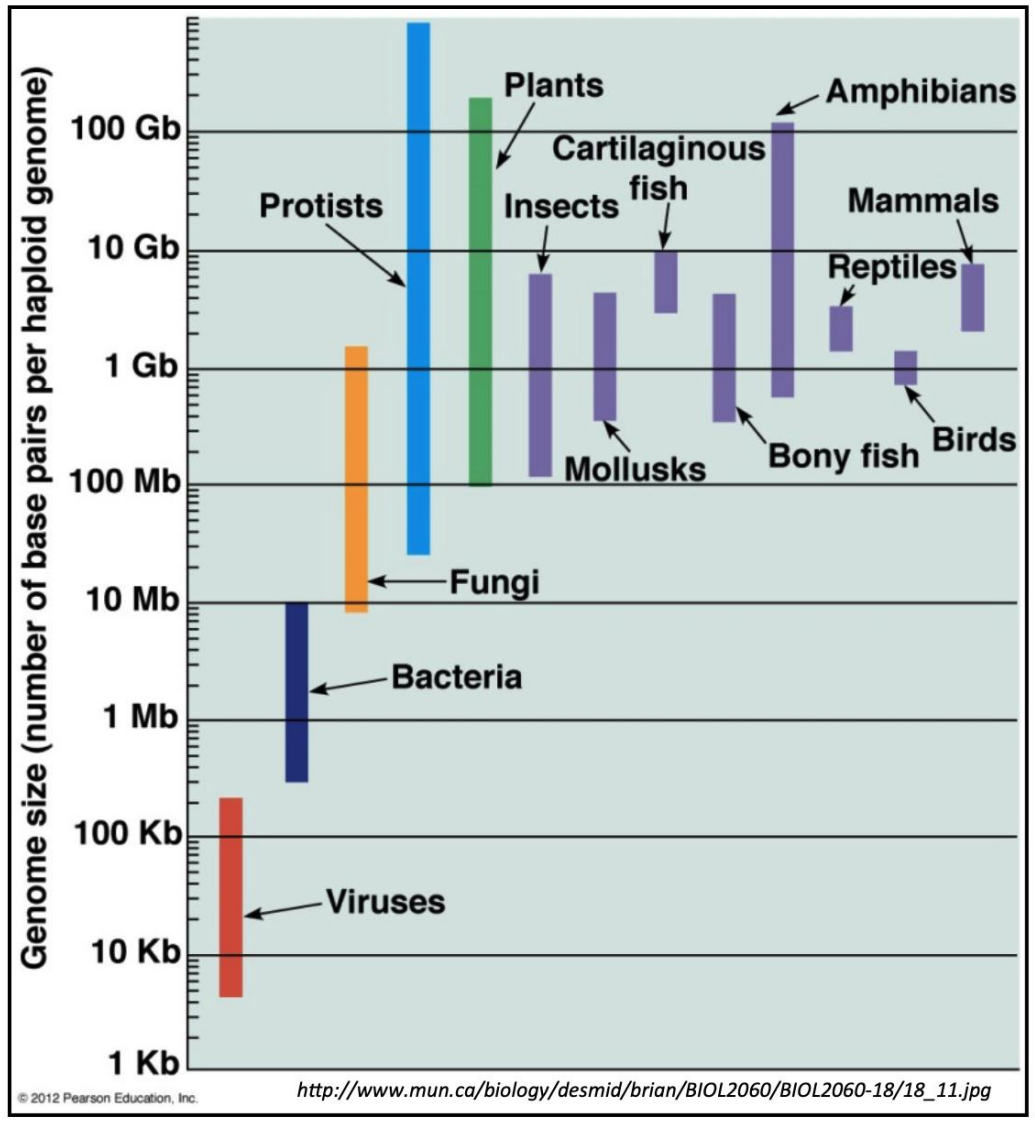
СБОРКА

ГЕРАСИМОВ ЕВГЕНИЙ,
2020

КАКИМИ БЫВАЮТ ГЕНОМЫ?



http://www.mun.ca/biology/desmid/brian/BIOL2060/BIOL2060-18/18_11.jpg

Малые геномы:

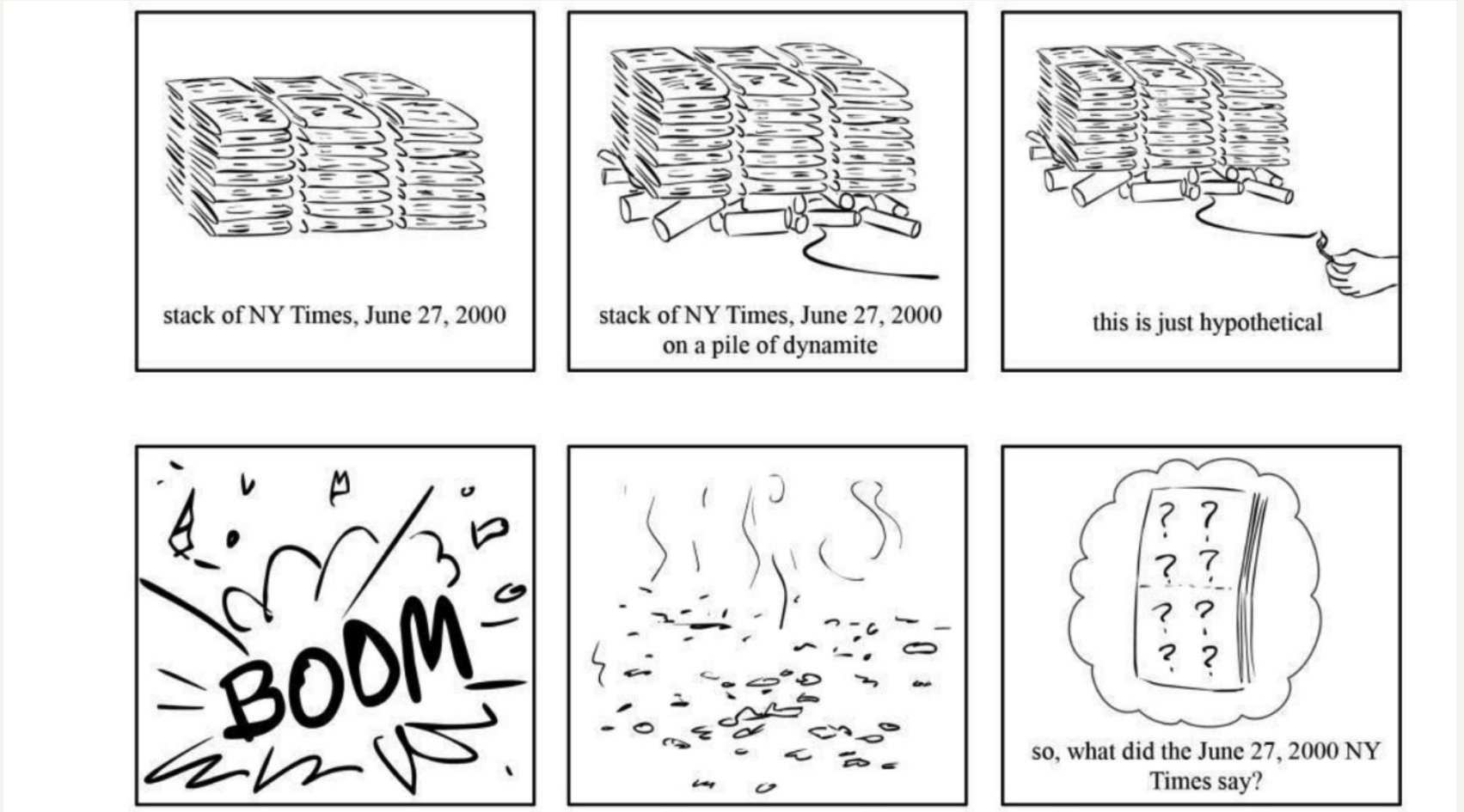
- вирусы
- бактерии
- органеллы

Несколько примеров

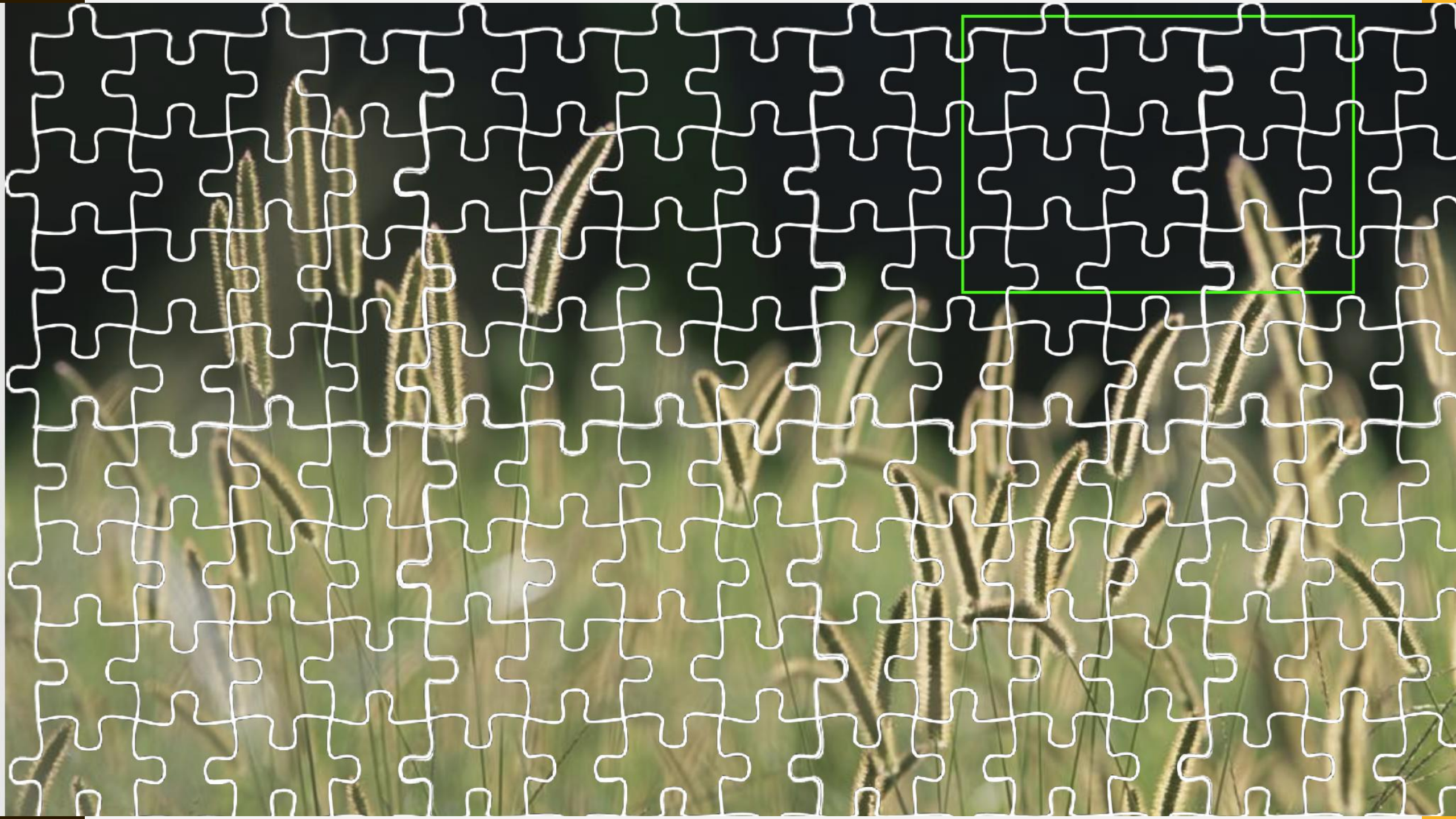
- человек (3.2 Гб) [23]
- *E. coli* (4.6 Мб) [1]
- дрозофила (139.5 Мб) [4]
- арабидопсис (135 Мб) [5]
- *Paris japonica* (150 Гб) [40]

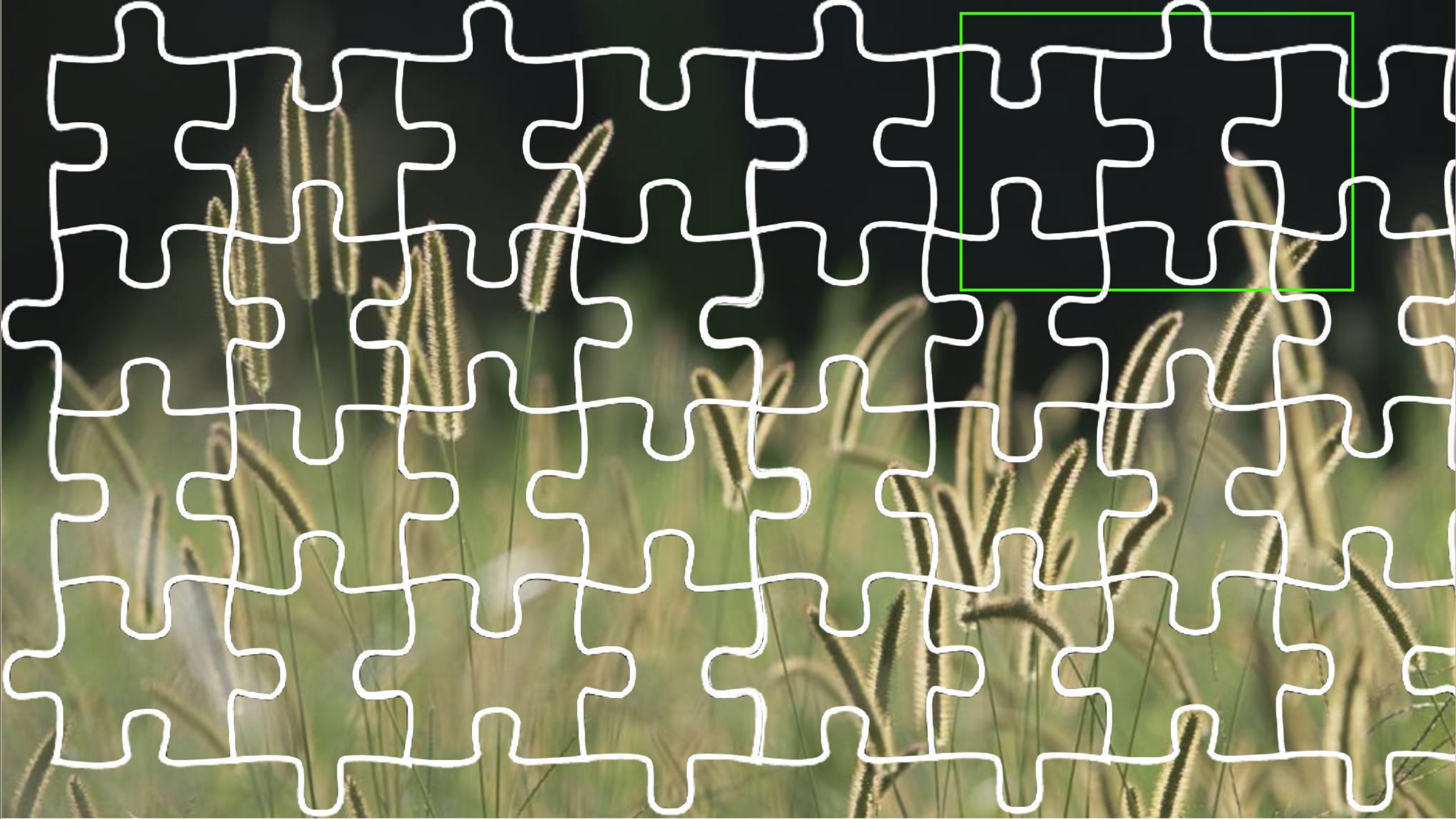
(Вороний глаз японский)

СУТЬ СБОРКИ DE NOVO



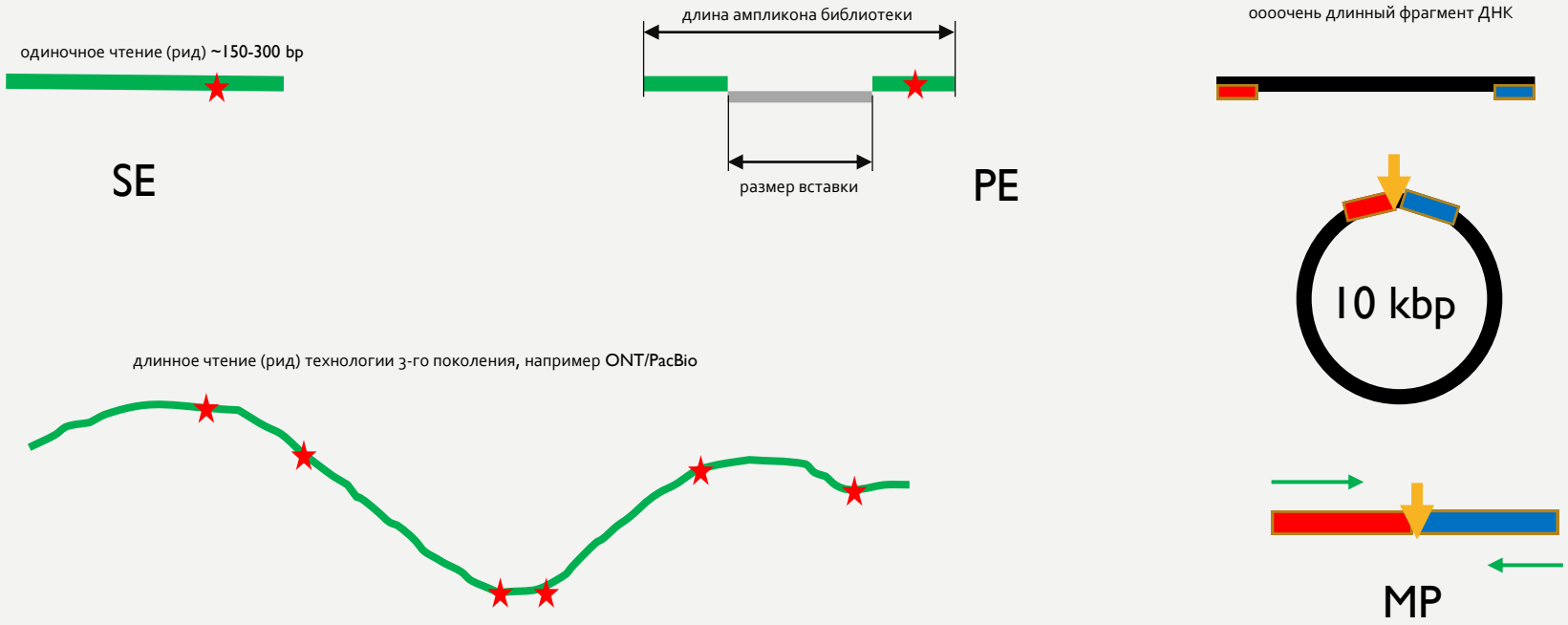
https://compscicenter.ru/media/slides/bioinformatics_2013_spring/2013_04_21_bioinformatics_2013_spring.pdf



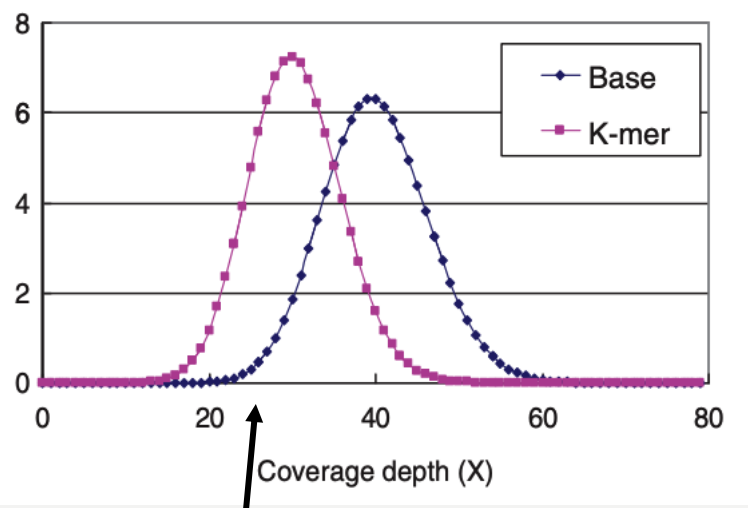


ВХОДНЫЕ ДАННЫЕ

- Чтения одной из платформ NGS, обычно Illumina или PacBio (реже 454 или ONT).



ПОКРЫТИЕ



d_k = ?

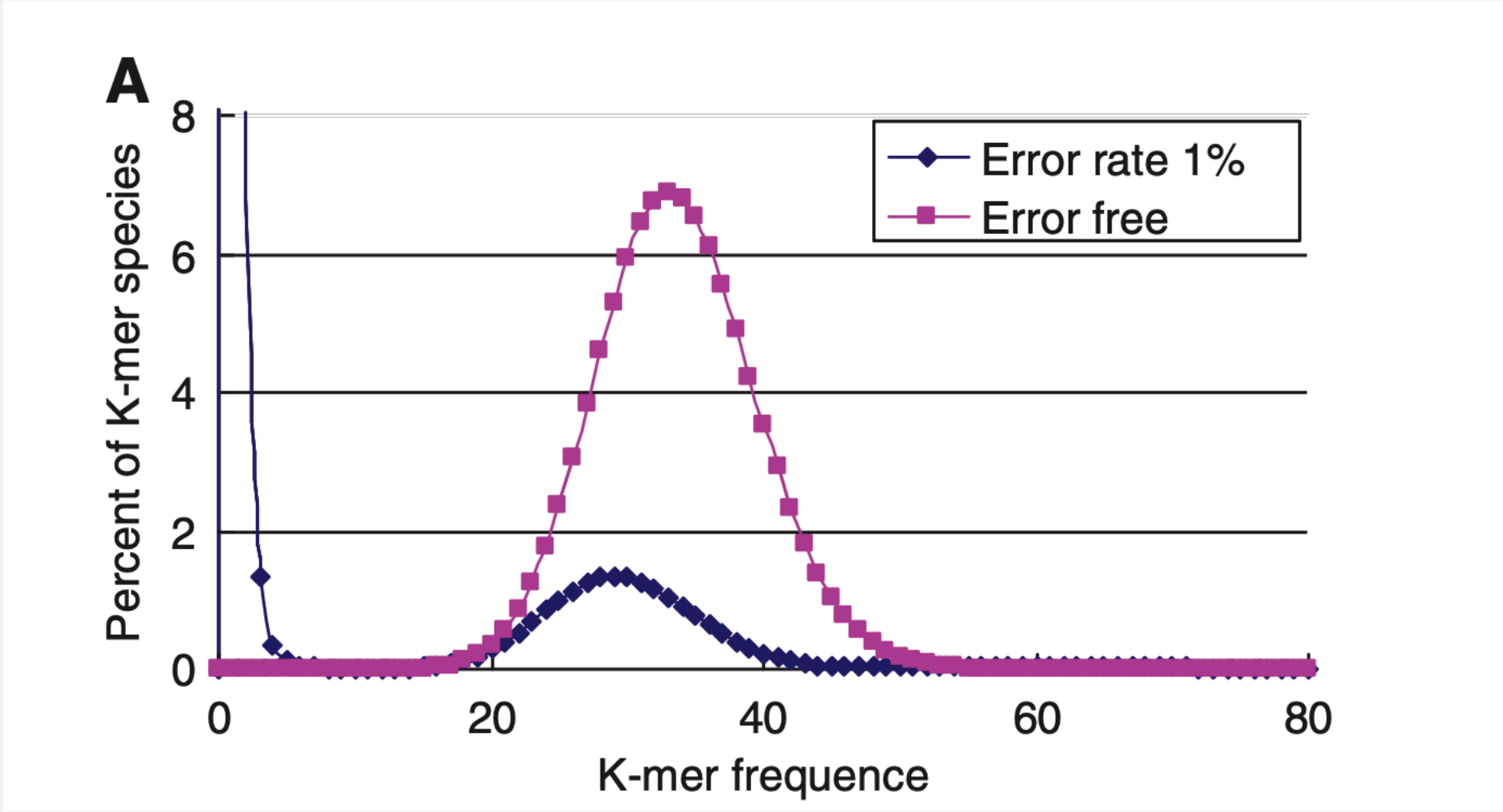
N – число ридов
L – длина рида
G – длина генома
K – длина k-мера
d_b – среднее покрытие нуклеотида
d_k – среднее покрытие (частота) k-мера
n_b – число прочитанных нуклеотидов
n_k – число прочитанных k-меров

1. $n_b = N * L$
   $n_k = N * (L - K + 1)$
2. $d_b = n_b / G$
   $d_k = n_k / G$; $G = n_k / d_k$
3. $d_k = d_b * n_k / n_b = d_b * (L - K + 1) / L$

d_b = ?

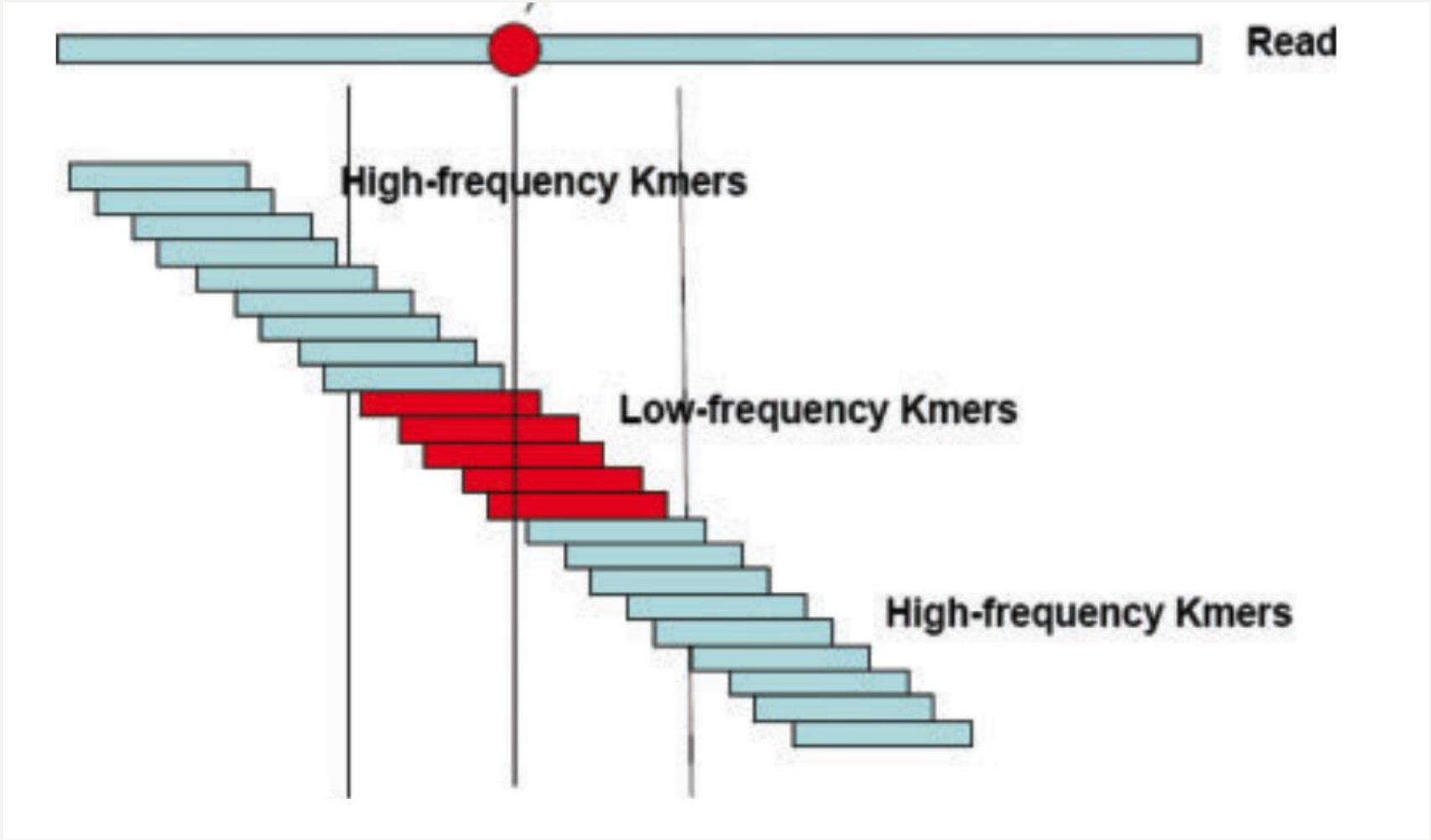
10.1093/bfgp/elr035

ОШИБКИ В РИДАХ



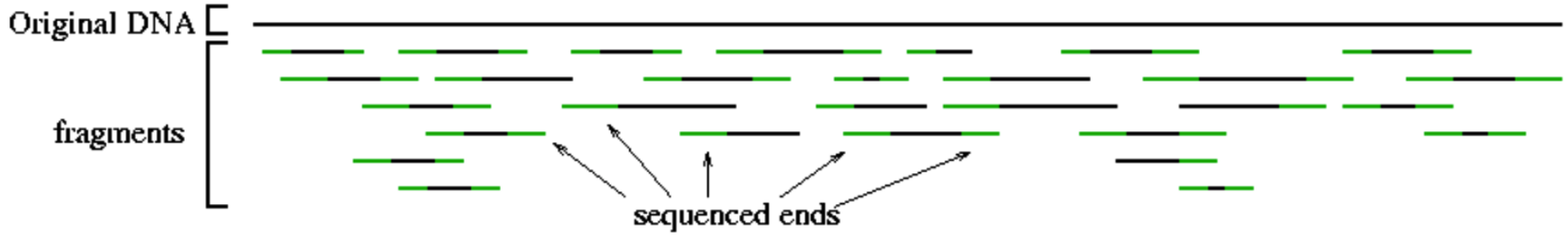
10.1093/bfgp/elr035

ИДЕЯ ДЛЯ КОРРЕКЦИИ ОШИБОК



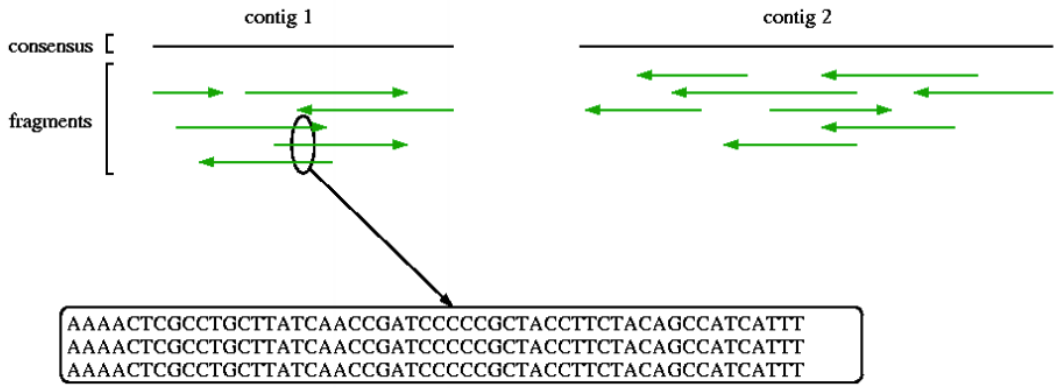
10.1093/bfgp/elr035

СБОРКА И СКАФФОЛДИНГ



(секвенирование)

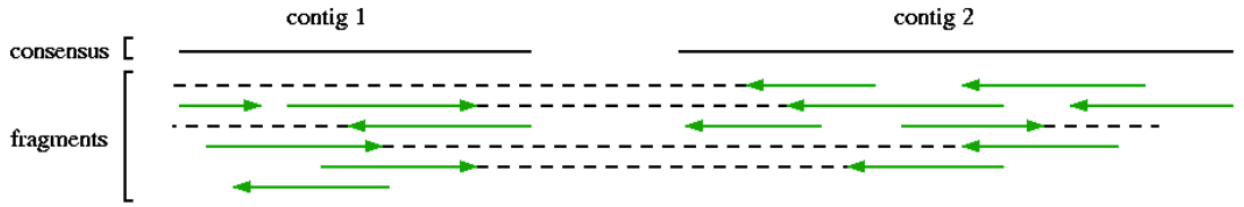
(контиг)



Assembly

(сборка)

(скаффолд из 2 контигов)

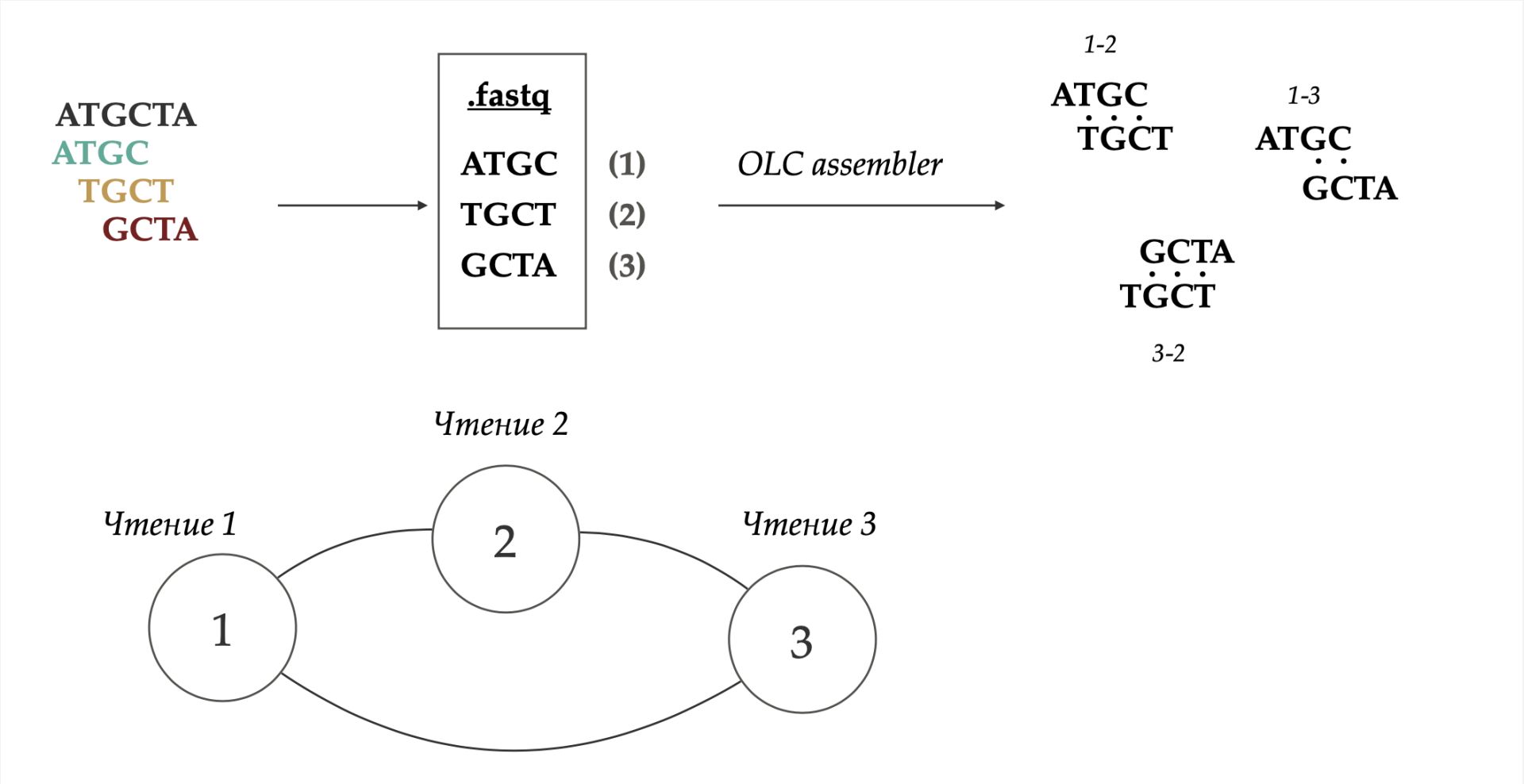


Scaffolding

(скаффолдинг)

http://www.cbcb.umd.edu/confcour/CMSC423-materials/Genome_assembly.pdf

OVERLAP-CONSENSUS-LAYOUT



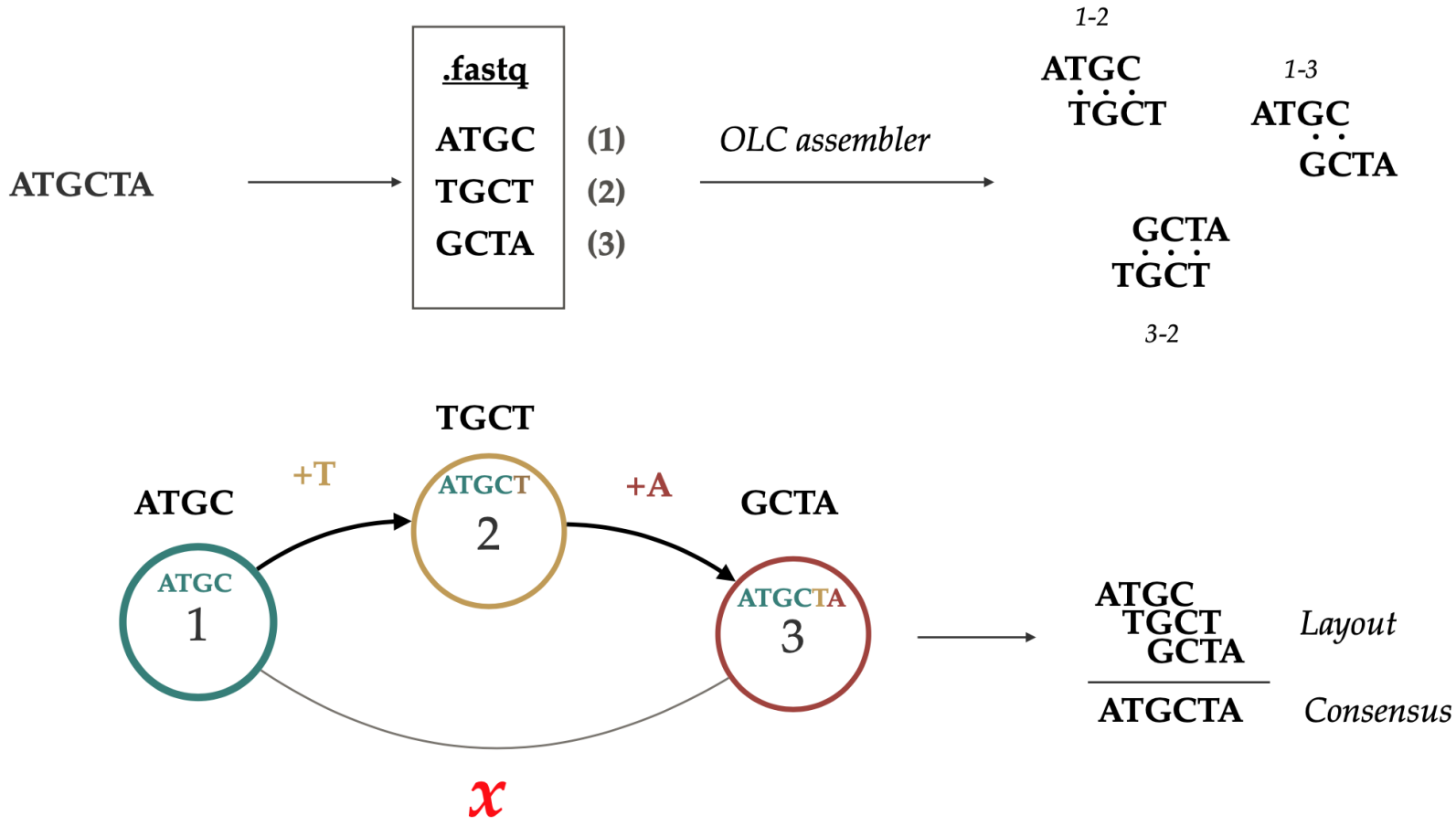
ПРИНЦИП OLC

Найти все возможные перекрывания между ридами (сложно).
На данном этапе используются суффиксные деревья, динамическое программирование.
Допускается выравнивание, содержащее определенное число замен и гэпов
(один из ридов может содержать ошибки)

Строится таблица (граф) перекрываний, в процессе выбираются только надежные перекрывания (обычно устанавливается критерий минимальной длины, % идентичности)

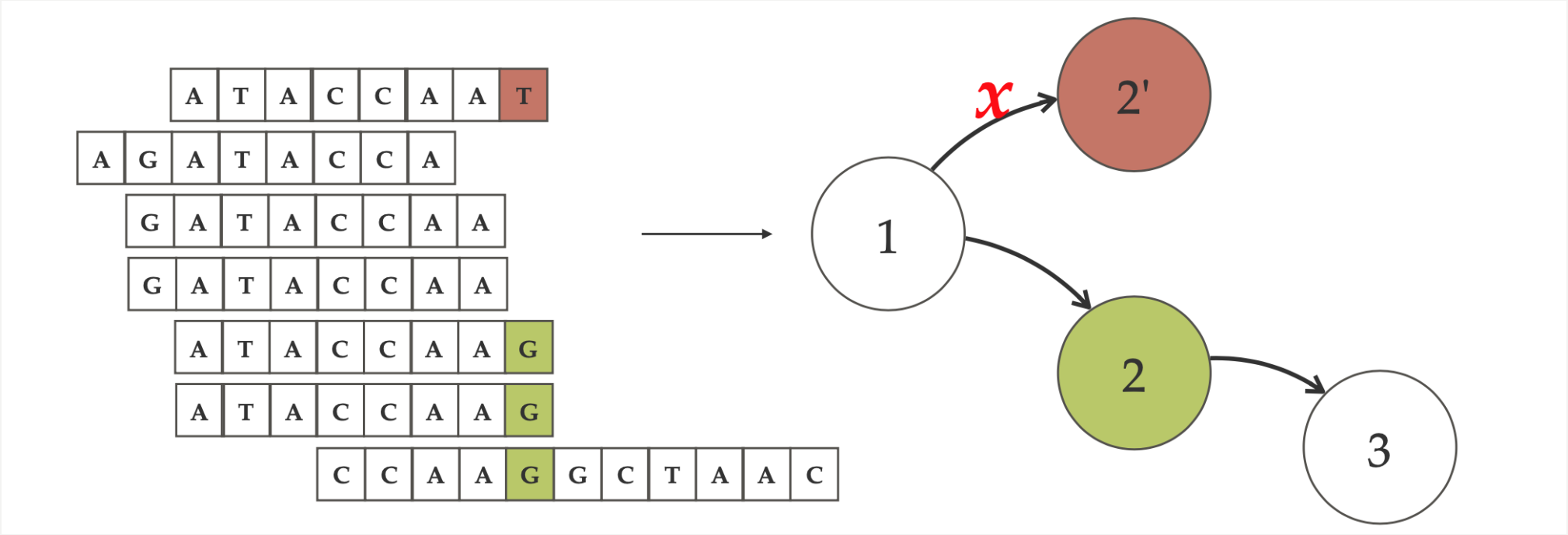
Риды выравниваются друг с другом согласно графу перекрываний, образуя консенсусную последовательность

OVERLAP-CONSENSUS-LAYOUT

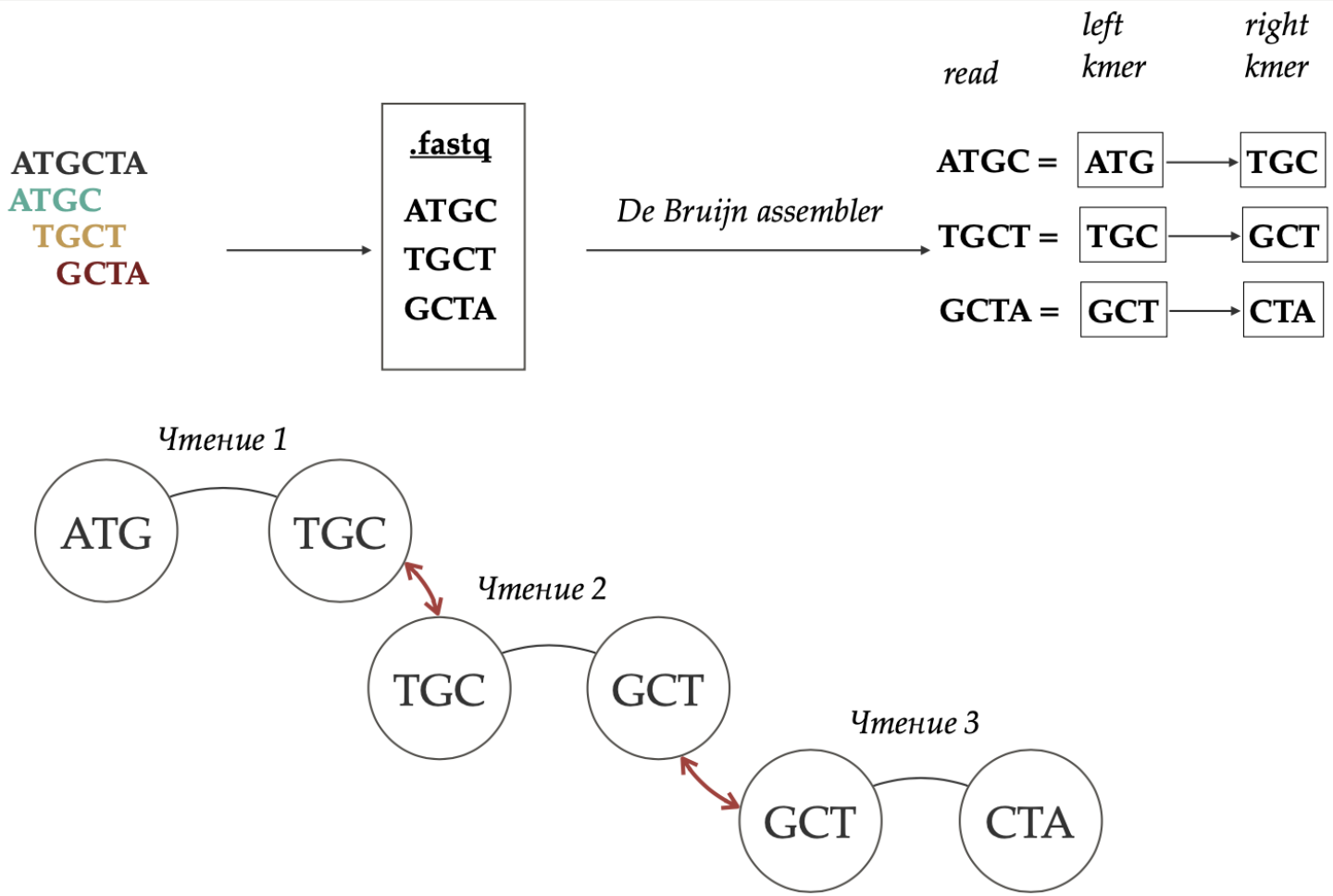


Минимальное перекрытие: 3

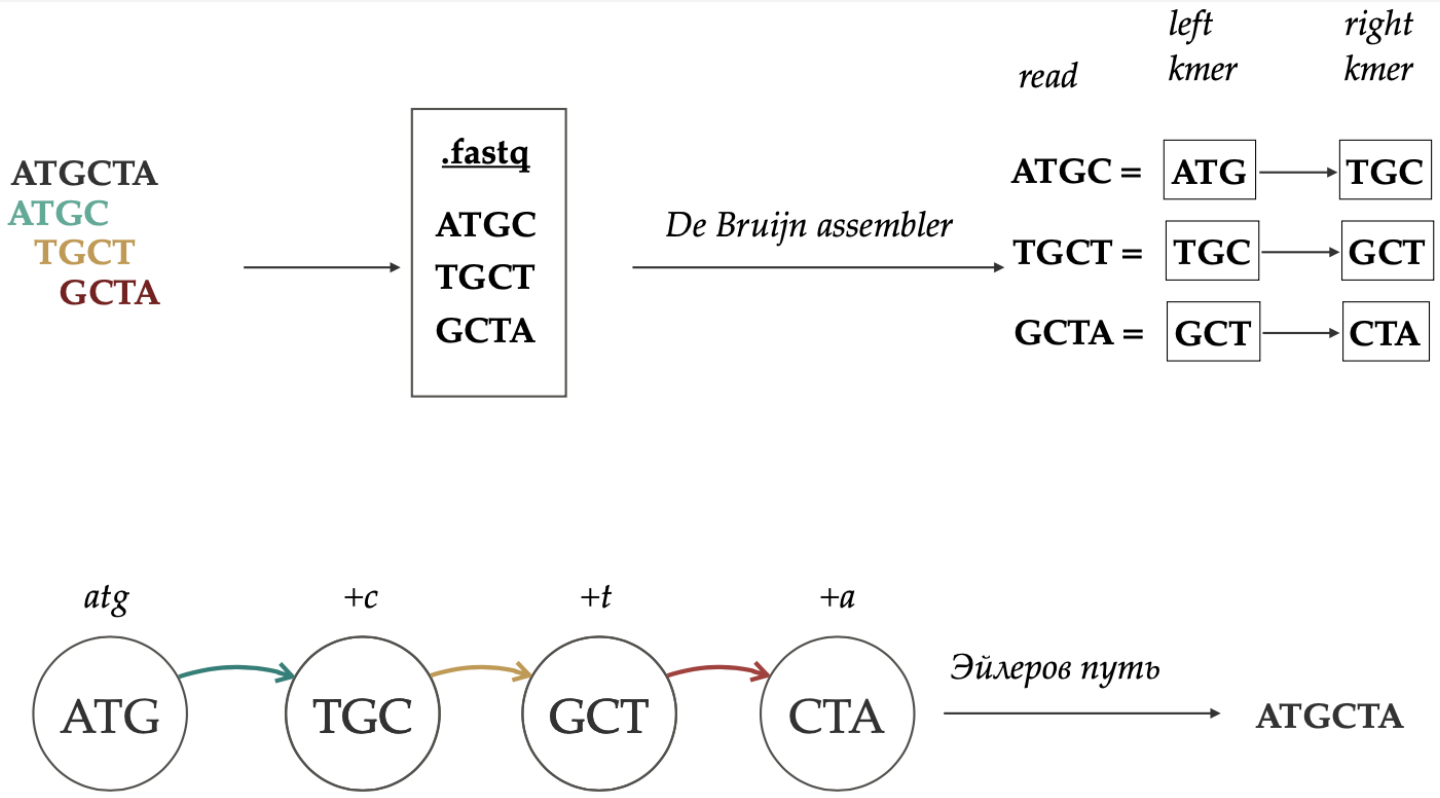
РАБОТА С LAYOUT



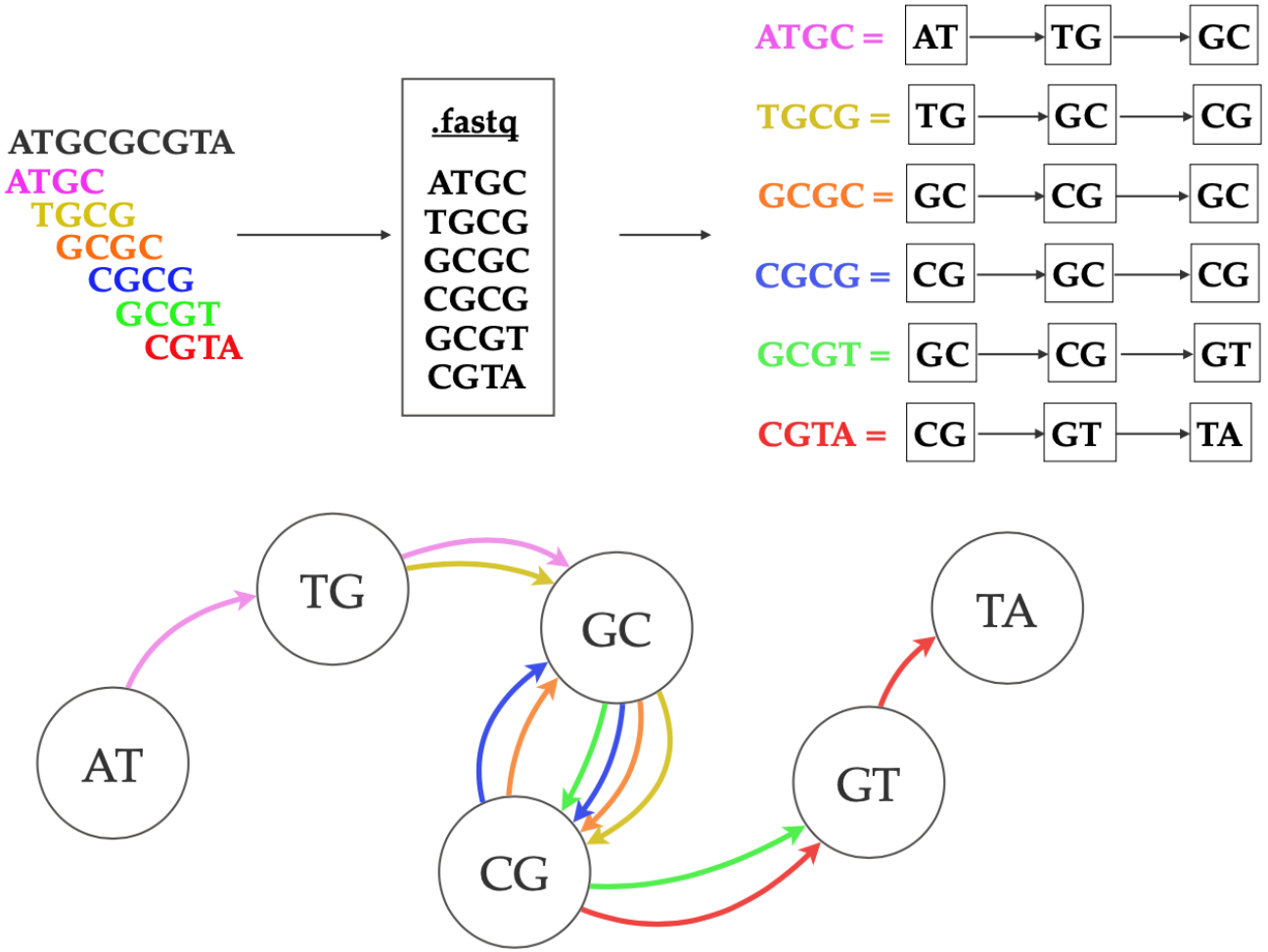
ГРАФЫ ДЕ БРЁЙНА (DBG)



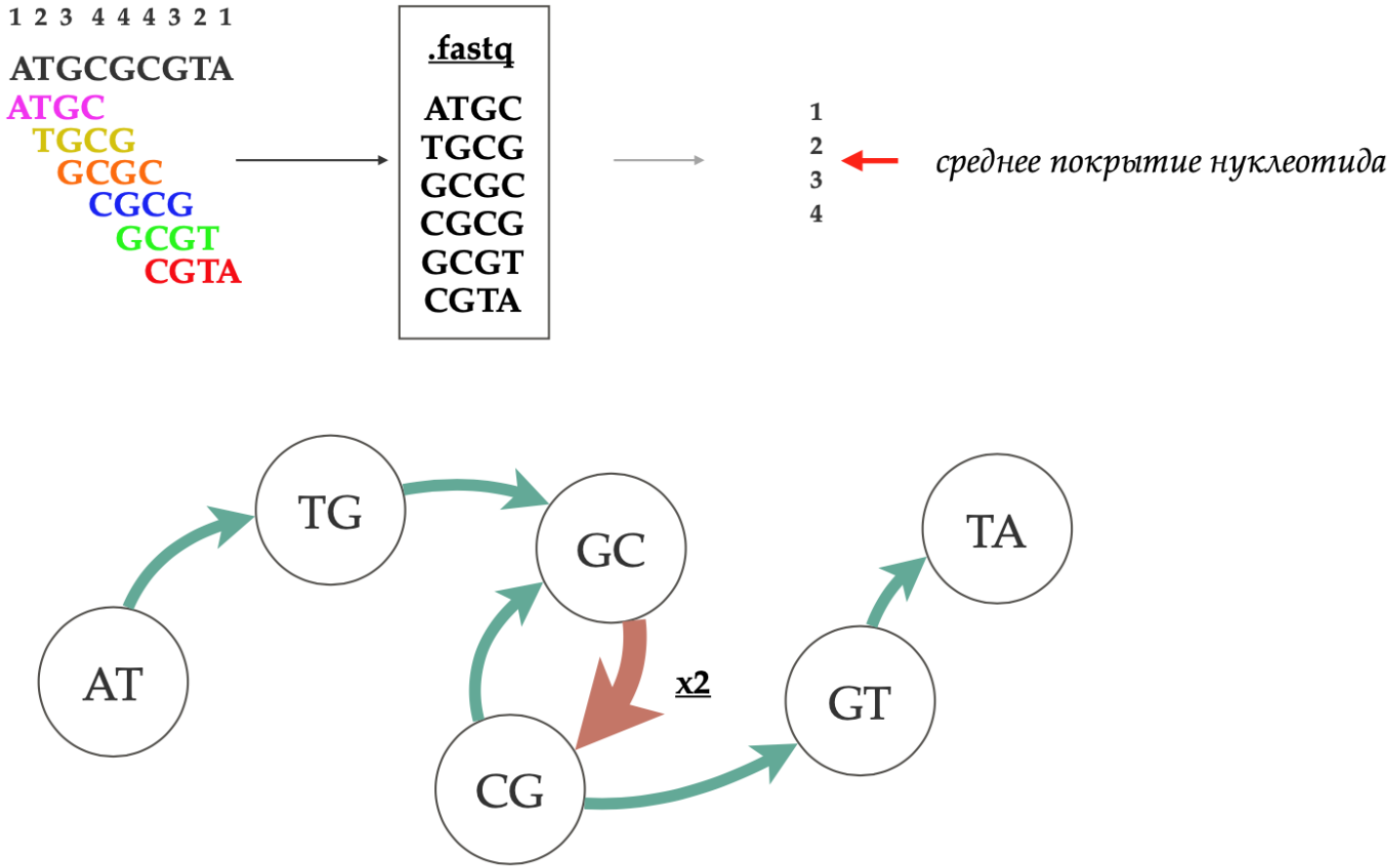
ГРАФЫ ДЕ БРЁЙНА (DBG)



ГРАФЫ ДЕ БРЁЙНА (DBG)



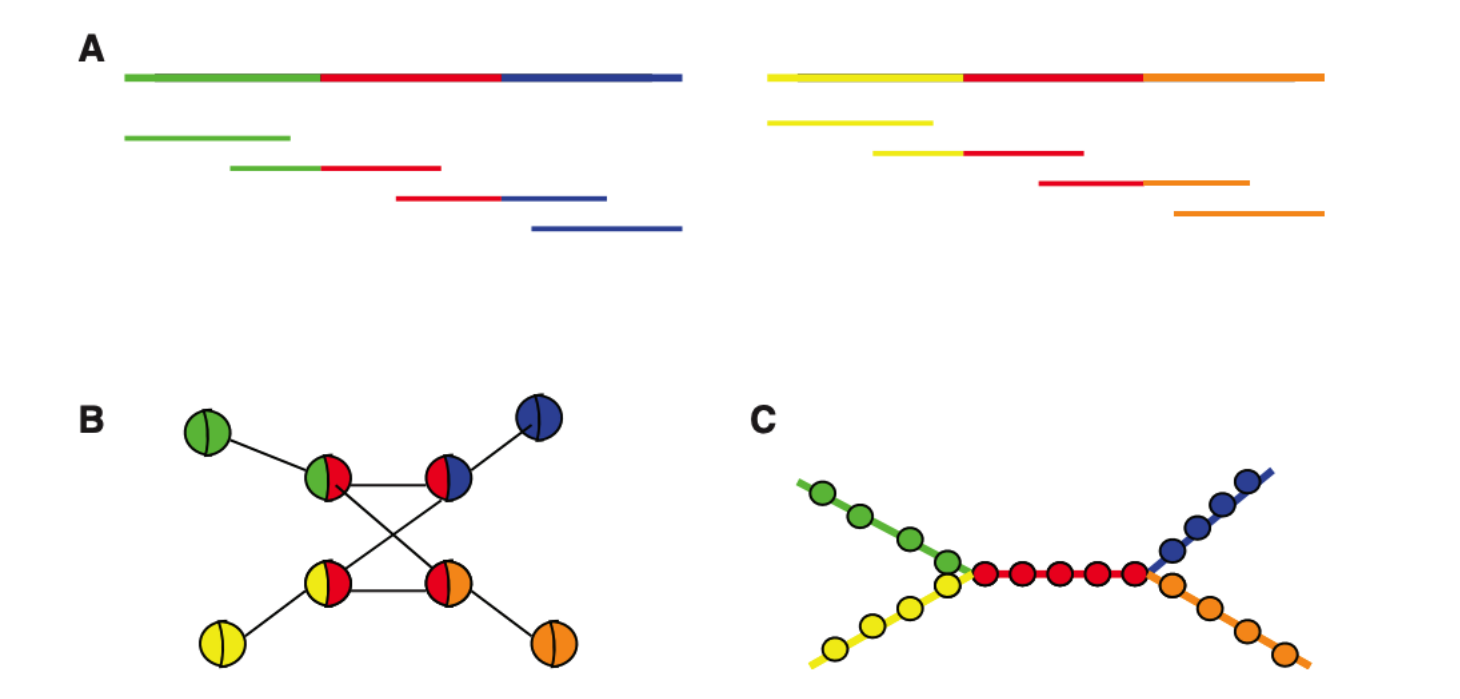
ГРАФЫ ДЕ БРЁЙНА (DBG)



OLC и DBG

OLC:
рид – узел, перекрывание - ребро

DBG:
рид – ребро или ребра, kmer - узел



10.1093/bfgp/elr035

РАБОТА С ПОВТОРАМИ

Кроме ошибок в данных и естественной гетерогенности (например, гетерозиготные состояния аллеля в диплойдном геноме или естественный полиморфизм в популяции), повторы – основной источник сложности при сборке

Существуют методы маскирования повторов до сборки

AGCTTTTTTGCA

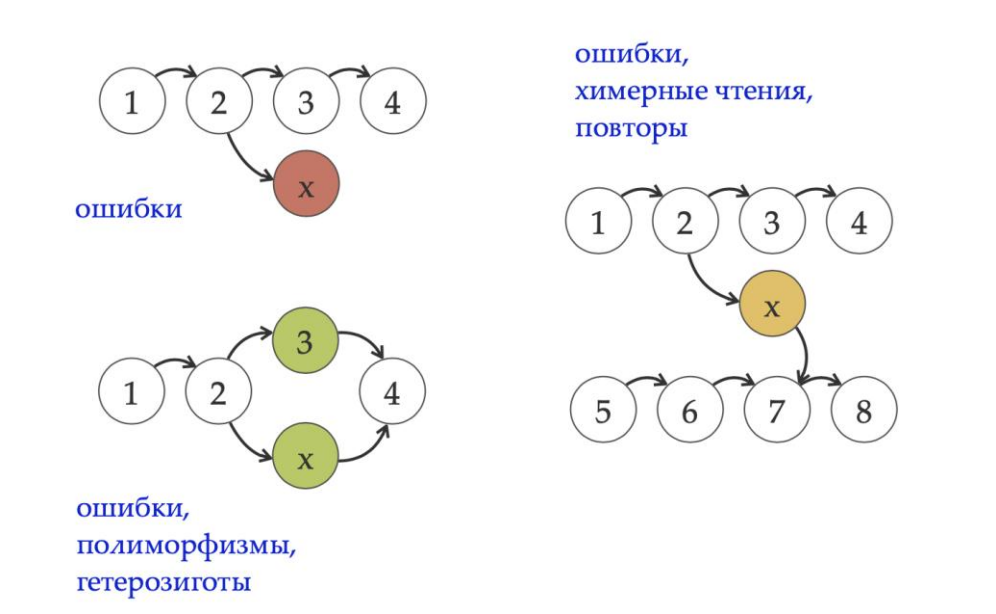
AGC → TTT → GCA

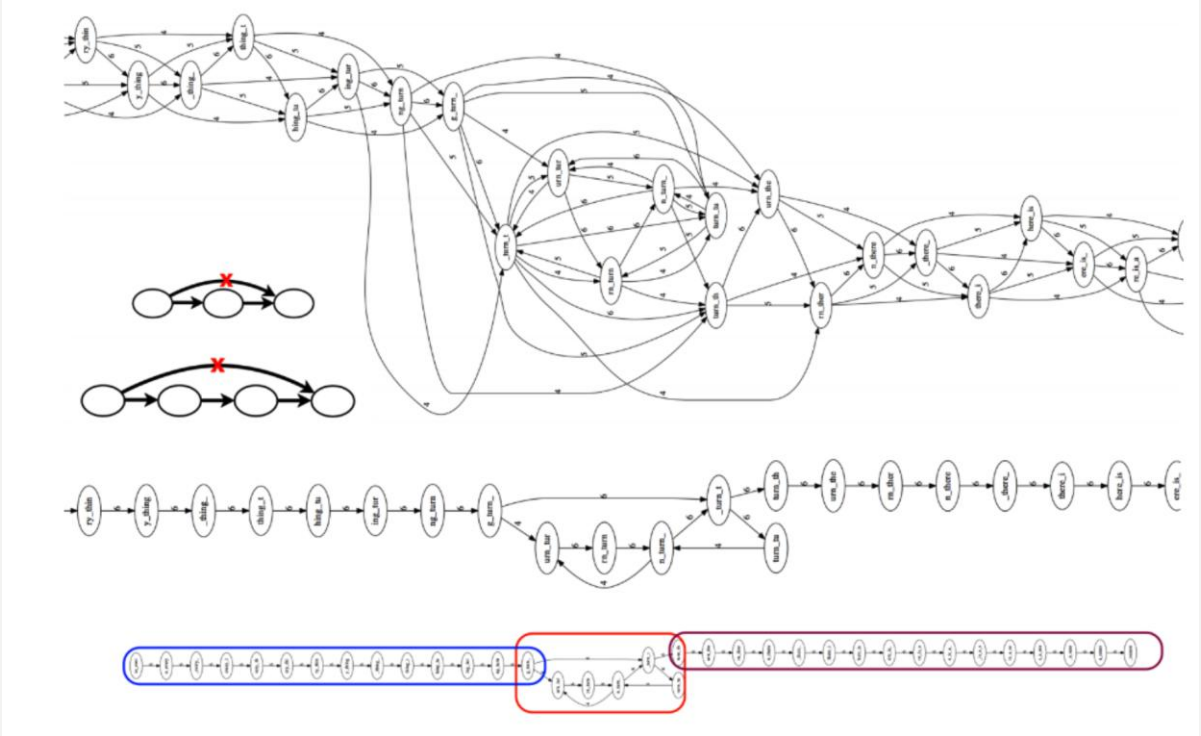
DBG: повторы не сильно усложняют граф, так как риды, приходящие из повтора, дают уже существующие в графе kmers

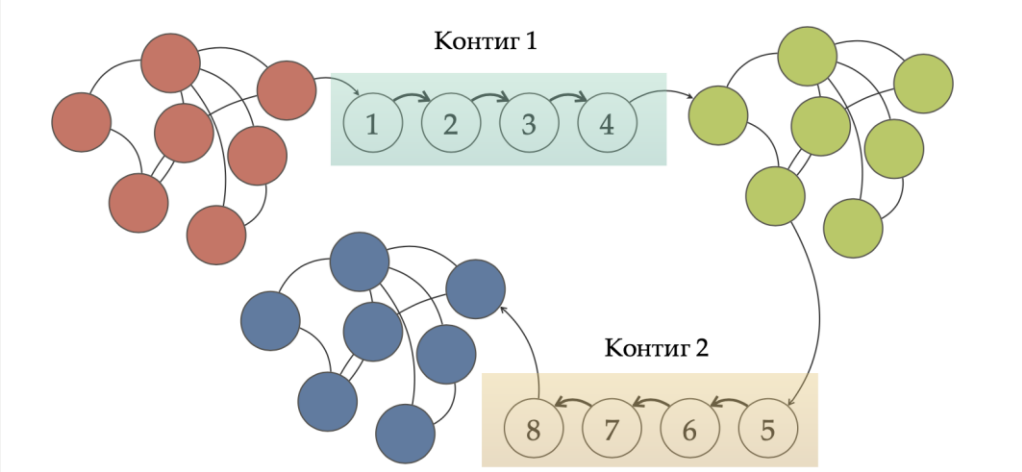
AGCTTTTTTGCA
CTTT
TTTTT
TTTTT
TTGCA

OLC: повторы усложняют граф, так как каждый рид дает в графе один новый узел; повторы увеличивают сложность вычисления, так как риды повторов имеют множество перекрываний (создают много ребер)

РАБОТА С ГРАФОМ: КОНТИГИ



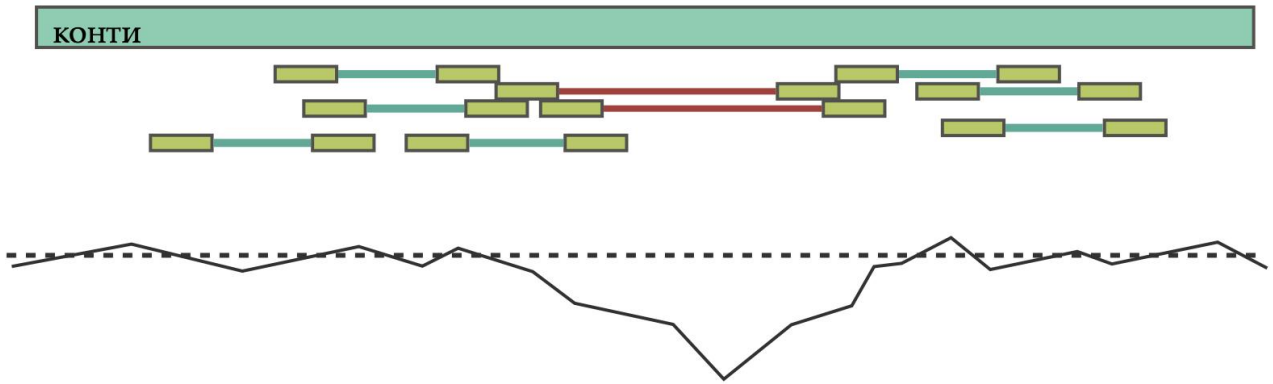




КАЧЕСТВО СБОРКИ

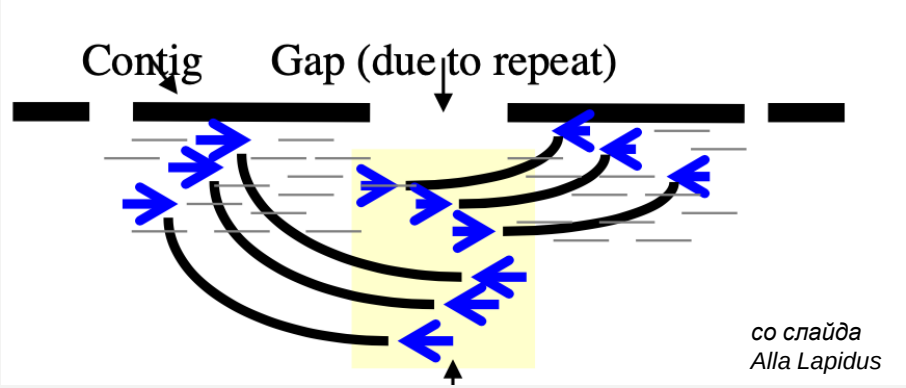
Статистика **N50**

Картирование чтений

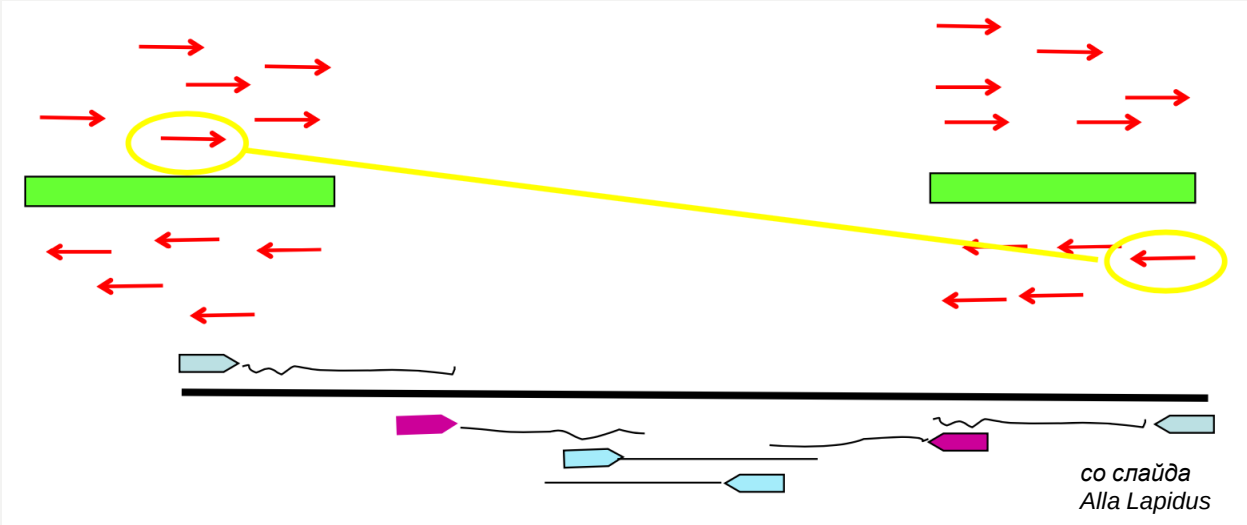


УЛУЧШЕНИЕ СБОРКИ

Локальная пересборка



Локальное досеквенирование



ДРУГИЕ ВИДЫ И СПОСОБЫ СБОРКИ

- Оптическое картирование
- Генетические карты
- Карты контактов HiC
- По имеющемуся референсу
- Транскриптомная сборка

КАЖДЫЙ БИОИНФОРМАТИК РАЗ В ЖИЗНИ…

Название	Алгоритм	Технологии	Авторы	Представлен
ABySS	De Bruijn	Solexa, SOLiD	Simpson, J. et al.	2008
ALLPATHS-LG	De Bruijn	Solexa, SOLiD	Gnerre, S. et al.	2011
Celera WGA Assembler / CABOG	OLC	Sanger, 454, Illumina	Myers, G. et al.; Miller G. et al.	2004
CLC Genomics Workbench	String Graph	Sanger, 454, Solexa, SOLiD	CLC bio	2008
Edena	OLC	Illumina	D. Hernandez et al.	2008
Euler	De Bruijn	Sanger, 454 (,Solexa ?)	Pevzner, P. et al.	2001
Euler-sr	De Bruijn	454, Solexa	Chaisson, MJ. et al.	2008
IDBA	De Bruijn	Sanger,454,Solexa	Yu Peng, Henry C. M. Leung, Siu-Ming Yiu, Francis Y. L. Chin	2010
MIRA	OLC	Sanger, 454, Solexa	Chevreux, B.	1998
Newbler	String Graph	454, Sanger, Solexa, Ion	454/Roche	2009
PCAP	OLC	Sanger, 454	Huang et al.	2003
SGA	String Graph	Illumina, Ion Torrent	Simpson, J. et al.	2011
SOPRA		Illumina, SOLiD, Sanger, 454	Dayarian, A. et al.	2010
SOAPdenovo	De Bruijn	Solexa	Li, R. et al.	2009
SPAdes	De Bruijn	Illumina, Solexa	Bankevich, A et al.	2012
Velvet	De Bruijn	Sanger, 454, Solexa, SOLiD	Zerbino, D. et al.	2007