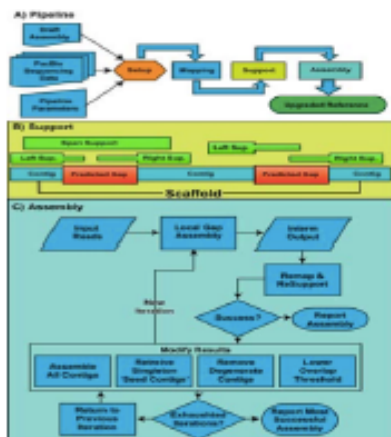


# Как собирали геном морковки?



# Сборка из чтений PacBio

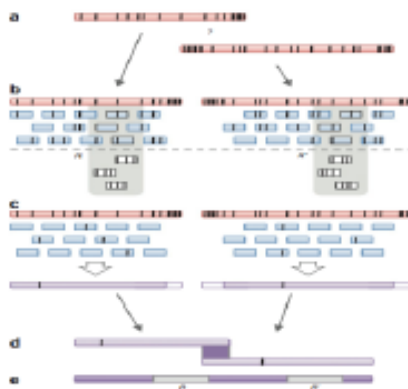
## PBJelly



**Gap Filling  
and Assembly Upgrade**

English et al (2012)  
*PLOS One*. 7(11): e47768

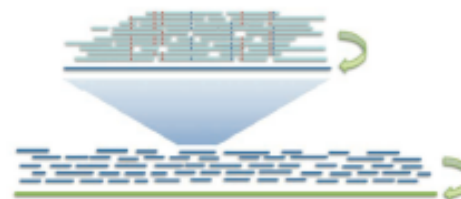
## PacBioToCA & ECTools



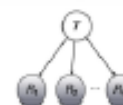
**Hybrid/PB-only Error  
Correction**

Koren, Schatz, et al (2012)  
*Nature Biotechnology*. 30:693–700

## HGAP & Quiver



$$\Pr(R | T) = \prod \Pr(R_k | T)$$



**Quiver Performance Results**  
Comparison to Reference Genome  
(M. ruber; 3.1 Mb; SMRT® Cells)

	Initial Assembly	Quiver Consensus
QV	43.4	54.5
Accuracy	99.99542%	99.99984%
Differences	143	12

**PB-only Correction &  
Polishing**

Chin et al (2013)  
*Nature Methods*. 10:563–569

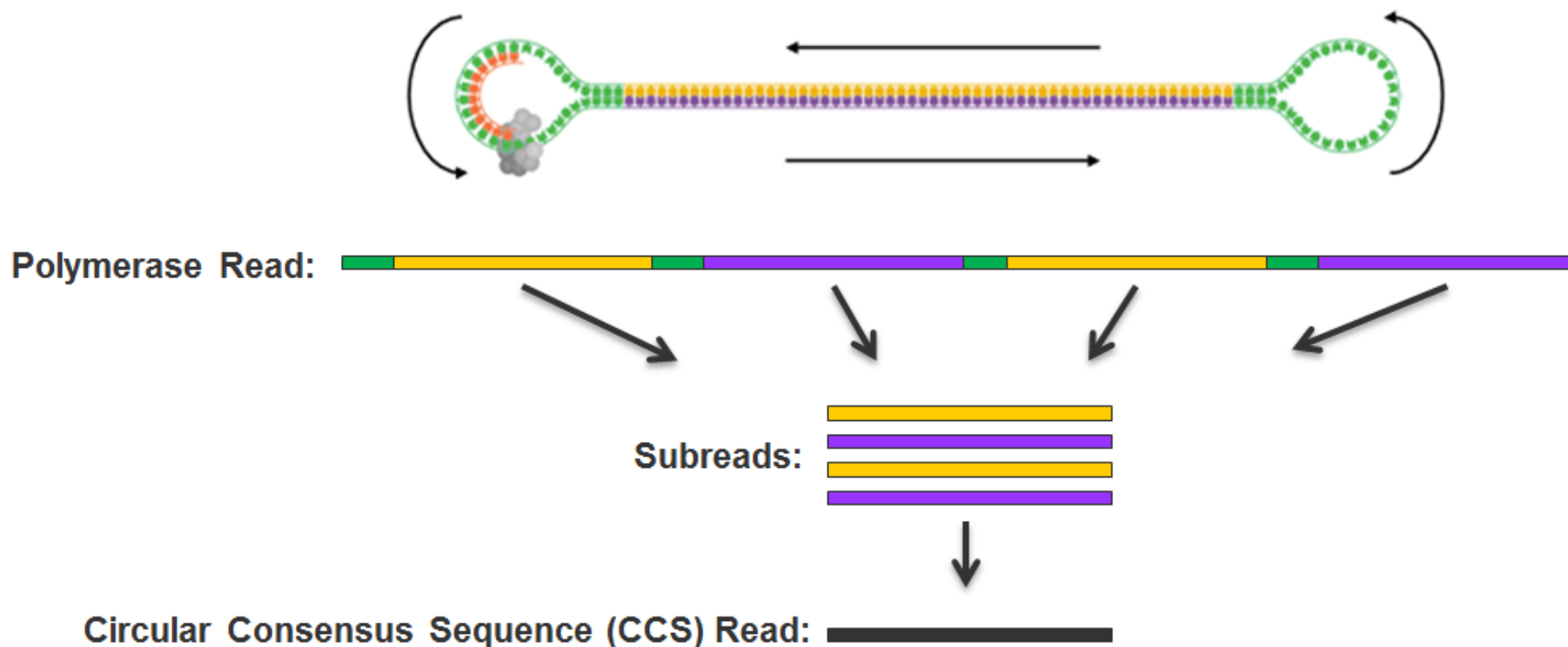
< 5x

PacBio Coverage

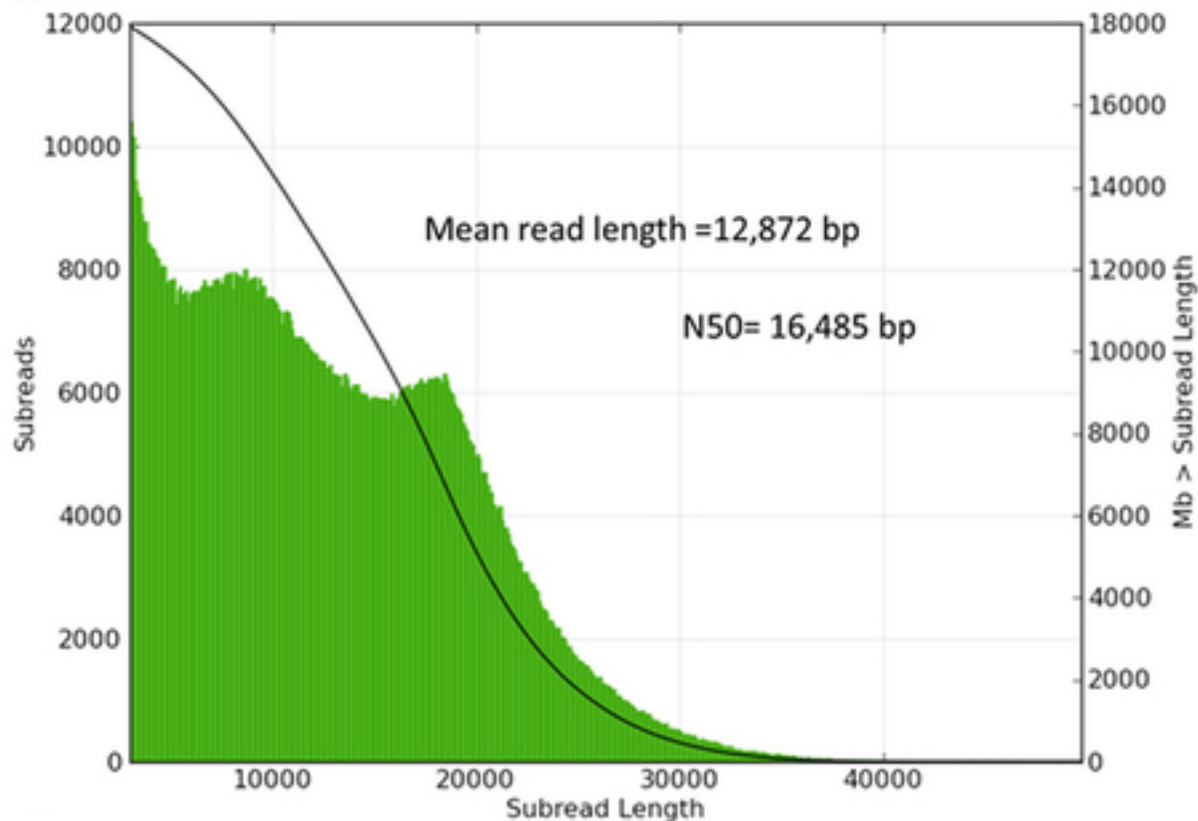
> 50x

# РасВіо.

## Терминология



# Сборка из чтений РасВіо



*Oropetium thomaeum*



Размер генома ~250 млн. п.н.

# Сборка из чтений РасВіо

---

**Raw Input:**

---

Mean Subread Length	12,872 bp
N50 (Subread Length)	16,485 bp
Total Number of sequenced Bases	18,022,966,707 bp
Number of Reads	1,400,150

---

**HGAP Preassembly (BLASR):**

---

Seed length cutoff	16,000 bp
Pre-Assembled Bases	6,281,202,330 bp
Pre-Assembled Reads	464,567
Pre-Assembled N50	18,572 bp

---

**Output (Celera Assembler):**

---

Number of Polished Contigs	625
Max Contig Length	7,984,151 bp
N50 Contig Length	2,386,328 bp
Sum of Contig Lengths	243,174,629 bp

---


# Сборка из длинных ридов

- PacBio
  - HGAP, HGAP2, Falcon, Canu, Spades, Celera assembler, DBG2OLC
- minION
  - Spades, Celera assembler, NanoPolish, Canu, DBG2OLC

# The genome assembly of celery

Article | [Open Access](#) | [Published: 06 January 2020](#)

## **The genome sequence of celery (*Apium graveolens* L.), an important leaf vegetable crop rich in apigenin in the Apiaceae family**

Meng-Yao Li, Kai Feng, Xi-Lin Hou, Qian Jiang, Zhi-Sheng Xu, Guang-Long Wang, Jie-Xia Liu, Feng Wang & Ai-Sheng Xiong 

*Horticulture Research* **7**, Article number: 9 (2020) | [Cite this article](#)

**3339** Accesses | **11** Citations | **2** Altmetric | [Metrics](#)



[<https://doi.org/10.1038/s41438-019-0235-2>]

# The genome assembly of celery

Library insert size	Library number	Clean data (Gb)	Sequence coverage (×)
180 bp	18	211.47	66.50
500 bp	14	152.35	47.91
800 bp	18	136.35	42.88
2 kb	8	47.78	15.03
5 kb	8	27.78	8.74
10 kb	2	25.07	7.88
Total	68	600.80	188.93

CutAdapt+SOAPdenovo2  
+GapCloser



# The genome assembly of celery

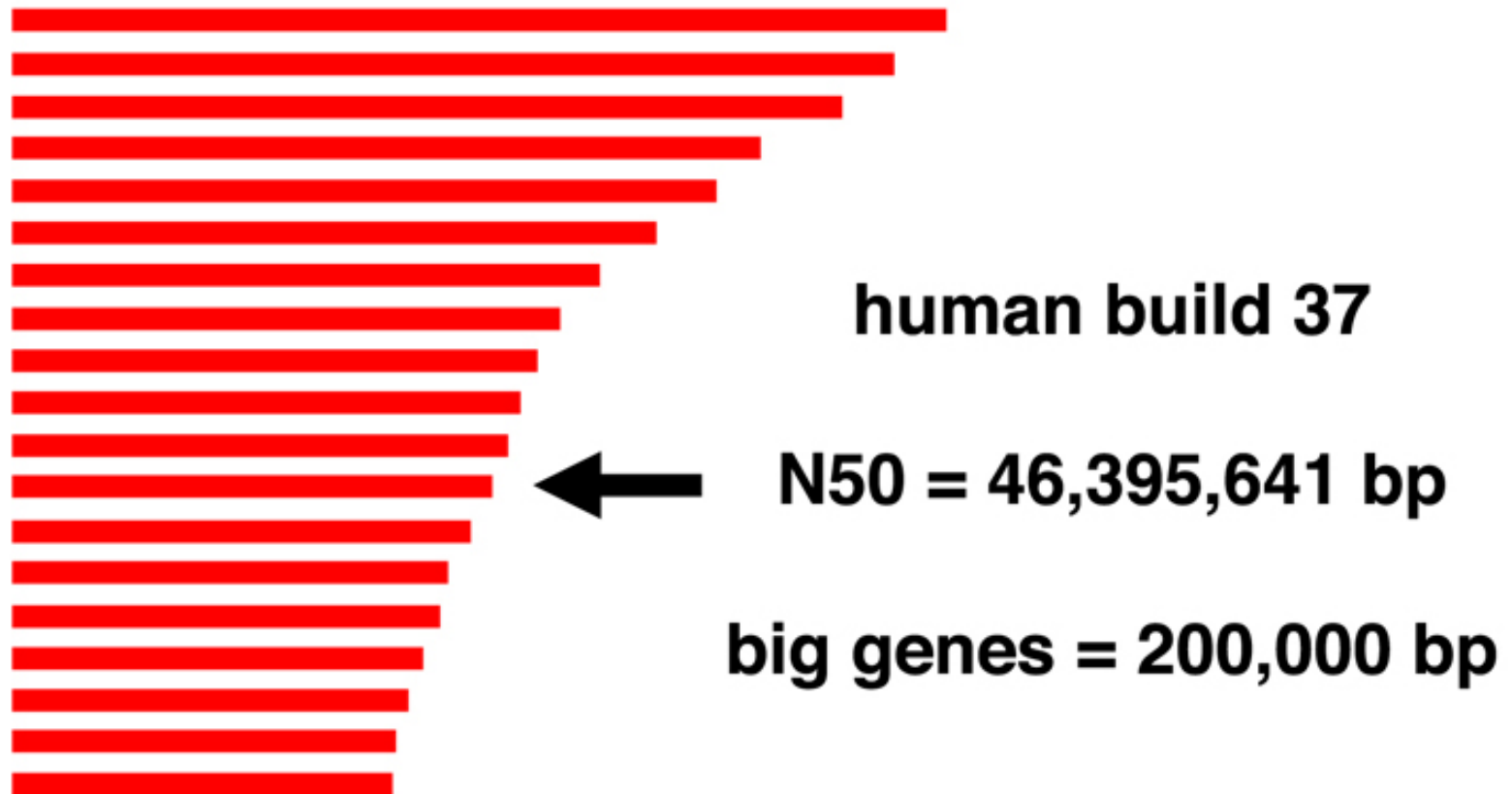
Feature	Value
<b>Genome size</b>	2.21 Gb
<b>Genome GC%</b>	35.35%
<b>Gene number</b>	34,277
<b>Gene no. per 100 kb</b>	1.44
<b>Average gene length (bp)</b>	3267
<b>Exon region GC (%)</b>	42.06%
<b>Exon number</b>	180,591
<b>Average exon length (bp)</b>	243.48
<b>Exon no. per gene</b>	5.27

Property	Contig	Scaffold
<b>Min sequence length (bp)</b>	500	500
<b>Max sequence length (bp)</b>	228,328	556,749
<b>Total sequence number</b>	432,762	257,842
<b>N50 length (bp)</b>	13,108	35,567
<b>N90 length (bp)</b>	1136	4841
<b>N number</b>	648,982	280,637, 212
<b>N rate (%)</b>	0.031	11.8
<b>Total sequence length (bp)</b>	2,017,581,028	2,372,941,895

# What is N50?

N50 statistic defines assembly quality

What is the smallest piece at 50% of genome

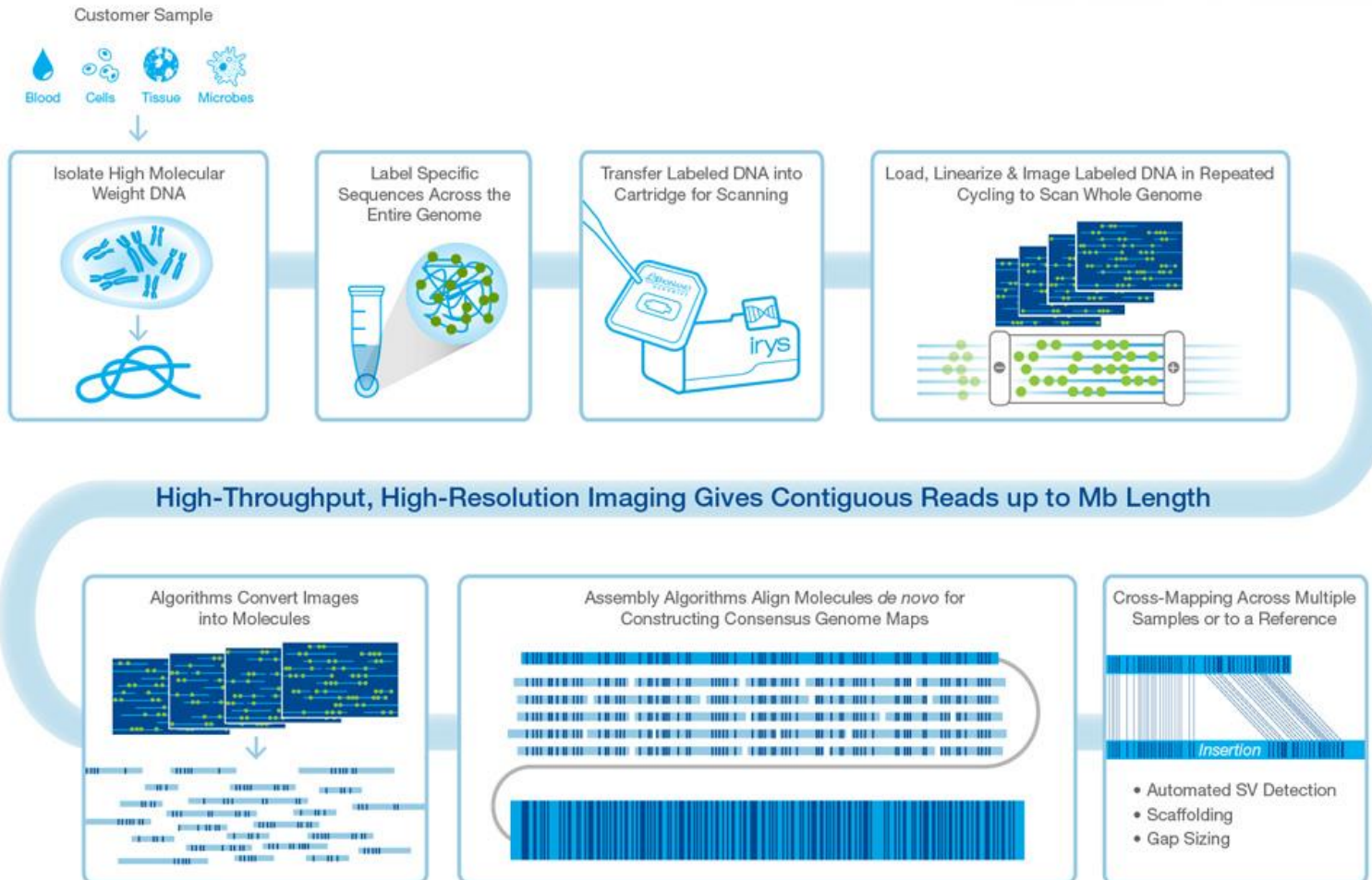


# The genome assembly of celery

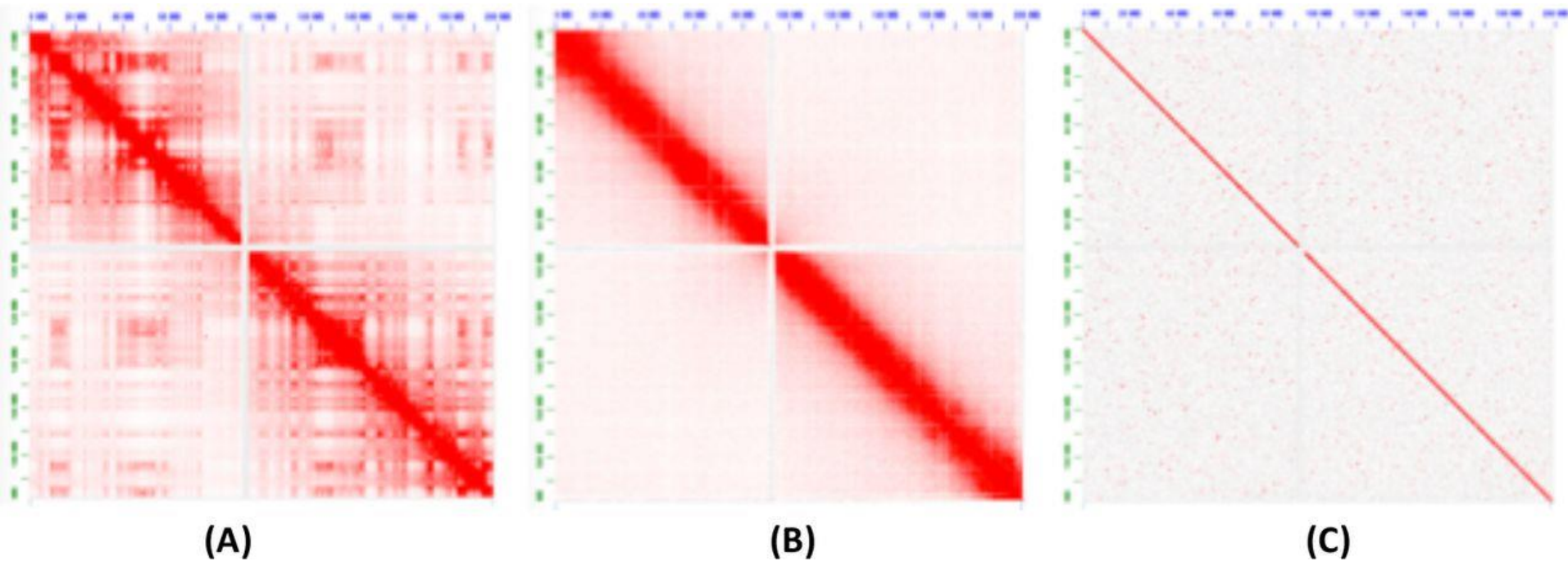
Feature	Value
<b>Genome size</b>	2.21 Gb
<b>Genome GC%</b>	35.35%
<b>Gene number</b>	34,277
<b>Gene no. per 100 kb</b>	1.44
<b>Average gene length (bp)</b>	3267
<b>Exon region GC (%)</b>	42.06%
<b>Exon number</b>	180,591
<b>Average exon length (bp)</b>	243.48
<b>Exon no. per gene</b>	5.27

Property	Contig	Scaffold
<b>Min sequence length (bp)</b>	500	500
<b>Max sequence length (bp)</b>	228,328	556,749
<b>Total sequence number</b>	432,762	257,842
<b>N50 length (bp)</b>	13,108	35,567
<b>N90 length (bp)</b>	1136	4841
<b>N number</b>	648,982	280,637, 212
<b>N rate (%)</b>	0.031	11.8
<b>Total sequence length (bp)</b>	2,017,581,028	2,372,941,895

# Irys technology



# Scaffolding by HiC



[The cells sequenced in (A) normal conditions, (B) during mitosis, and (C) Dovetail Chicago]

# How was the carrot genome assembled?

Data was used

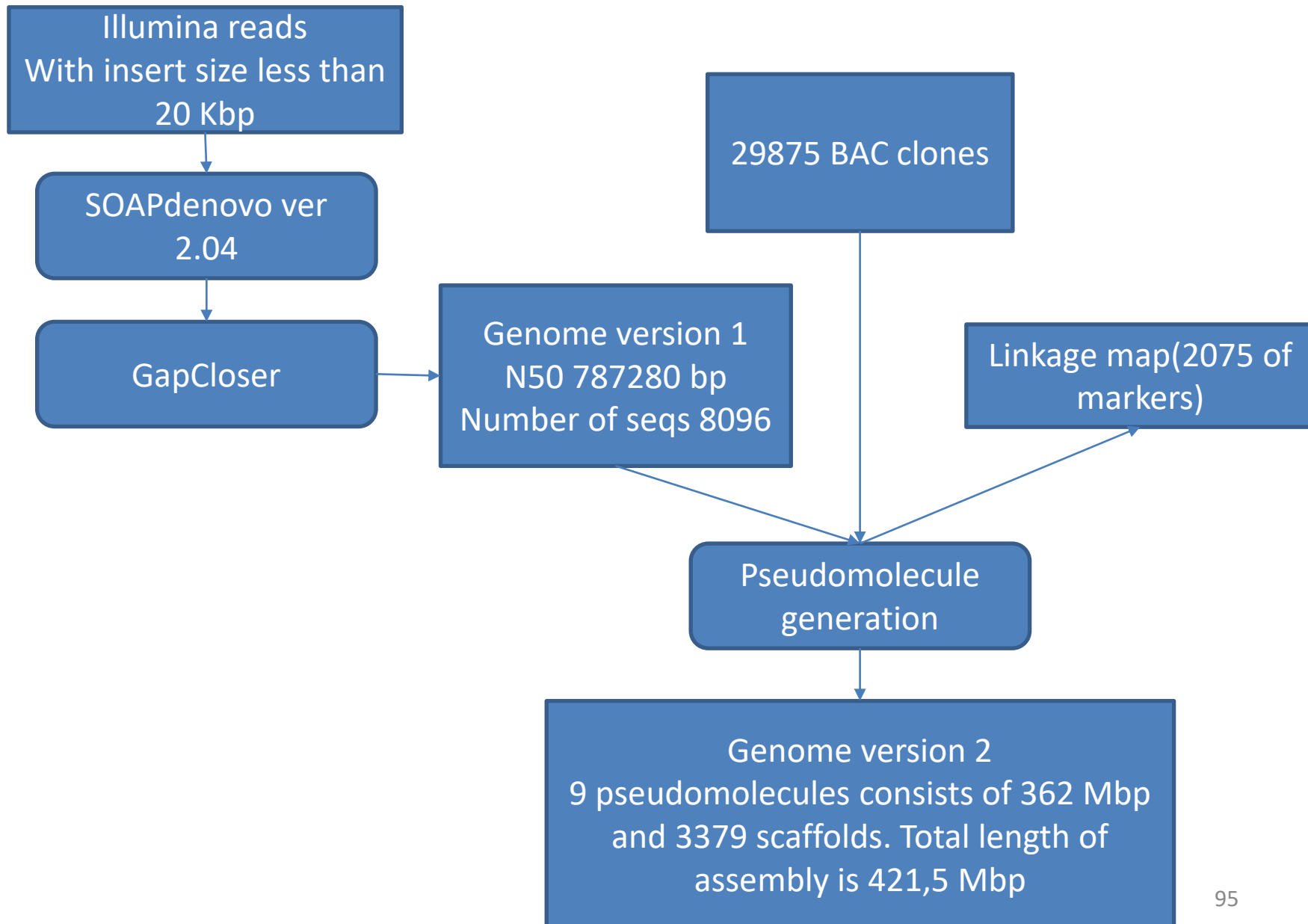


Genome size  
~470 Mbp

*Daucus carota*

Sequencing method	Insert Size	Read Length (nt)	Total Data (nt×10 <sup>9</sup> )	Sequence Depth (×)
Illumina, Paired-ends	170nt	100	29,2	61,7
	285nt	100	25,1	53,2
	800nt	100	15,5	32,8
	2knt	49	12,9	27,2
	5knt	49	7,1	14,9
	10knt	49	20,5	43,3
	20knt	49	22,4	47,4
	40knt	49	14,4	30,5
Total	—	—	147,2	311,1
Sanger, BAC	148±70knt	566	0,04	0,08

# How was the carrot genome assembled?



# Assembly from PacBio reads

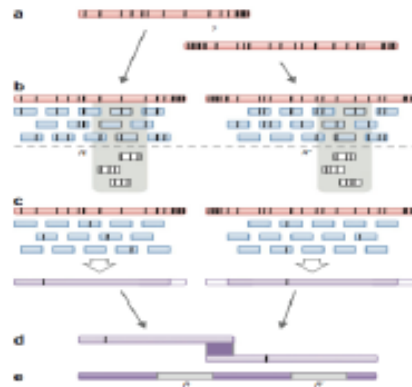
## PBJelly



### Gap Filling and Assembly Upgrade

English et al (2012)  
PLOS One. 7(11): e47768

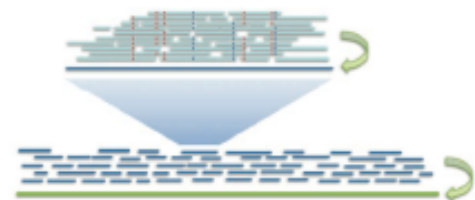
## PacBioToCA & ECTools



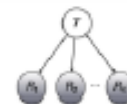
### Hybrid/PB-only Error Correction

Koren, Schatz, et al (2012)  
Nature Biotechnology. 30:693–700

## HGAP & Quiver



$$\Pr(R | T) = \prod \Pr(R_k | T)$$



**Quiver Performance Results**  
Comparison to Reference Genome (M. ruber; 3.1 Mb; SMRT® Cells)

	Initial Assembly	Quiver Consensus
QV	43.4	54.5
Accuracy	99.99542%	99.99984%
Differences	143	12

### PB-only Correction & Polishing

Chin et al (2013)  
Nature Methods. 10:563–569

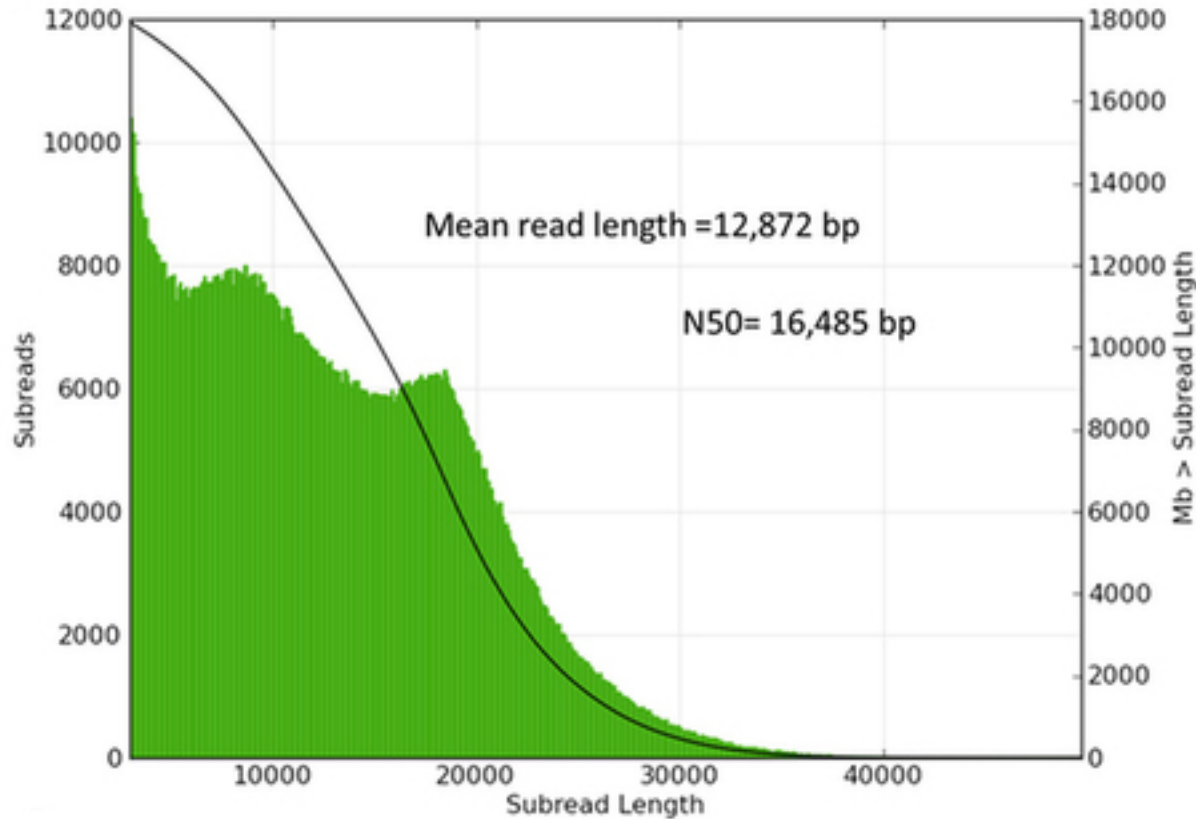
< 5x

PacBio Coverage

> 50x



# Assembly from PacBio reads



*Oropetium thomaeum*



Genome size ~250 Mbp

# Assembly from PacBio reads

---

**Raw Input:**

---

Mean Subread Length	12,872 bp
N50 (Subread Length)	16,485 bp
Total Number of sequenced Bases	18,022,966,707 bp
Number of Reads	1,400,150

---

---

**HGAP Preassembly (BLASR):**

---

Seed length cutoff	16,000 bp
Pre-Assembled Bases	6,281,202,330 bp
Pre-Assembled Reads	464,567
Pre-Assembled N50	18,572 bp

---

---

**Output (Celera Assembler):**

---

Number of Polished Contigs	625
Max Contig Length	7,984,151 bp
N50 Contig Length	2,386,328 bp
Sum of Contig Lengths	243,174,629 bp



---

# Genome assembly of *Ophiorrhiza pumila*



Article | [Open Access](#) | Published: 15 January 2021

## **Chromosome-level genome assembly of *Ophiorrhiza pumila* reveals the evolution of camptothecin biosynthesis**

Amit Rai , Hideki Hirakawa, Ryo Nakabayashi, Shinji Kikuchi, Koki Hayashi, Megha Rai, Hiroshi Tsugawa, Taiki Nakaya, Tetsuya Mori, Hideki Nagasaki, Runa Fukushi, Yoko Kusuya, Hiroki Takahashi, Hiroshi Uchiyama, Atsushi Toyoda, Shoko Hikosaka, Eiji Goto, Kazuki Saito & Mami Yamazaki 

*Nature Communications* **12**, Article number: 405 (2021) | [Cite this article](#)

**1768** Accesses | **48** Altmetric | [Metrics](#)

[doi: 10.1038/s41467-020-20508-2]

# Genome assembly of *Ophiorrhiza pumila*

**Table 1** *O. pumila* reference genome assembly statistics at different stages and combinations of scaffolding.

Assembly	Number of contigs	Number of scaffolds	Number of contigs assigned to scaffolds	Contig N50 (Mb)	Scaffold N50 (Mb)	Number of gaps	Assembly size (Mb)
PacBio <sup>a</sup> only (Canu assembly)	243	—	—	9.38	—	—	449.00
Bionano de novo Optical Map	—	458	—	—	1.68	—	442.00
PacBio + Optical <sup>b</sup> Map	108	45	83	8.21	21.05	117	442.00
PacBio + Hi-C <sup>c</sup>	213	34	198	9.39	40.80	96	441.00
PacBio + Hi-C + Optical Map <sup>d</sup>	239	26	208	8.21	24.17	91	441.90
PacBio + Optical Map + Hi-C <sup>e</sup>	108	13	108	8.21	37.11	85	439.00
PacBio + Optical Map + Hi-C + Pbjelly (PacBio) + genome polishing (final <i>O. pumila</i> reference genome)	31	13 (11 Chromosomes + 1 MT + 1 CP)	31	18.49	40.06	21	439.90

*O. pumila* is a medicinal plant that can produce the anticancer monoterpene indole alkaloid (MIA) camptothecin. Here, the authors report its genome assembly, and propose a working model for MIA evolution and biosynthesis through comparative genomics, synteny, and metabolic gene cluster analyses.

<sup>a</sup>PacBio refers to contig assembly derived using Pacbio reads only and Canu<sup>22</sup> assembler.

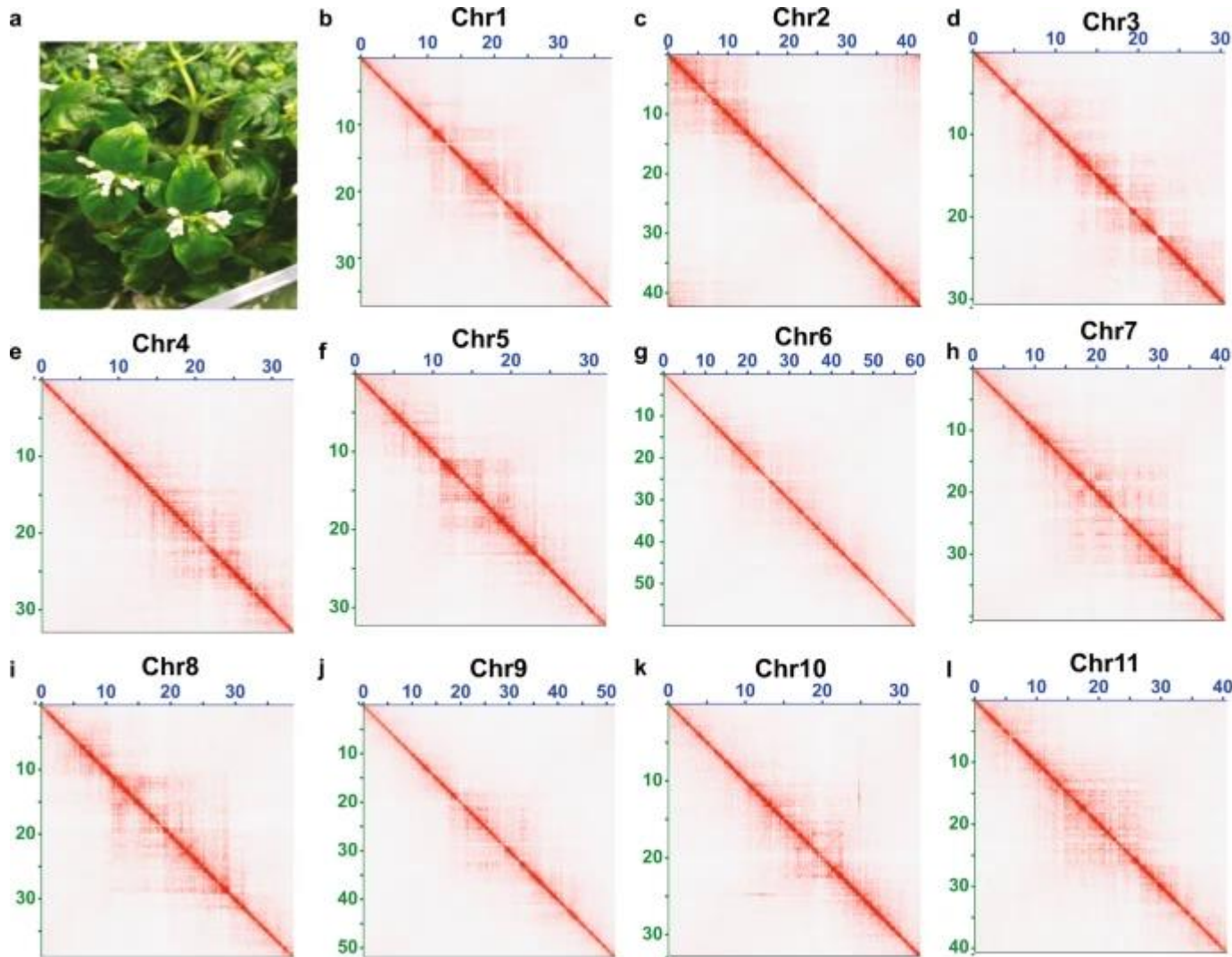
<sup>b</sup>PacBio + Optical Map refers to Pacbio contig-level assembly scaffolded by Bionano de novo assembly.

<sup>c</sup>PacBio + Hi-C refers to Pacbio contig-level assembly scaffolded by Hi-C library sequencing datasets.

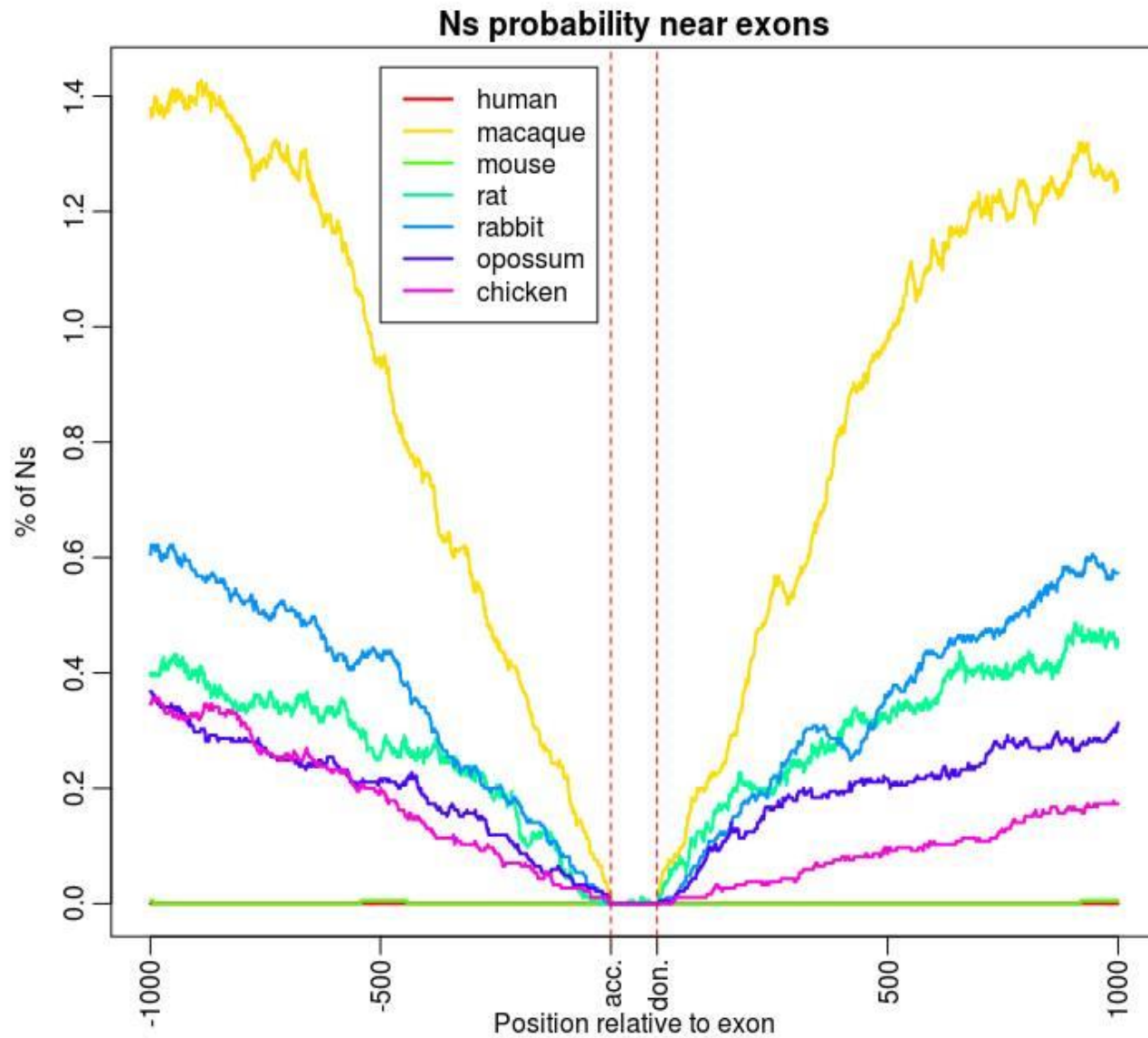
<sup>d</sup>PacBio + Hi-C + Optical Map refers to Pacbio + Hi-C assembly scaffolded by Bionano de novo assembly.

<sup>e</sup>PacBio + Optical Map + Hi-C refers to Pacbio + Optical Map assembly scaffolded by Hi-C library sequencing datasets.

# Genome assembly of *Ophiorrhiza pumila*







[<https://www.facebook.com/photo.php?fbid=959980017480213&set=a.320211138123774.1073741828.100004046713050&type=3>]

# Советы



- Не надо **разрабатывать** свой собственный **сборщик**.
- **Быстро** получите **первую версию генома** – сразу станет понятно есть ли у проекта шансы.
- **Пробластуйте** порцию **чтений** против **NT** – нет ли у вас значительных загрязнений.
- **Бластуйте** собранные **контиги** против **NT** – получили ли тот вид, что ожидали?
- Если вы занимаетесь **сборкой больше двух месяцев** то скорее всего вы застряли. Переходите к **следующей стадии проекта**.

# Полезные ссылки

Monya Baker. De novo genome assembly: what every biologist should know. Nature Methods 9, 333-337 (2012).

<http://bsc2010.bioinformatics.ucdavis.edu/handson/index.html>