De novo сборка транскриптомов

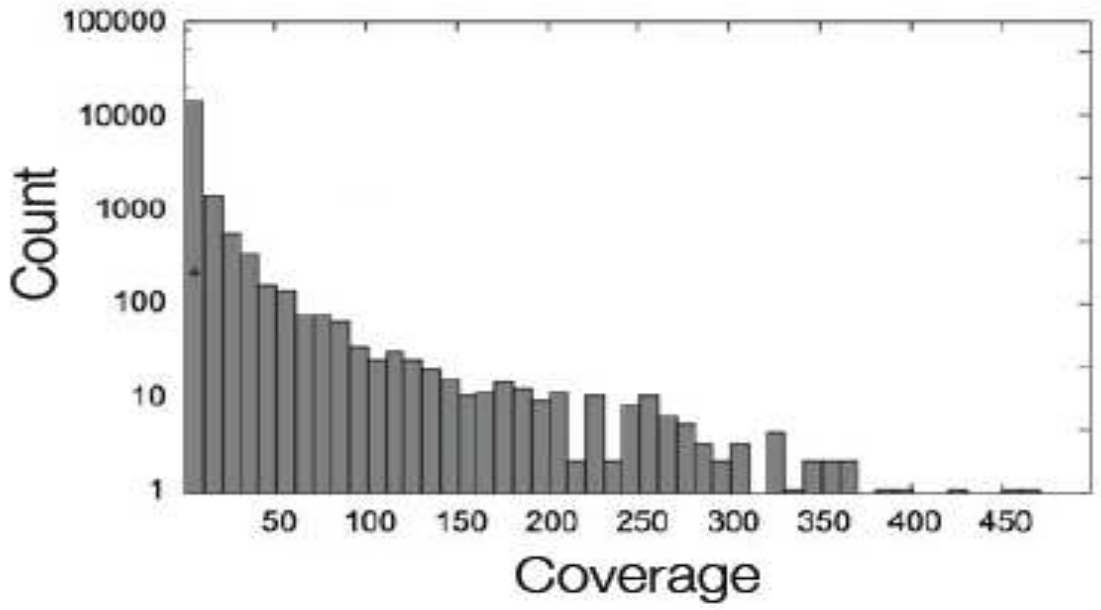
Касьянов Артем
н.с. Лабораториии вычислительной генетики и системной биологии ИОГен РАН

De novo секвенирование транскриптома vs *de novo* секвенирование генома

- **Геномы не модельных организмов** могут быть достаточно **сложными** для **восстановления(большое число повторов**, **полиплоидность**, **большой размер**).
- Секвенирование транскриптома позволяет **быстро получить доступ к информации о генах и белках**, использующихся для функционирования организма.
- В большинстве случаев для более **точной аннотации генома** все равно потребуются **транскриптомные данные**.
- Секвенирование траснкриптома **дешевле полногеномного секвенирования**.
- Вследствие развития технологий секвенирования **растет длина рида**. На данный момент **«слитые» риды Illumina Miseq** достигают **трети средней длины транскрипта**, что значительно упрощает сборку и позволяет **использовать OLC сборщики**.

De novo секвенирование транскриптома vs *de novo* секвенирование генома

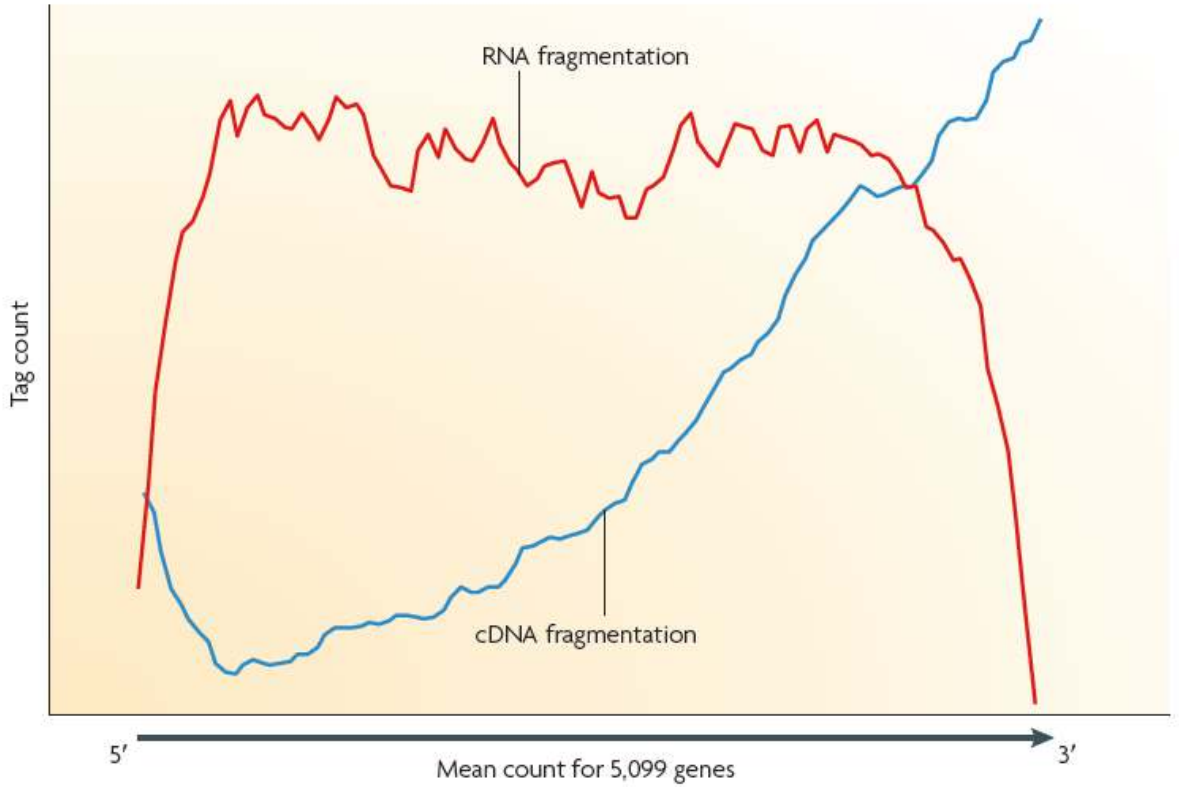
- **Транскриптомы различных тканей** взятые в разные промежутки времени **могут очень сильно отличаться**.
- **20% генов** дают **80% ридов**.



[O'Neil *et al. BMC Genomics* 2010, **11**:310]

De novo секвенирование транскриптома vs *de novo* секвенирование генома

- **Неравномерность покрытия** транскриптов.

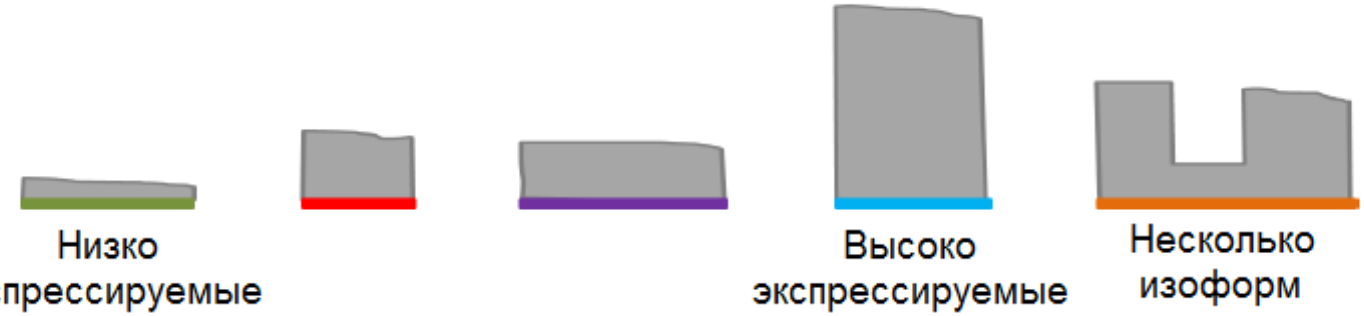


[Zhong Wang et al. Nat. Rev. Gen. 2009, Vol.10]

Сборка транскриптомов vs сборка геномов

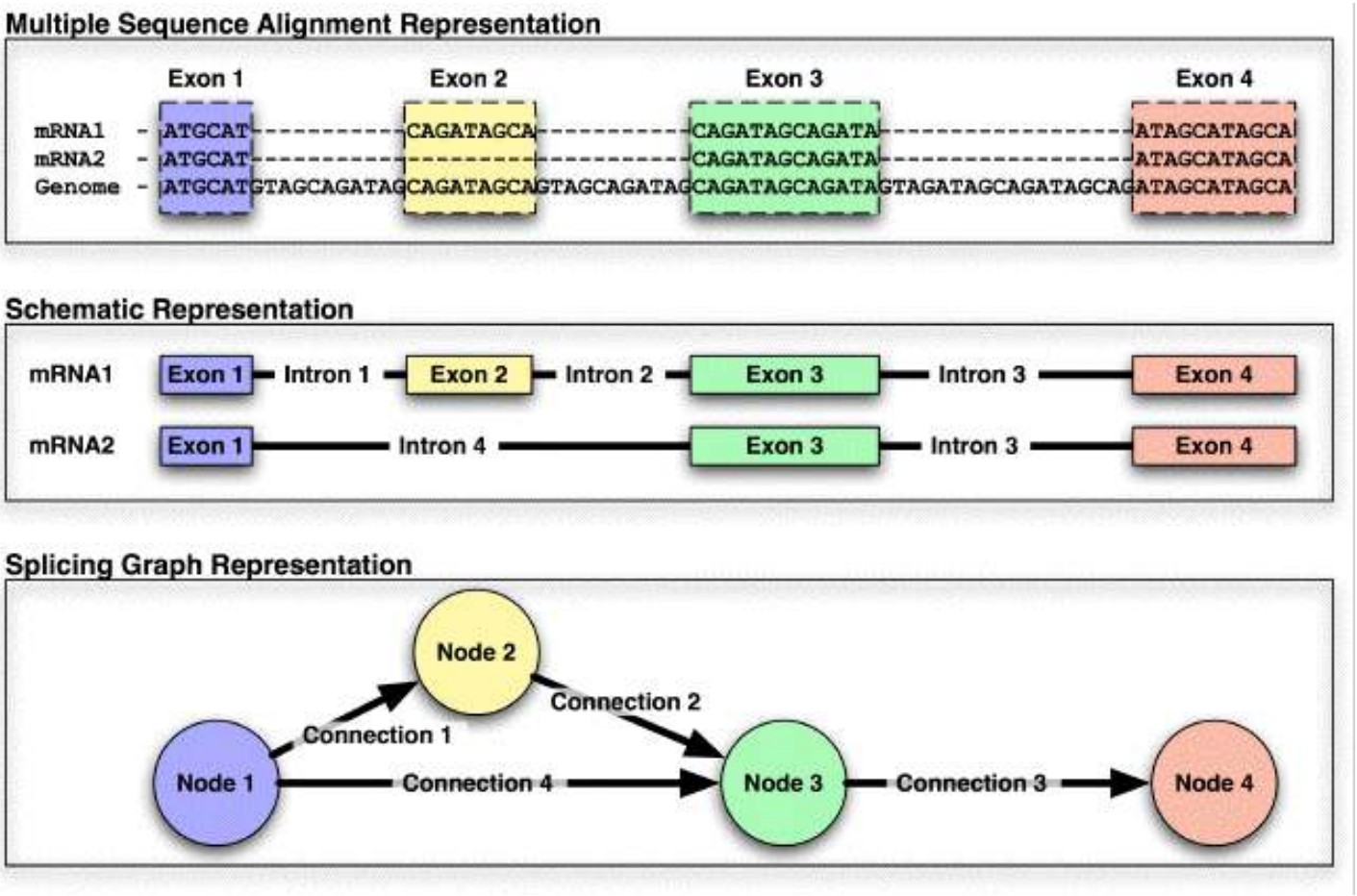
Программы для сборки геномов ожидают более-менее **равномерное распределение покрытия**.
Собранные регионы для которых наблюдается **возрастание покрытия** принимаются за **повторы**.

Но **индивидуальные гены** внутри транскриптома могут иметь **очень разное покрытие.**



[http://training.bioinformatics.ucdavis.edu/docs/2013/09/short-course-2013/_downloads/MB_RNASeq_Trans_Assembly_SC_2013.pdf]

Splicing graph

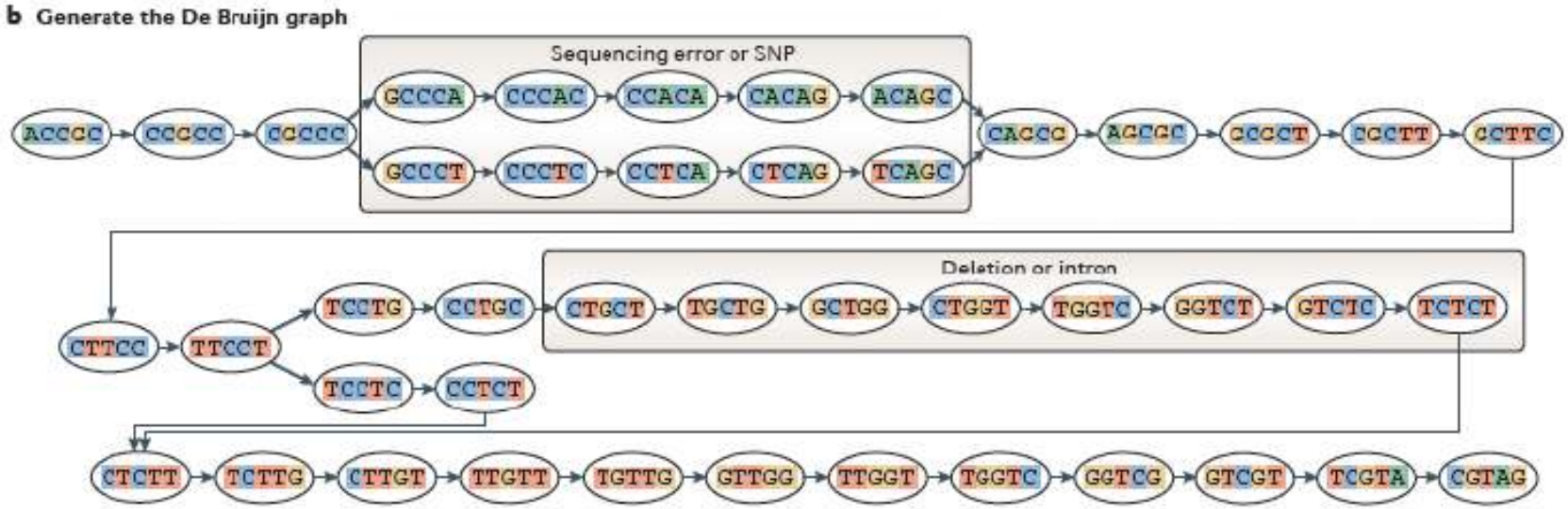


[http://proline.bic.nus.edu.sg/dedb/methodology.html]

Общая схема de novo сборки транскриптома

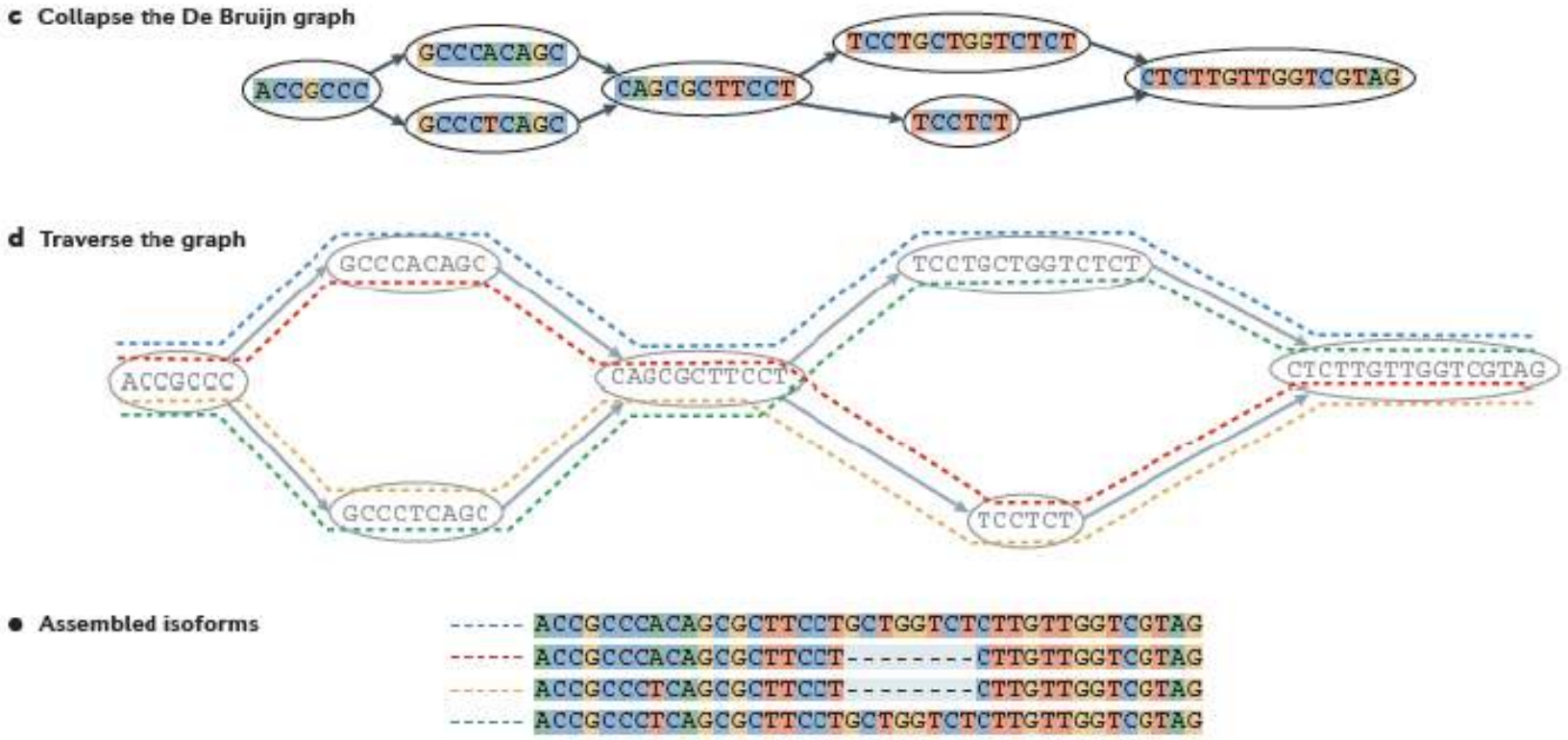
[Martin & Wang (2011) Nat. Rev. Gen. 12,671]

Общая схема de novo сборки транскриптома



[Martin & Wang (2011) Nat. Rev. Gen. 12,671]

Общая схема de novo сборки транскриптома



[Martin & Wang (2011) Nat. Rev. Gen. 12,671]

Транскриптомные сборщики

Сборщики, основанные
на DeBruijn графах.
(Illumina,SOLiD,IonTorrent)

- Trinity(Broad)
- Velevet(Oases)
- TransAbyss
- SOAPtrans

Сборщики, основанные
на OLC подходе(454, Sanger, PacBio)

- Mira3
- Est2assembly
- GS/Newbler(Roche)
- SMRT Pipe(PacBio)

De novo сборка транскриптома

- **Предобработка**

FASTQC, prinseq, trmmomatic, kmc2, kmernator2…

- **Сборка**

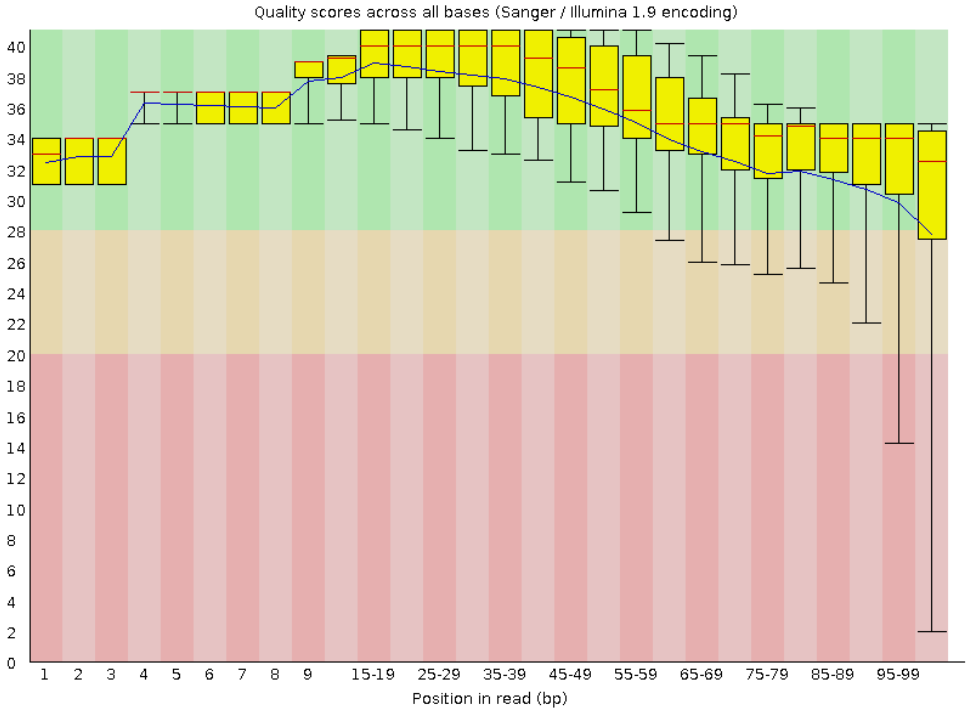
Oases, Newbler, Trinity…

- **Оценка качества сборки**

QUAST, BLAST, Prinseq, Bowtie, Transrate...

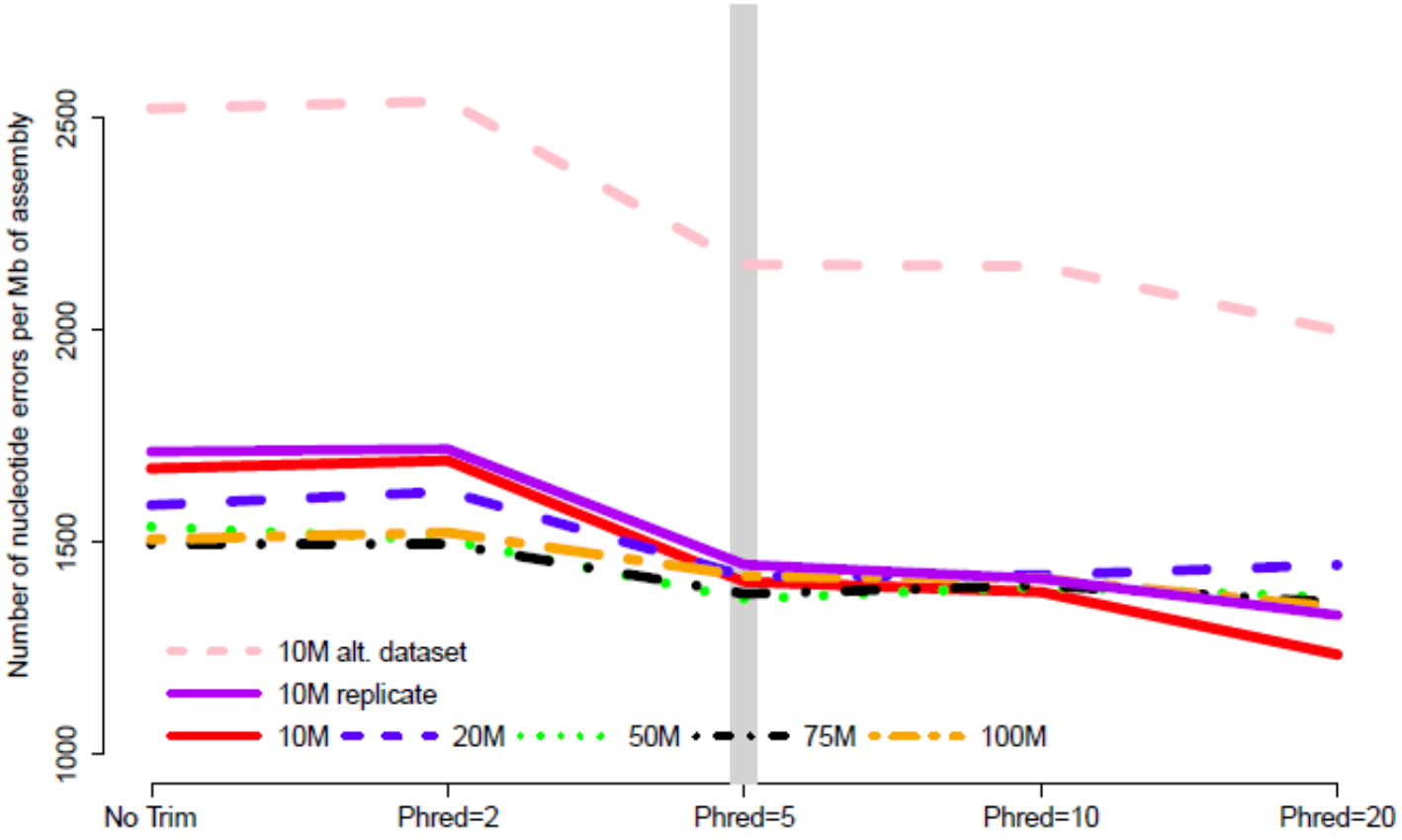
Предобработка

- **Оценка качества набора чтений.**(FASTQC)



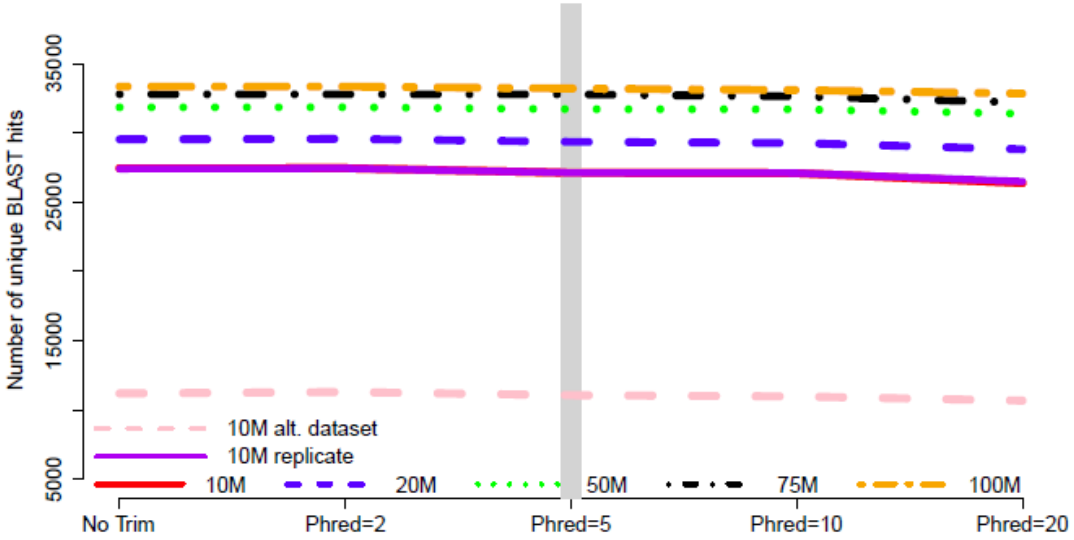
- **Триммирование ридов.**(Prinseq, Trimmomatic)

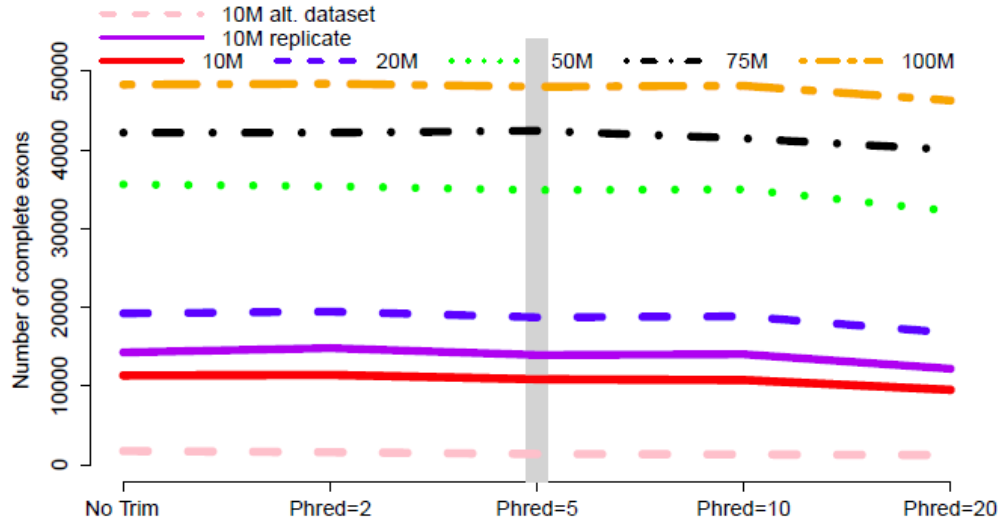
Предобработка



[Matthew D. MacManes, Front. Genet., 31 January 2014]

Предобработка



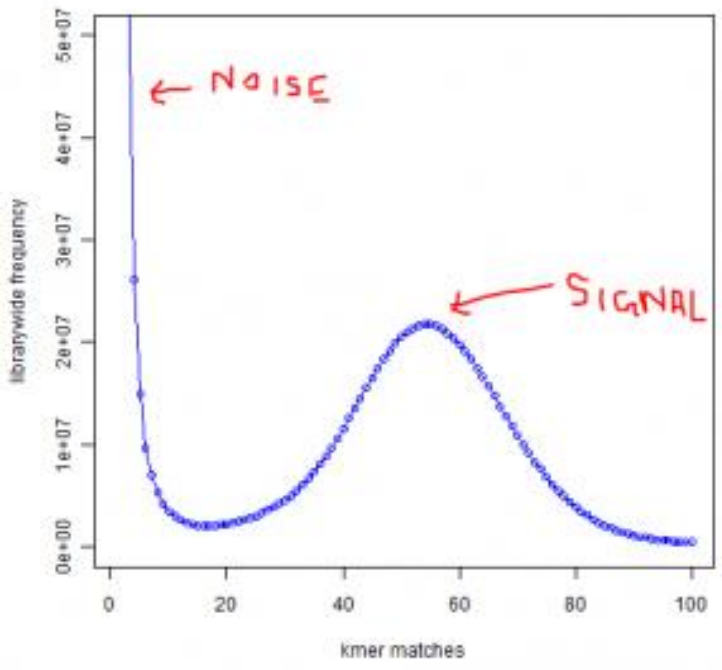


[Matthew D. MacManes, Front. Genet., 31 January 2014]

Предобработка

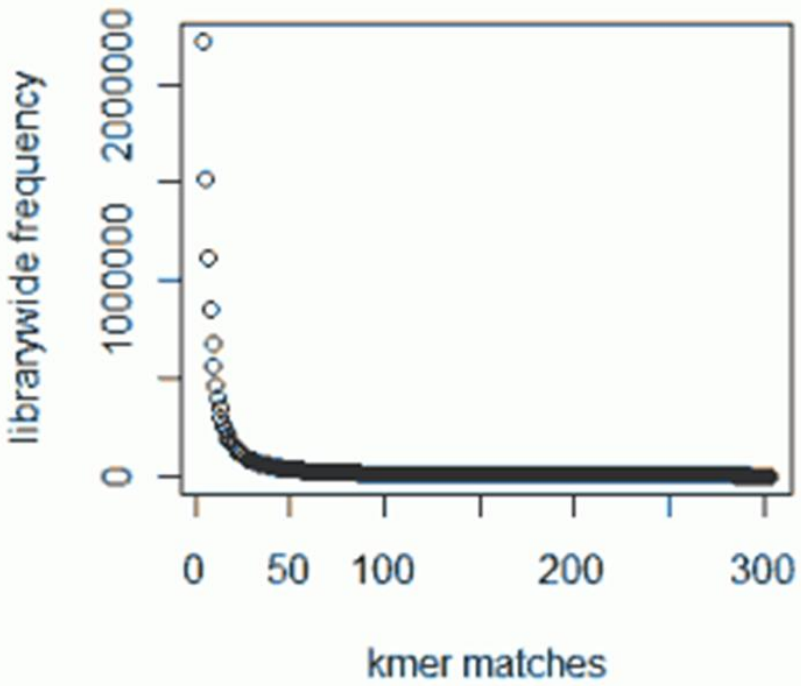
- **Фильтрация по кмерам**(kmernator2…)

Для геномных данных



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

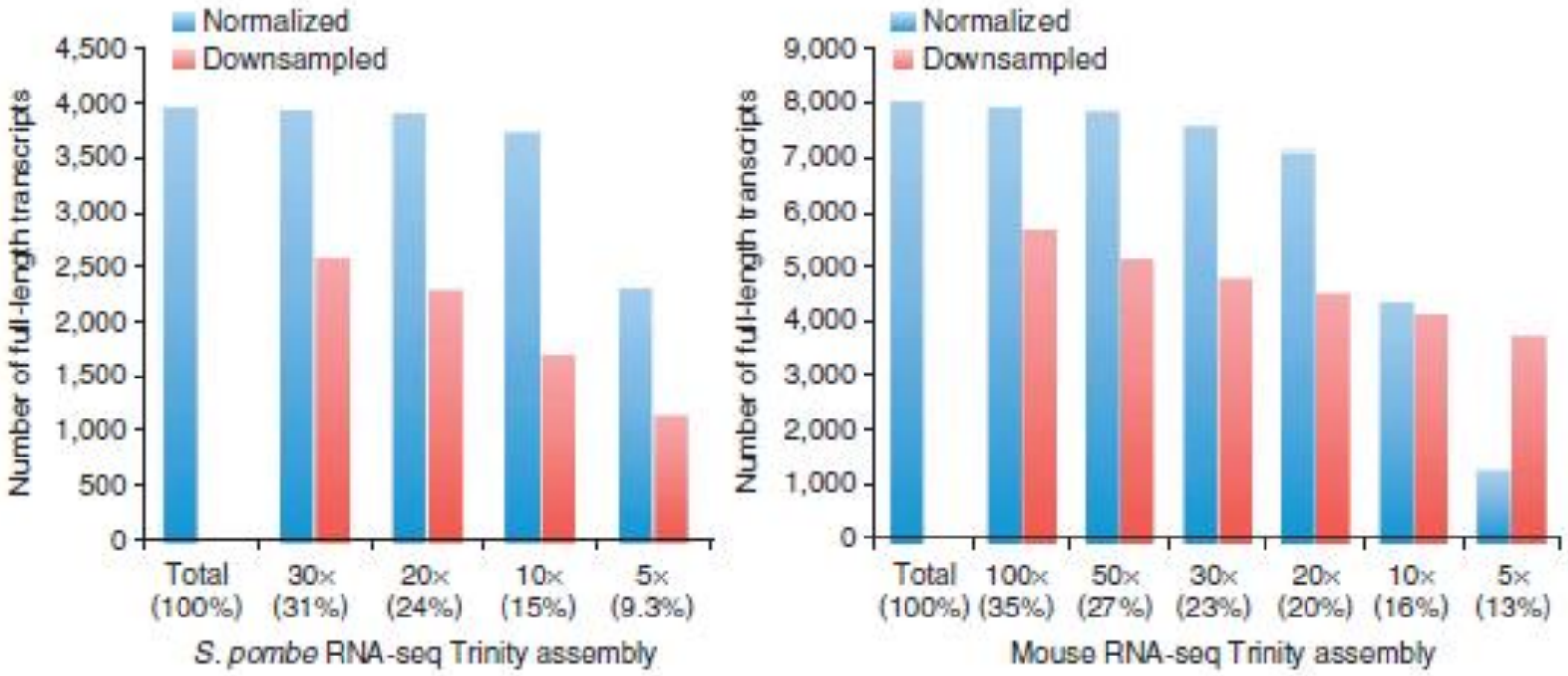
Для транскриптомных данных



[http://www.homolog.us/blogs/blog/2011/10/26/k-mer-distribution-of-a-transcriptome/]

Предобработка

- «Цифровая» нормализация



[Haas BJ et al. Nat Protoc. 2013 Aug;8(8):1494-512]

Сборка транскриптомов с помощью Trinity

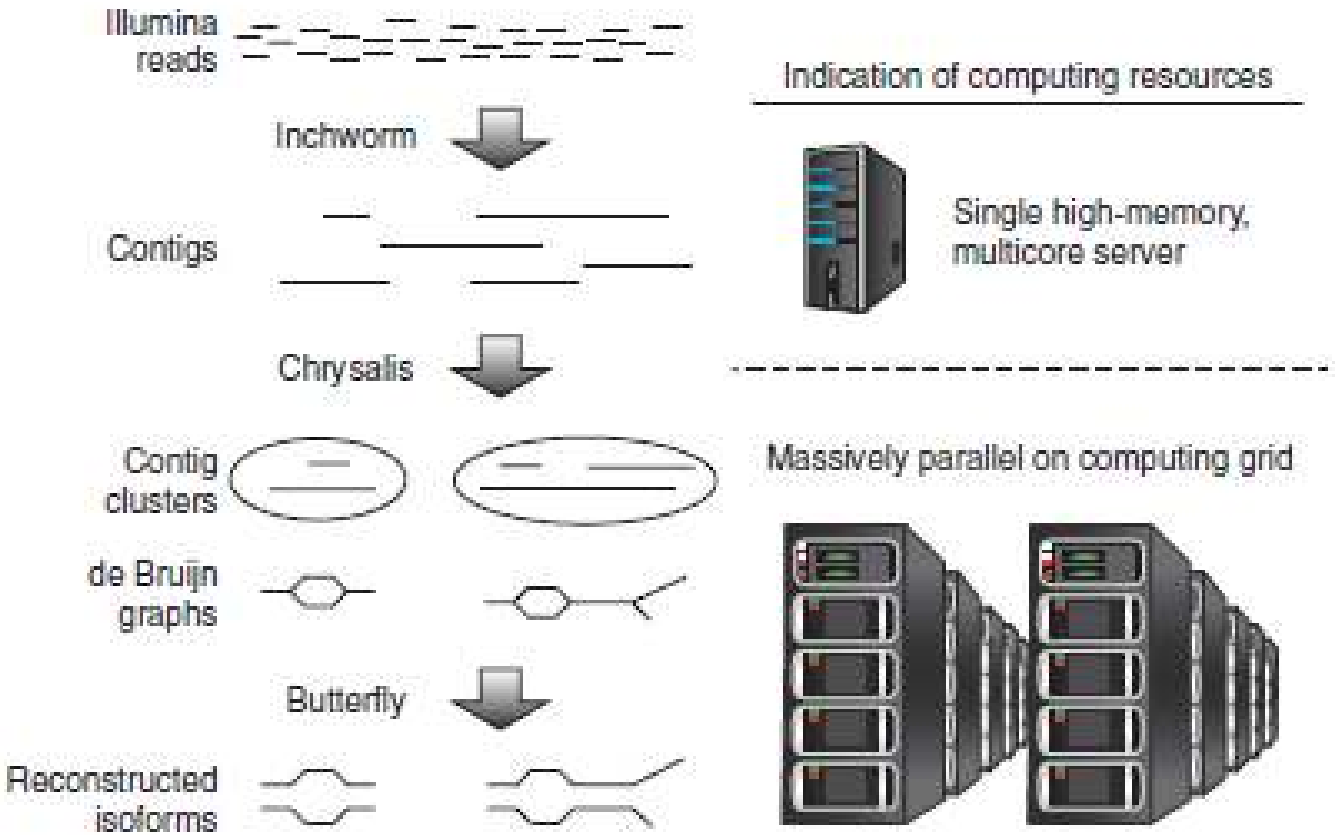
- Имеет модуль сборки на основе **референсного генома.**
- В пакете Trinity присутствует модуль для **нормализации ридов in silico** (normalize_by_kmer_coverage.pl).
  - Алгоритм **связан с «core» алгоритмами** Trinity
  - **Снижает время** работы и использование **памяти**
  - **Улучшает сборку**, так как исключаются **к-меры/риды**, вероятнее всего содержащие **ошибки.**

http://trinityrnaseq.sourceforge.net/

Сборка транскриптомов с помощью Trinity

- Написан для работы с **ридами Illumina.**
- Не «**пересобирает**» сборку **геномного сборщика**.
- Требования по памяти – **1G** для **1 млн.100 п.н. ридов Illumina**
- Время работы – от **½ часа до 1 часа** на **1 млн. ридов**. Самый **время затратный** этап **Butterfly**.

Сборка транскриптомов с помощью Trinity



[Haas BJ et al. Nat Protoc. 2013 Aug;8(8):1494-512]

Сборка транскриптомов с помощью Trinity

- **Подготовка парных ридов(Если данные не получены непосредственно с Illumina).**
  - В случае использования парных ридов. Необходимо, чтобы пары были распредленны по двум файлам. Причем первый член пары должен иметь суффикс – “/1”, второй – “/2”. Пример:

(запись в файле left.fq)

@61DFRAAXX100204:1:100:10494/1

CCGTTCACATGGAACCTTTCCCCTCTTCGGCCTTCAAAGTTCTCATTTGAATATTTGCTACTACCACCAAGATCTGCACCGACGGCCGCTCCGCCCGGGCTCGCGCCCT

+

CCCDFFFFHGHHHJJJJJJJJJJIIJJJJJIJGIGIJJIFHHIJIJIJJHJJJIJJIIJJJJIJJJJJJIJJHHHEHHFFFDDDDDDDBDDDDDDD5<@35<59909>5

(запись в файле right.fq)

@61DFRAAXX100204:1:100:10494/2

CTTTGAGGTGAAGGCAACTGCTGGTGACACCCATCTTGGTGGGGAAGATTTTGACAACAGAATGGTTAACCACTTTGTCCAAGAGTTCAAGAGGAAGAGTAAGAAGGAT

+

?B@FDDDFCF?FHIJJJIIJIJJJFHBAHHJG@DBFGFG0BFHIHIIGIJJJIIJIJGCHGGCHH)?B>CFEDDD@C>CC;ACAD:AAC>CCAD@?BBB3:+::ACDD:

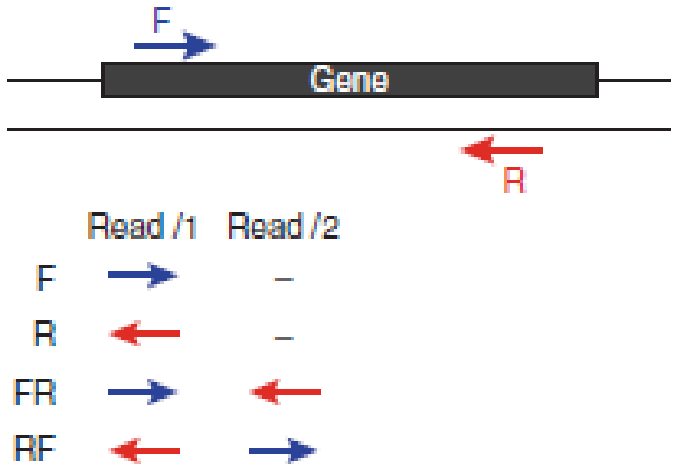
Сборка транскриптомов с помощью Trinity

- Как запустить Trinity:
  - Парные чтения в файлах - reads_left.fq, reads_right.fq:

    /path/to/Trinity.pl --seqtype fq --JM 4G --left reads_1.fq --right reads_2.fq
  - **--seqtype** – формат входного файла(fq – FASTQ; fa - FASTA).
  - **--left и --right** – имена файлов с ридами для спаренных ридов.
  - **--JM** – максимальный объем памяти для алгоритма Jellyfish.
- Другие полезные параметры:
  - **--CPU** – указывает Trinity, что можно использовать многопоточность и задает число используемых ядер. В основном затрагивает алгоритм Butterfly.
  - **--output** – директория для сохранения выходных файлов. По умолчанию trinity_out_dir.
  - **--full_cleanup** – полностью удалять все промежуточные файлы.

Сборка транскриптомов с помощью Trinity

– --SS_lib_type. Указывает Trinity, что используется нити-специфичная библиотека.

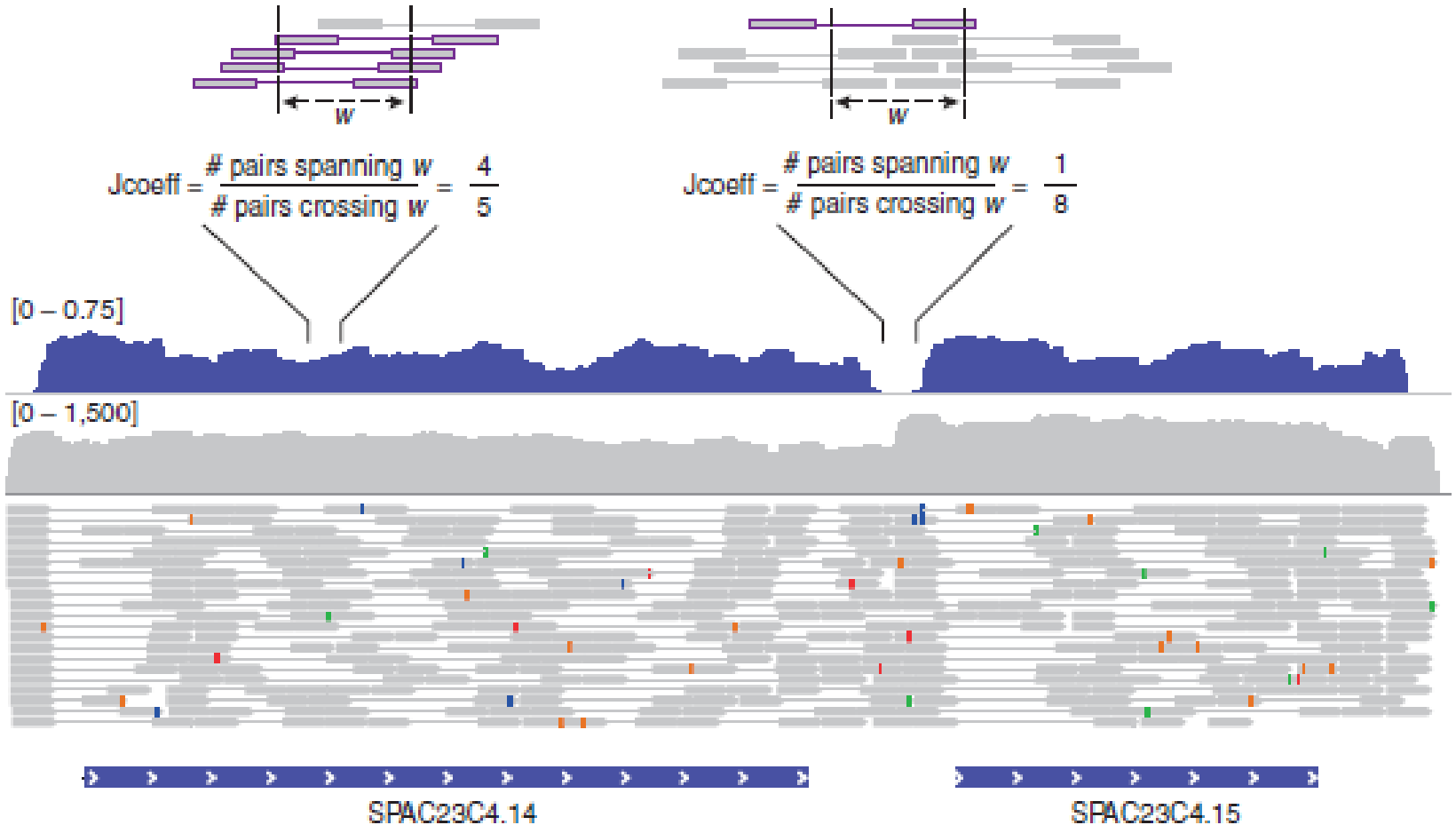


[Haas BJ et al. Nat Protoc. 2013 Aug;8(8):1494-512]

– --min_kmer_cov=2. Указывает, что необходимо убрать из рассмотрения кмеры с покрытием меньше 2.

Сборка транскриптомов с помощью Trinity

– --jaccard_clip



[Haas BJ et al. Nat Protoc. 2013 Aug;8(8):1494-512]

Сборка транскриптомов с помощью Trinity

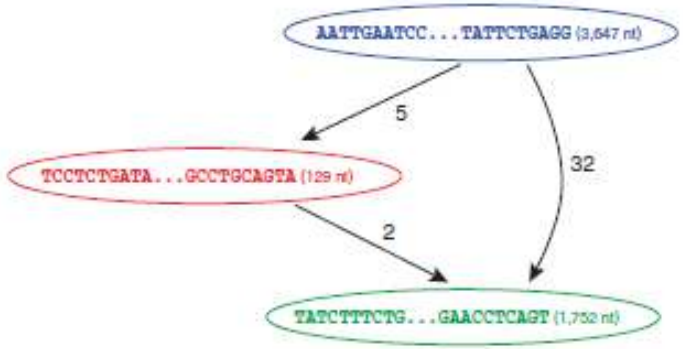
- Результаты сборки.
  - Результат сборки(**Trinity.fa**) лежит в директории заданной ключом **–output** или если ключ не задан в директории **trinity_out_dir**.
  - Если ключ **--full_cleanup** не задан то папка с результатами будет содержать результаты выполнения всех стадий алгоритма. При большом объеме входных данных она может занимать очень много места(для 30Гб входных данных - ~1Тб).
  - Как понять, что значат названия последовательностей в Trinity.fa?

>comp0_c0_seq1 len=5528 [3647:0-3646 129:3647-3775 1752:3776-5527]

……

>comp0_c0_seq2 len=5399 [3647:0-3646 1752:3547-5398]

…..



[Haas BJ et al. Nat Protoc. 2013 Aug;8(8):1494-512]

Сборка транскриптомов с помощью Trinity

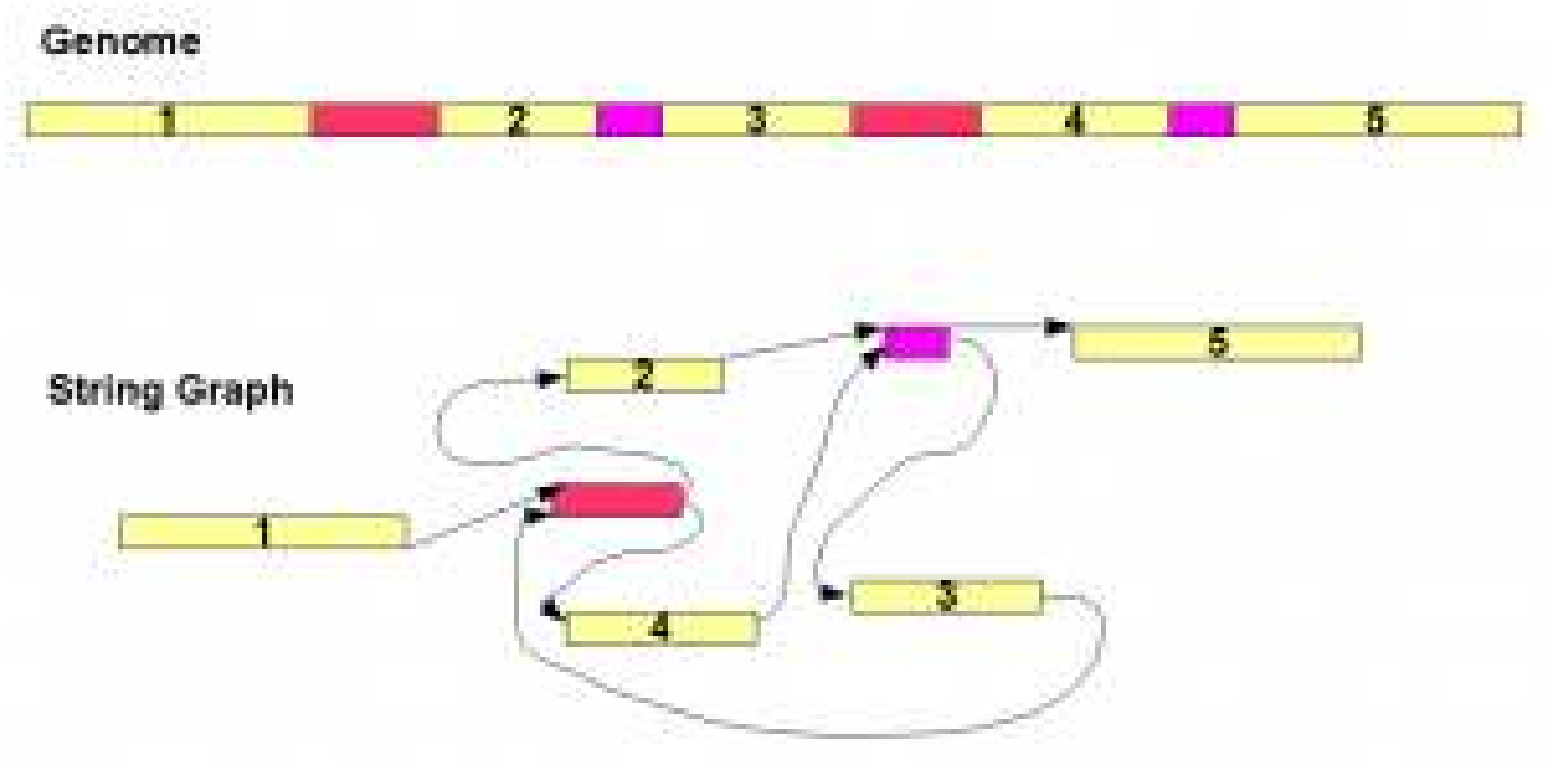
- Для **оценки качества и постпроцессинг сборки** в состав Trinity включено несколько инструментов:
  - **TrinityStats.pl** – вычисление N50 и количества контигов.
  - **alignReads.pl** и **SAM_nameSorted_to_uniq_count_stats.pl** – выравнивание ридов на сборку и оценка выравниваний.
  - **TransDecoder** – инструмент для предсказания ORF в результатах сборки.
  - **Trinotate** – инструмент для проведения функциональной аннотации

Сборка транскриптомов с помощью Newbler

- Разработан фирмой **Roche**.
- Разрабатывался для ридов технологии **454**. Так же можно использовать с данными Illumina.
- **String graph сборщик.**
- Работает с данными форматов **FASTA и SFF**.

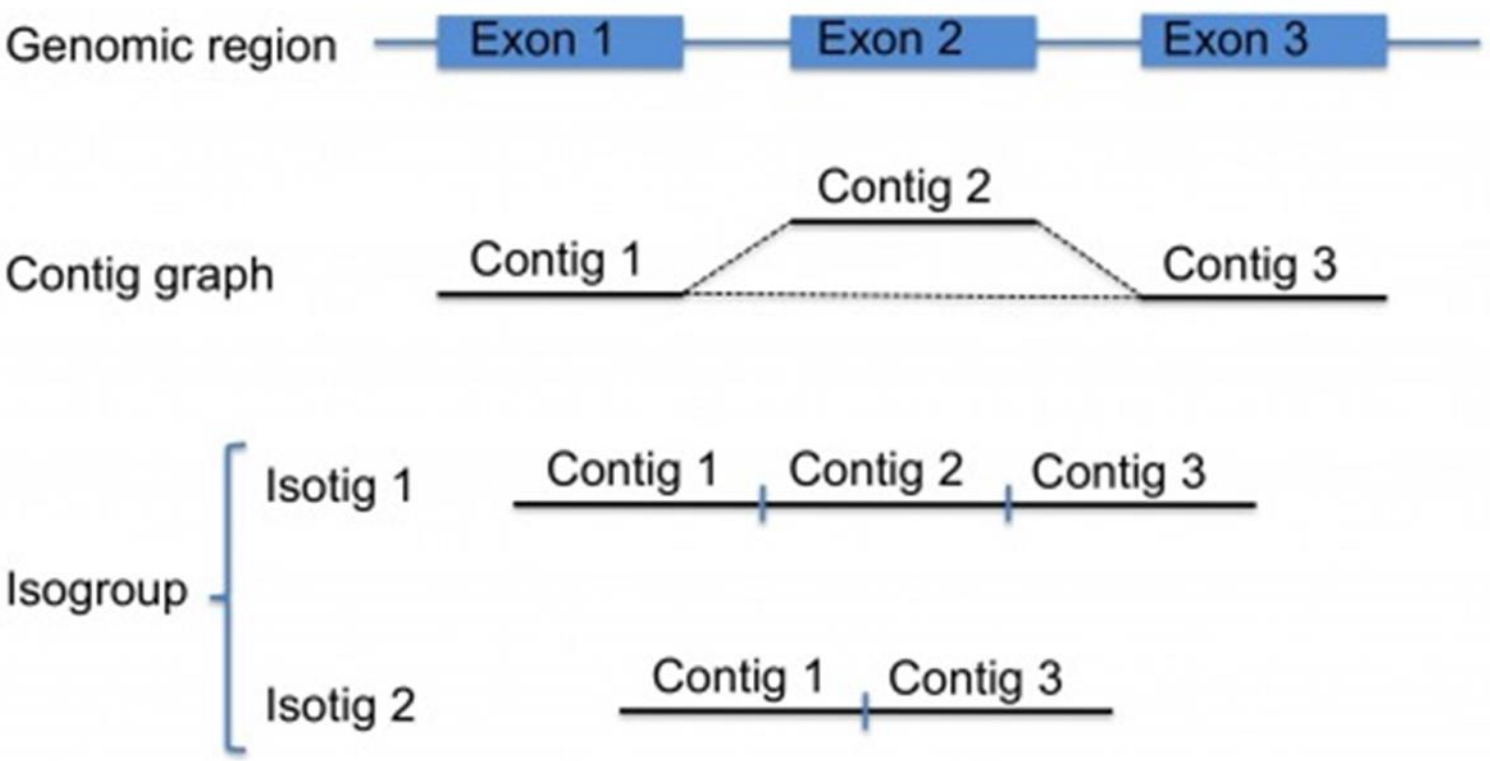
http://www.454.com/products/analysis-software/

Сборка транскриптомов с помощью Newbler



[http://contig.wordpress.com/2010/08/31/running-newbler-de-novo-transcriptome-assembly-i/]

Сборка транскриптомов с помощью Newbler



[http://contig.wordpress.com/2010/08/31/running-newbler-de-novo-transcriptome-assembly-i/]

Сборка транскриптомов с помощью Newbler

- Как запустить Newbler?

```
/path/to/newAssembly project1
cd project1
/path/to/addRun -lib libname -p reads_left.fasta
/path/to/addRun -lib libname -p reads_right.fasta
runProject –cdna
```

С помощью команды **newAssembly** создается проект сборки. Далее необходимо перейти в папку с новым созданным проектом. С помощью команды **addKun** добавляются файлы с ридами.(**-lib** – задает имя библиотеки, **-p** указывает, что библиотека с парными чтениями). **runProject** запускает сборку(-cdna указывает, что сборка производится RNA-seq данных).

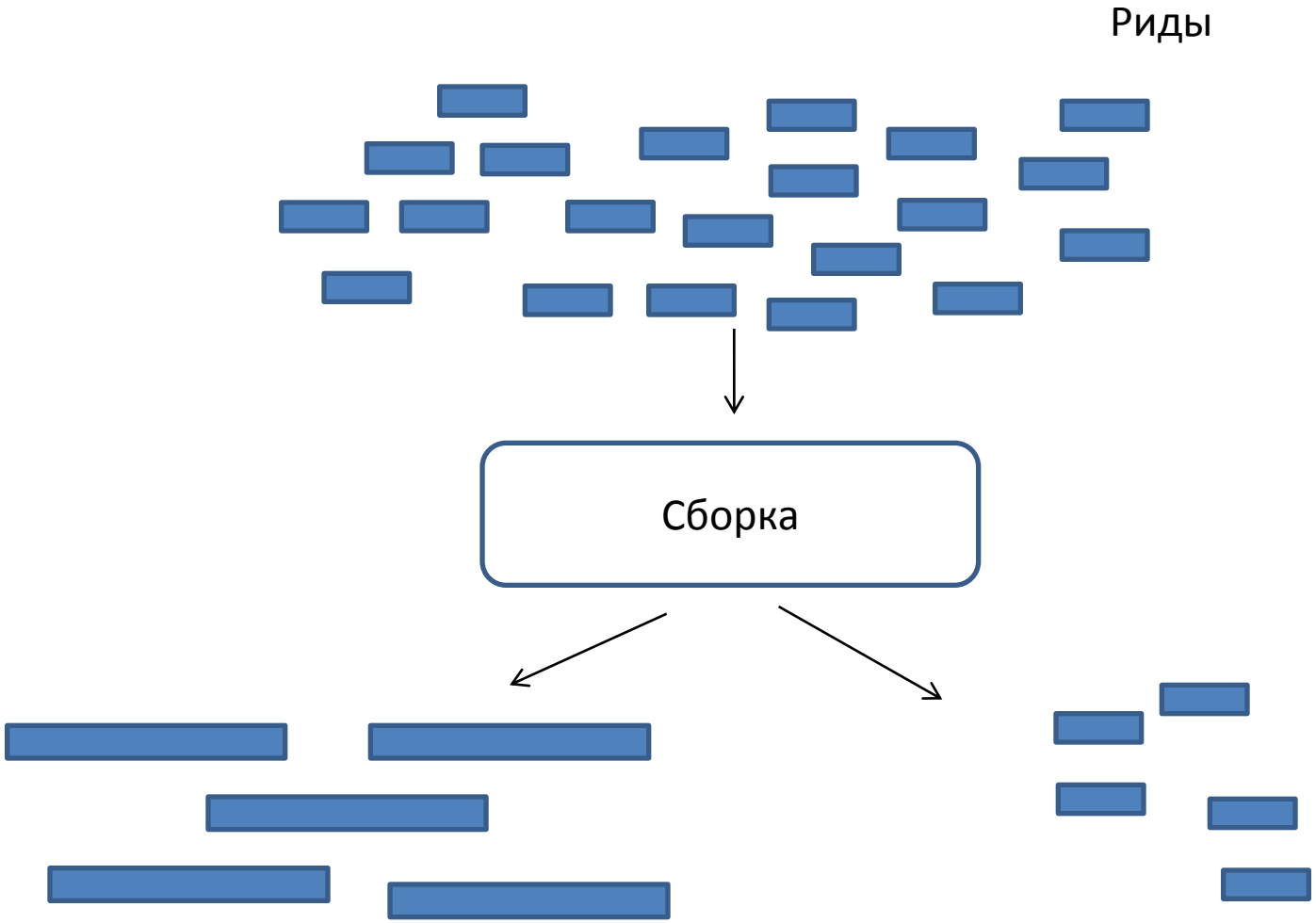
Сборка транскриптомов с помощью Newbler

- Для **runProject** можно задать еще ряд дополнительных параметров:
  - **-cpu num** – число вычислительных ядер, которое может использовать сборщик.
  - **-m** – позволяет сохранить информацию о ридах в памяти, что ускорит сборку, но значительно увеличит требование к памяти.

Сборка транскриптомов с помощью Newbler

- После выполнения сборки в попдпапке **Assembly** создаются следующие файлы:
  - **454AllContigs.fna** - fasta файл, содержащий все контиги размером больше 100 п.н.
  - **454LargeContigs.fna** - fasta файл, содержащий контиги >500п.н.
  - **454NewblerMetrics.txt** – статистические данные о результатах сборки.
  - **454Isotigs.fna** - fasta файл с изотигами.

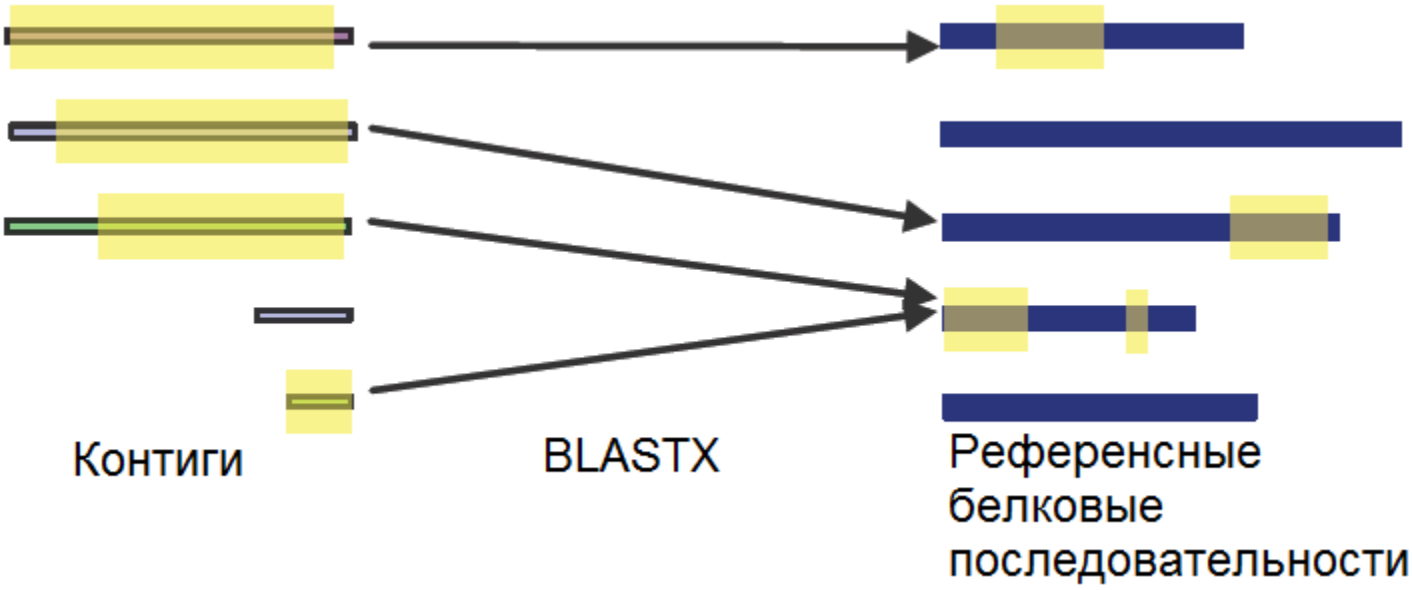
Результаты сборки



Оценка качества сборки

- **Картирование чтений** обратно на сборку.
  - **>60%** ридов картируется - **норма**.
  - **>80%** ридов картируется - **очень хороший результат**.
- Оценка **числа контигов**.
  - Должно быть в пределах размера транскриптома.(A.Thaliana ~25 000 генов, H.sapiens ~30 000 генов)
- Оценка **среднего покрытия контигов**.
- Оценка **числа уникальных п.н..**
  - чем больше тем лучше.
- Оценка **N50 контигов**. Должно соответствовать N50 транскриптома. (A. Thaliana ~1900 п.н., H.sapiens ~2500 п.н.)
- Оценка **числа контигов >1 т.п.н.**.
  - Чем больше тем лучше.

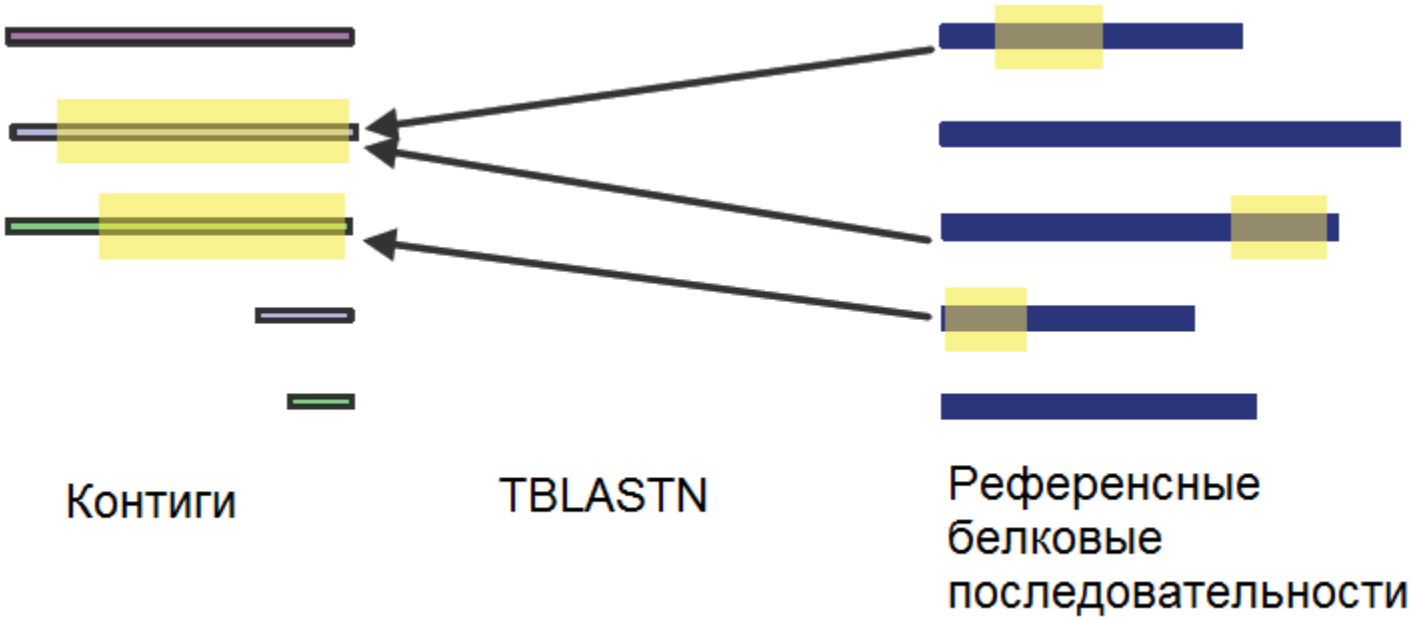
Оценка качества сборки



- **Blast** на **белковую/транскриптомную базу** близкого организма.
  - Поможет понять какая часть транскриптома была собрана.
- **Число** найденных **совпадений** в **белковой/транскриптомной базе**.

[https://www.abrf.org/Committees/Education/Activities/ABRF2013_SW1_oneil_DeNovo-transccript-Assembly.pdf]

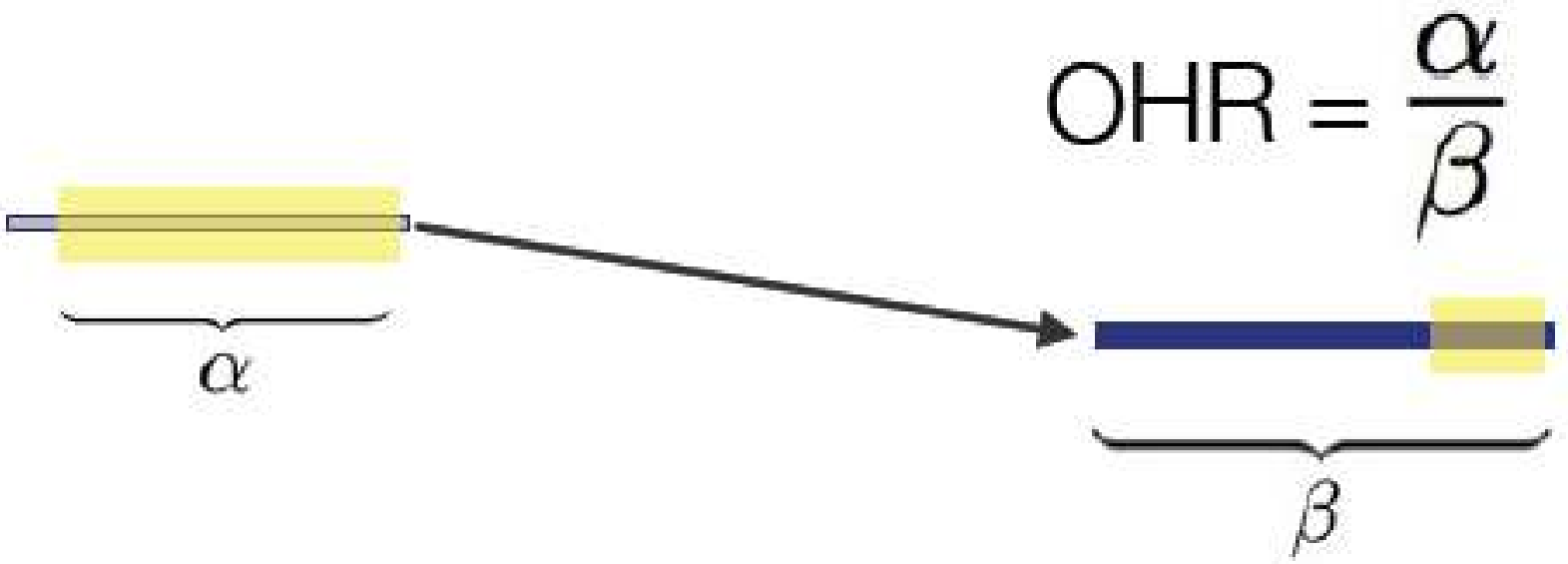
Оценка качества сборки



- **Обратная аннотация**: Картирование референсных белков на контиги.
- **Число** найденных **совпадений** при **обратной аннотации**.

[https://www.abrf.org/Committees/Education/Activities/ABRF2013_SW1_oneil_DeNovo-transccript-Assembly.pdf]

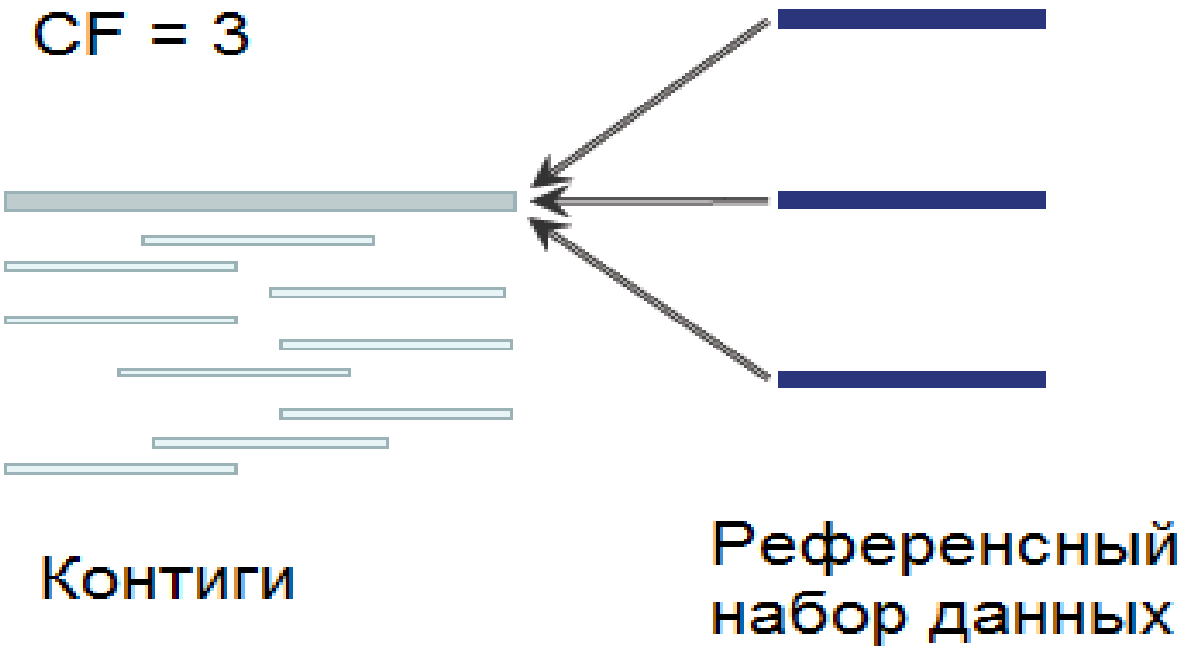
Оценка качества сборки



Ortholog hit ratio(OHR) – мера **полноты** сборки транскрипта.

[https://www.abrf.org/Committees/Education/Activities/ABRF2013_SW1_oneil_DeNovo-transccript-Assembly.pdf]

Оценка качества сборки



Collapse Factor(CF) - мера **пересобранности транскриптов**.

[https://www.abrf.org/Committees/Education/Activities/ABRF2013_SW1_oneil_DeNovo-transccript-Assembly.pdf]

Оценка качества сборки

- Из метрик используемых при **оценке геномных сборок** наиболее надежно отражают **качество транскриптомной сборки**:
  - **% используемых в сборке ридов.**
  - **Число контигов > 1 т.п.н.**.
  - **Число уникальных п.н.**.
- Для использования таких метрик как **N50**, **число контигов** необходимо иметь **оценки размера исследуемого транскриптома**.
- **Метрики основанные на аннотации** имеется возможность применять только в случае наличия **достаточно полной белковой базы**, либо **исследуемого организма**, либо **близкородственного** ему.
- **%картирующихся на сборку ридов,N50, среднее покрытие, среднее OHR** необходимо использовать на всем наборе результатов сборщика(**контиги + синглетоны**).
- **Средний CF, число совпадений с белковой базой при прямой и обратной аннотации** только на **контигах**.

Постобработка транскриптомной сборки

«У меня слишком много контигов, что делать дальше?»

- Многие **транскриптомные сборщики**(в частности Trinity) дают **большое количество контигов**(>100K)
- Возникает вопрос, как уменьшить число контигов.
- Можно сформулировать следующее правило:

Не нужно кластеризовать, нужно фильтровать.

- **Кластеризация** приведет к **схлопыванию паралогов**, **альтернативных изоформ** и **семейств генов**.
- **Кластеризация** приводит к возникновению **химер**.
- **Фильтруйте** по **% изоформ**, **покрытию**, **ORF**, **бластовым хитам** и т. д. Оценивать насколько **велики потери данных** в **результате фильтрации** можно **картируя риды на контиги**.

Сложности транскриптомной сборки

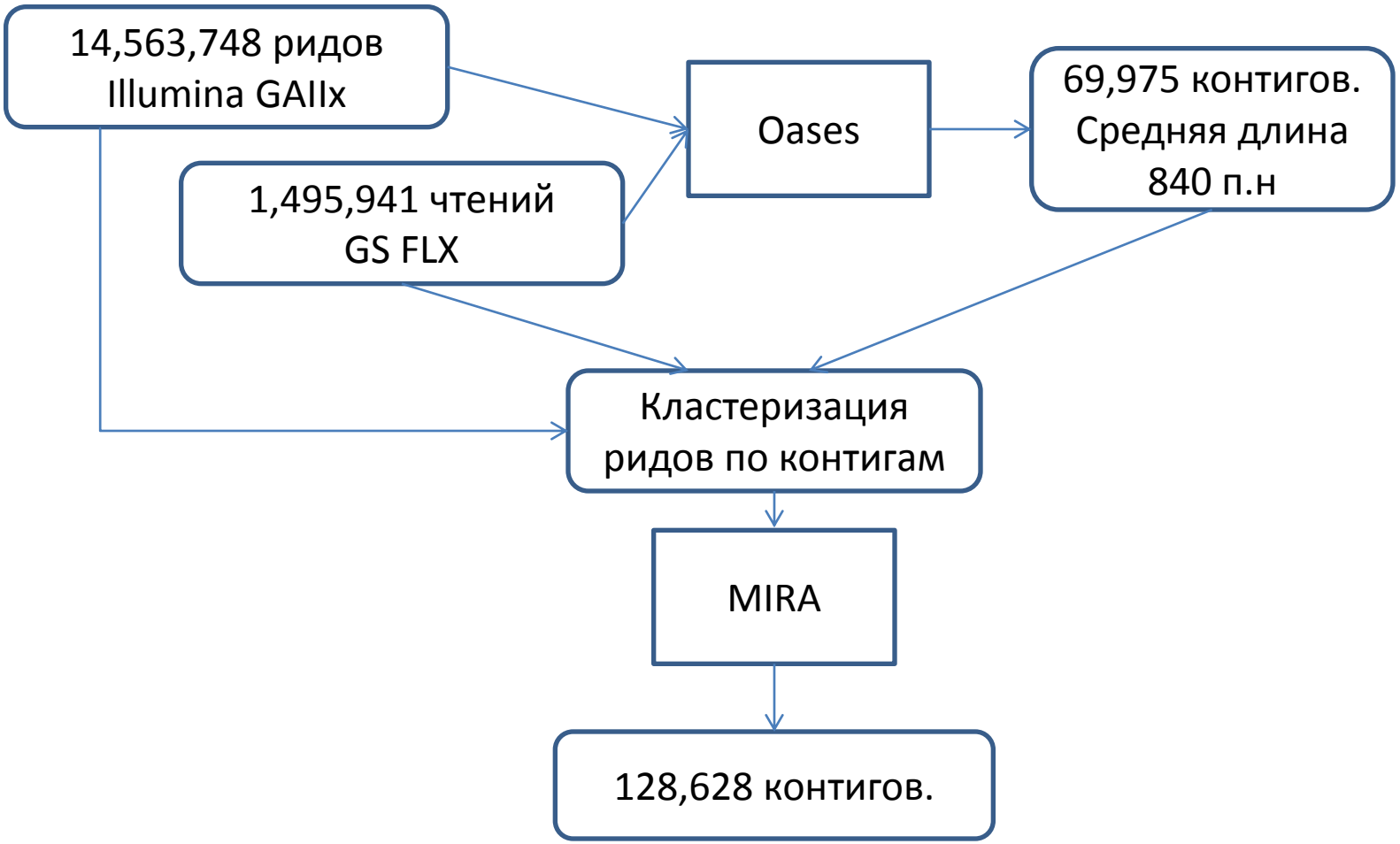
- **Загрязнение** в исходном образце
  - Приводит к очень **фрагментированным сборкам**. **Число контигов** значительно **больше ожидаемого** и **N50 мало**.
  - Можно пытаться **фильтровать риды** путем сравнения их с **транскриптомами/белками возможного загрязнителя**. Процедура очень долгая и мучительная, требующая **много вычислительных ресурсов**. **Не гарантирует результата**.

Сложности транскриптомной сборки

- **Паралоги**
  - В случае большого числа **паралогов** в целевом транскриптоме,например если сборка производилось для **транскриптома полиплоида**, будет наблюдаться **две ситуации** при разных наборах параметров сборщика/сборщиков(либо **число контигов** значительно меньше ожидаемого при более-менее **достигнутом целевом N50** , либо **огромное число контигов** при **малом N50**).
  - Пример. Статья [Schreiber et al. BMC Genomics 2012, 13:492]. Собирался транскриптом T. aestivum.

Сложности транскриптомной сборки

Число генов у риса ~41000



Сложности транскриптомной сборки

- Альтернативные изоформы.
- Неравномерность покрытия.
- Повторяющиеся последовательности.
- Химерные контиги.

Софт для финализации сборки

- STM – скаффолдинг по белкам.

[http://www.surgetgroba.ch/downloads/stm.tar.gz]

Объединение сборок(Assembly Reconcilation)

- MIX[https://github.com/cbib/MIX]

Оценка качества сборок.

- ALE[http://sc932.github.io/ALE/]
- Transrate[http://hibberdlab.com/transrate/]

Вопросы?