Оценка качества транскриптомных сборок

Елена Бушманова

Руководитель: Андрей Пржибельский
Лаборатория алгоритмической биологии
СПбАУ РАН

Проект по *de novo* сборке данных РНК

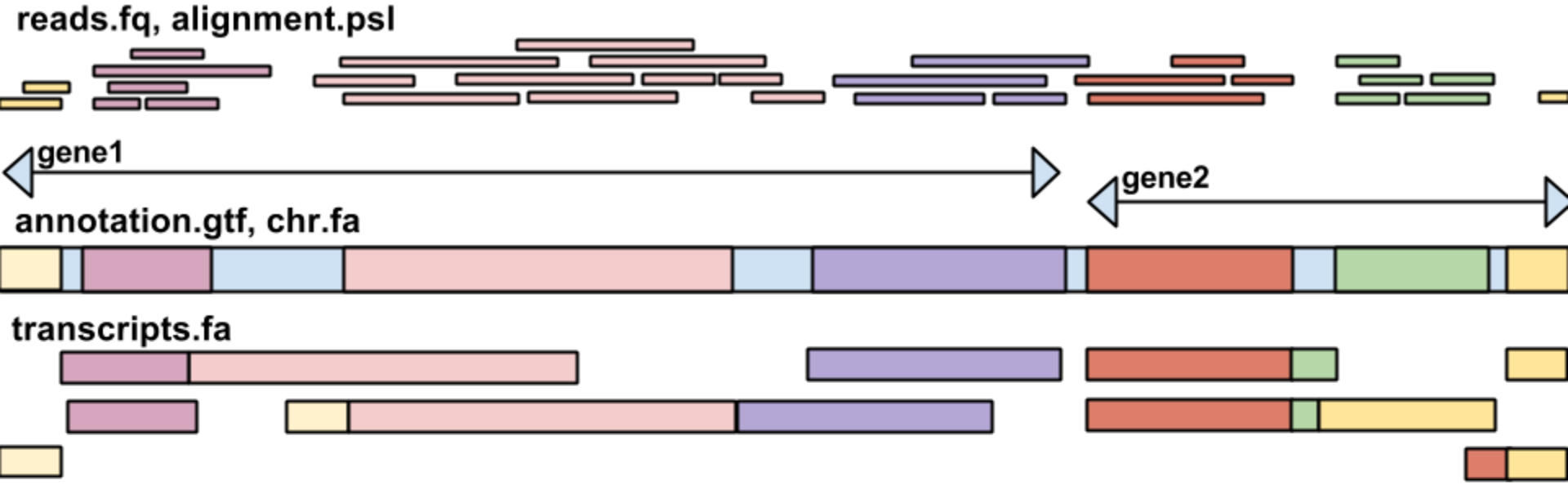
- Создание РНК-ассемблера на базе SPAdes
- Поиск fusion генов
- **Разработка Trans-QUAST**
- Поддержка данных масс-спектрометрии (утилита Enosi)

Trans-QUAST

- **Использование метрик и методов из существующих статей (Oases, Trinity, Scripture, Cufflinks)**
- **Получение статистик, характеризующих качество РНК-сборки**
- Учет специфики проекта (поиск fusion генов)

Метрики

- Частота ошибок (делеций, вставок, замен)
- Количество непокрытых транскриптами аннотированных генов и экзонов
- Процент полностью и частично восстановленных генов и их экзонов
- Суммарная длина генов и транскрибируемых экзонов, процент восстановленных оснований
- Процент не аннотированных генов
- Инверсии, релокации и транслокации



Coverage of annotated exons:

count of annotated exons = 7

count of covered annotated exons = 6

count of complete covered annotated exons = 5

Coverage of annotated genes:

count of annotated genes = 2

percent of covered annotated genes = 100%

count of complete covered annotated genes = 1

count of uncovered annotated genes = 0

Gene1:

count of covered exons = 4

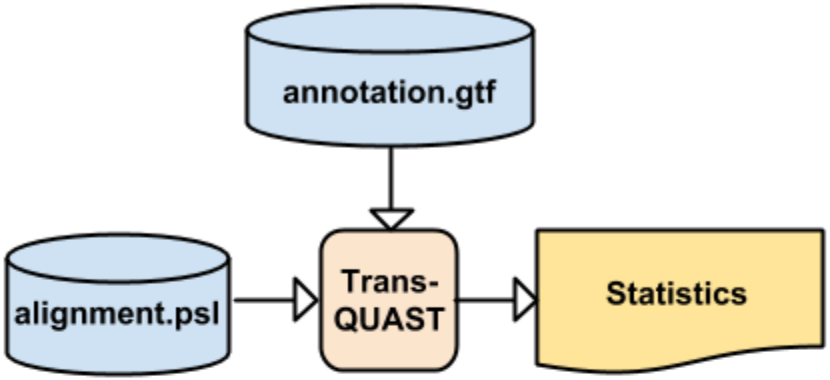
count of complete covered exons = 4

Gene2:

count of covered exons = 2

count of complete covered exons = 1

Pipeline 1




```
Trans-Quast.py --filePSL alignment.psl --fileGTF annotation.gtf
```

GTF:

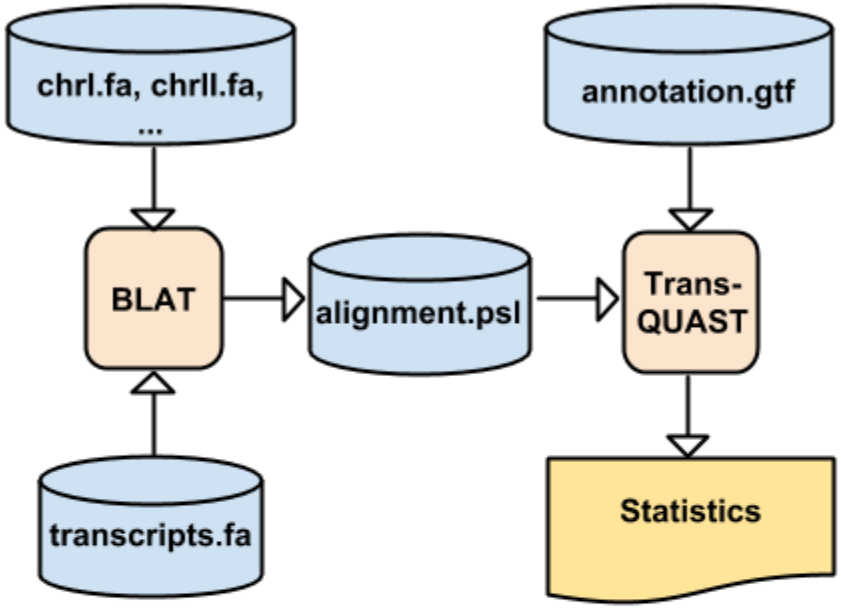
```
chrI    temp2.txt   stop_codon  1807    1809    .   -   0   gene_id "YAL068C"; transcript_id "YAL068C"; exon_number "1"; exon_id "YAL068C.1"; gene_name "Y
chrI    temp2.txt   exon    2480    2707    .   +   .   gene_id "YAL067W-A"; transcript_id "YAL067W-A"; exon_number "1"; exon_id "YAL067W-A.1"; gene_name
chrI    temp2.txt   CDS 2480    2704    .   +   0   gene_id "YAL067W-A"; transcript_id "YAL067W-A"; exon_number "1"; exon_id "YAL067W-A.1"; gene_name "YAL
chrI    temp2.txt   start_codon 2480    2482    .   +   0   gene_id "YAL067W-A"; transcript_id "YAL067W-A"; exon_number "1"; exon_id "YAL067W-A.1"; gene_n
chrI    temp2.txt   stop_codon  2705    2707    .   +   0   gene_id "YAL067W-A"; transcript_id "YAL067W-A"; exon_number "1"; exon_id "YAL067W-A.1"; gene_n
chrI    temp2.txt   exon    7235    9016    .   -   .   gene_id "YAL067C"; transcript_id "YAL067C"; exon_number "1"; exon_id "YAL067C.1"; gene_name "YAL0
chrI    temp2.txt   CDS 7238    9016    .   -   0   gene_id "YAL067C"; transcript_id "YAL067C"; exon_number "1"; exon_id "YAL067C.1"; gene_name "YAL067C"
chrI    temp2.txt   start_codon 9014    9016    .   -   0   gene_id "YAL067C"; transcript_id "YAL067C"; exon_number "1"; exon_id "YAL067C.1"; gene_name "Y
```

PSL:

```
psLayout version 3

match   mis-    rep.    N's Q gap   Q gap   T gap   T gap   strand  Q           Q       Q       Q   T           T       T       T   block   blockSizes  qS
        match   match       count   bases   count   bases           name        size    start   end name        size    start   end count
---------------------------------------------------------------------------------------------------------------------------------------------------------
1978    0   0   0   0   0   0   0   +   c0_g1_i1    1995    0   1978    tpg|BK006943.2| 745751  243280  245258  1   1978,   0,  243280,
1297    0   0   0   0   0   0   0   -   c0_g2_i1    1297    0   1297    tpg|BK006943.2| 745751  242003  243300  1   1297,   0,  242003,
1580    0   0   0   0   0   0   0   -   c1_g1_i1    1580    0   1580    tpg|BK006948.2| 1091291 470099  471679  1   1580,   0,  470099,
772 0   0   0   0   0   0   -   c1_g2_i1    788 0   772 tpg|BK006948.2| 1091291 469342  470114  1   772,    16, 469342,
```

Pipeline 2

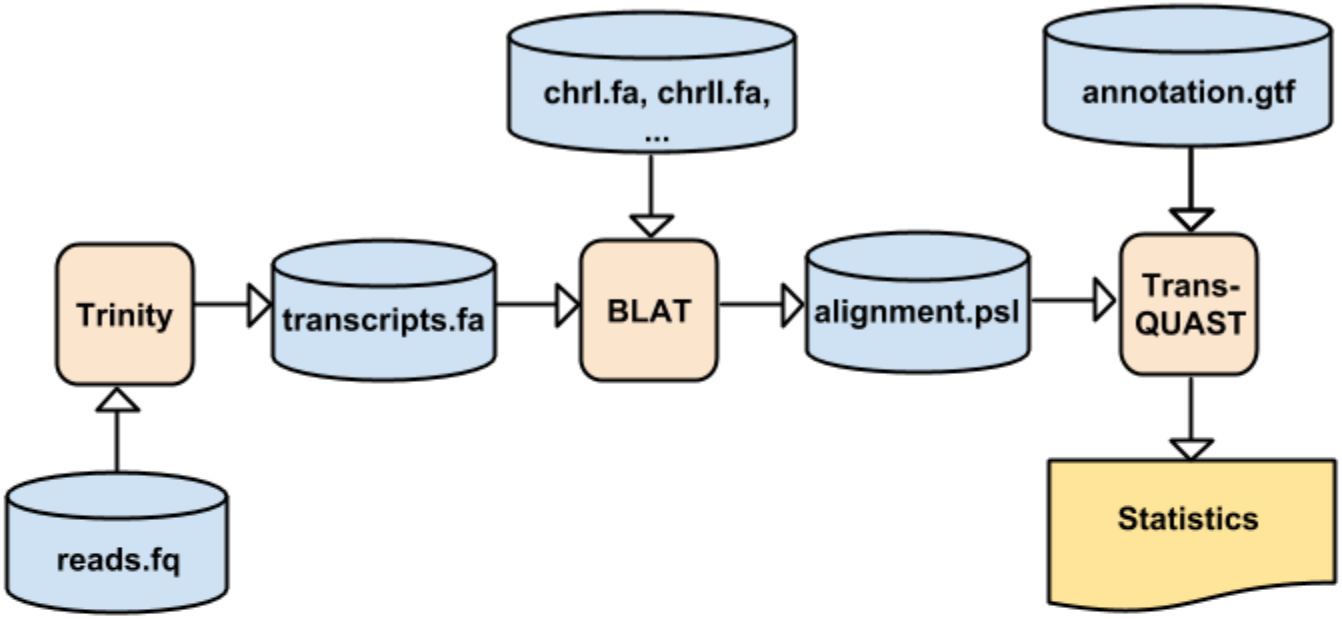



```
Trans-Quast.py --database DATABASE --rna transcripts.fa --fileGTF
annotation.gtf
```

FASTA:

```
TTCCGCCGCATATGCATTGTGTAGATCCAAAAGTAAGGACAAGATATCATGGGATGAAGA
AGAACAGGCGCGATTAATGGGCGTTGTAAAATTTAATTCAGAGCATTACAGGGACTAGAA
AAAAAAAAAAAAAAA
>c0_g2_i1 len=1297 path=[3945:0-1296]
TGTTTTTCCTTATCAGCTCAATGATAATGCCAACGCCATTATTTAACGAATTGCCTCCCT
TGAGCATTATATCCATAAGTTGTTTCATCATATTTGGAGAAACTAATTGCCTTGTTAATT
CATTAGGCCCTATGCTCGAAGTAATCTCATTAGGACAGTTACCGCTAATAGTAACCAGAG
CTTTCAGAAAATCACCGGCTGCAGATTGTGTGCAAGAATCAAAAACAGGATCGAGTAAAT
GAATCAACTTTGGAACCAAGTTCTGTTTTTTAAACAGTTGTATAACTCCATTTGATATTT
```

Pipeline 3

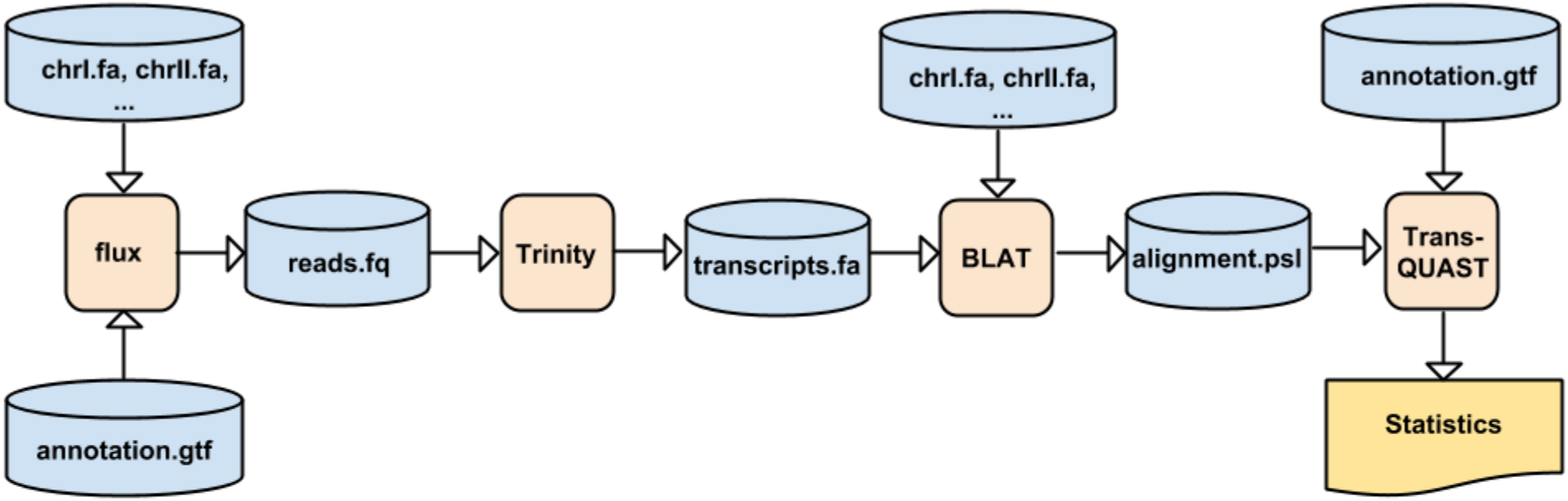



```
Trans-Quast.py --database DATABASE --reads reads.fq --fileGTF
annotation.gtf
```

FASTQ:

```
IIIIHIDDDFBDIIIIIIIIIHHIHIHIIIIIIGIIIIIH>EIIHIIGDE8HD>EGC<;<BHHHBBB@EHHHGB58<?A;CBDHHHIHHGHEHEHHD>AG
@chrI:335-792W:YAL069W:5:315:-19:158/1
CATATTGAAACGCTAACAAATGATCGTAAATAACACACACGTGCTTACCCTgCCACTTTAaACCACCgCaACgcngCAacgTgcgnncgcgaTcggTcgg
+
GIIEGEGBEDEGGGHHGGGEHIIIGCBBDEIIIHGIIIHHFDIIIDEB<=7#EEGGHG3-.A?A####################################
@chrI:335-792W:YAL069W:5:315:-19:158/2
ATGGAGaGAAGTGGAATCTGAGAGTAGGGTAAGTTTGAGATGGTATATACTGTAGCATCCGTGTGCGTACGTAAAATCAGTATACAAGTGAGGGTGAGTA
```

Pipeline 4




```
Trans-Quast.py --database DATABASE --fileGTF annotation.gtf
```

Планы

- Тестирование на различных сборщиках (Trinity, Spades, Trans-ABySS, Oases)
- Использование различных данных:
  - Симуляция (flux, dwgsim, BEERS)
  - Реальные геномы
- Получение статистик для fusion генов:
  - Химерические транскрипты
  - Реальные fusion гены

Спасибо за внимание!

Вопросы?