rnaQUAST: quality assessment tool for transcriptome assemblies

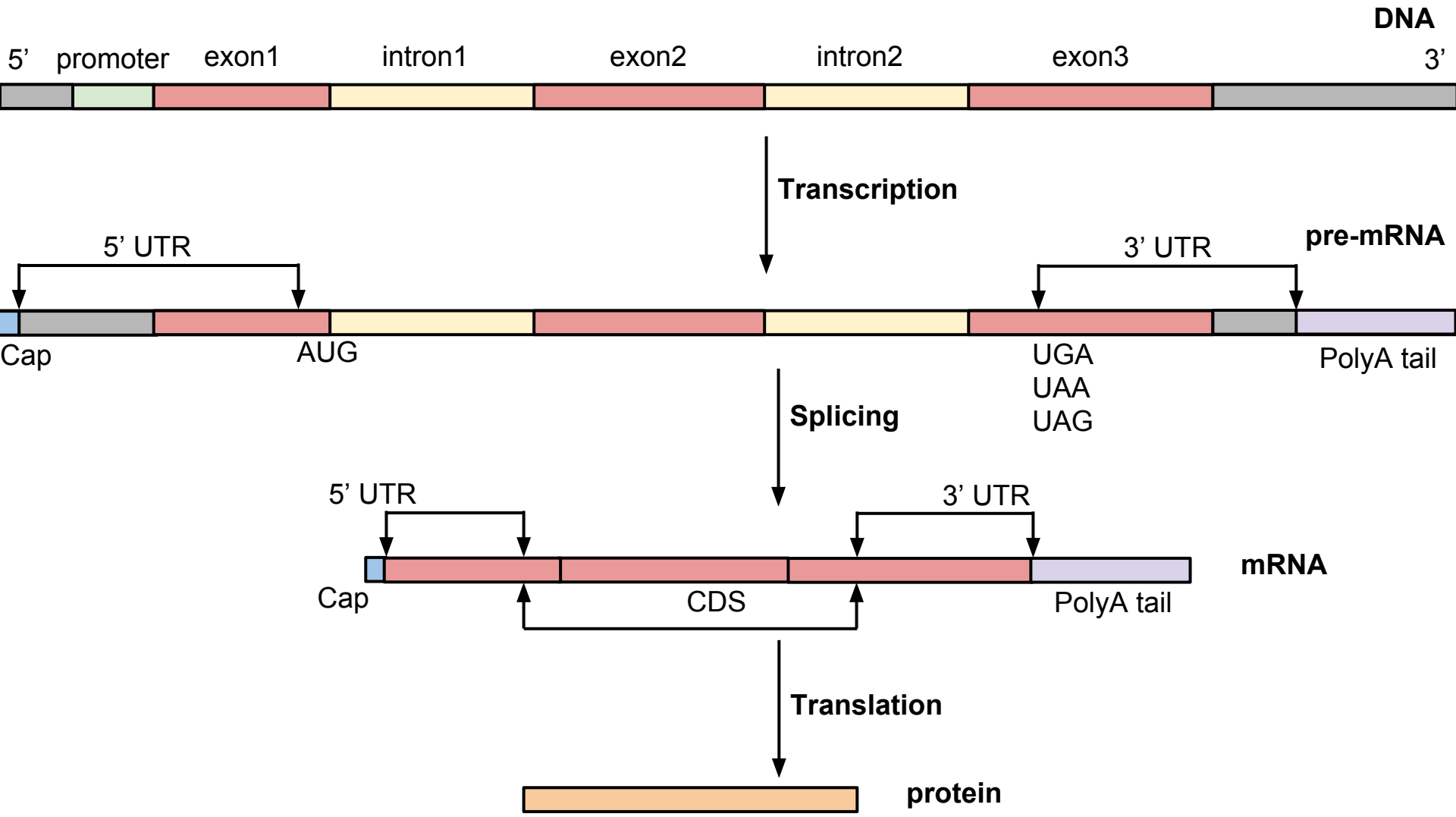
Elena Bushmanova

Scientific advisor: Andrey Prjibelski
Algorithmic Biology Lab, SPbAU RAS

Goal

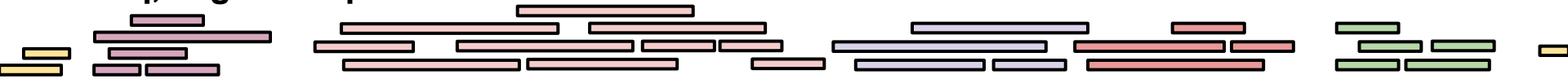
- A tool for analysing transcripts using
  - Reference genome
  - Genome annotation
- Various metrics
  - Annotation coverage
  - Transcripts completeness
  - Transcripts correctness
- Comparison of different assemblers

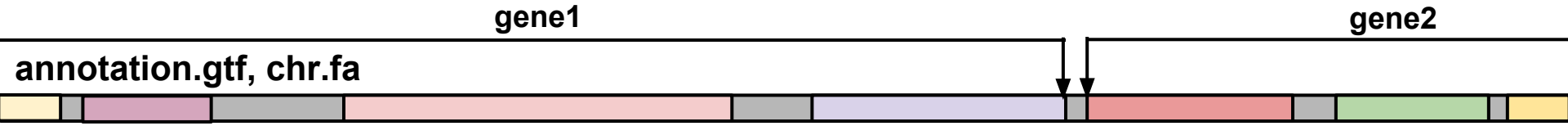
mRNA


DNA
5’
promoter
exon1
intron1
exon2
intron2
exon3
3’
Transcription
pre-mRNA
5’ UTR
3’ UTR
Cap
AUG
UGA
UAA
UAG
PolyA tail
Splicing
5’ UTR
3’ UTR
mRNA
Cap
CDS
PolyA tail
Translation
protein

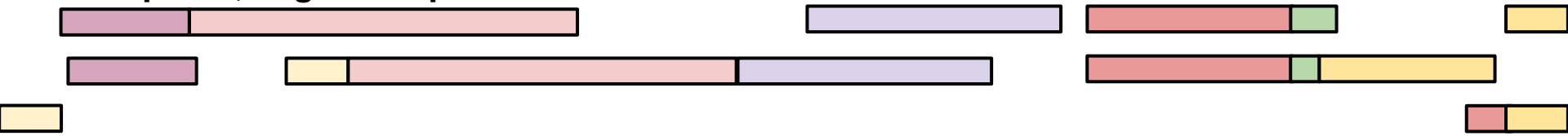
Data

reads.fq, alignment.psl

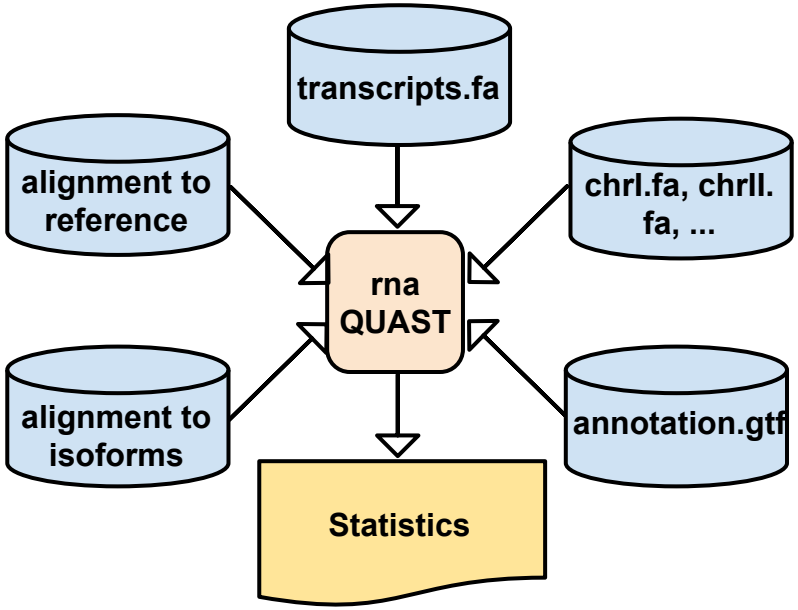




transcripts.fa, alignment.psl

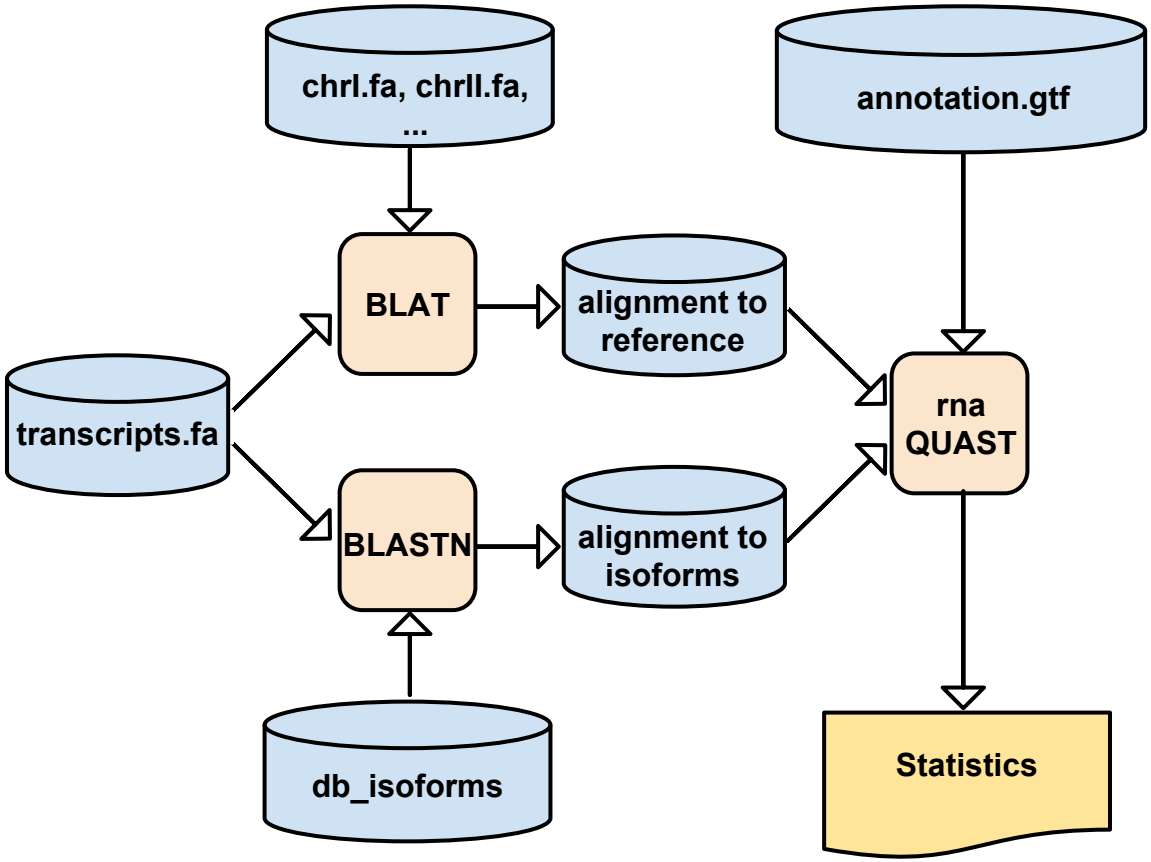


Pipeline 1



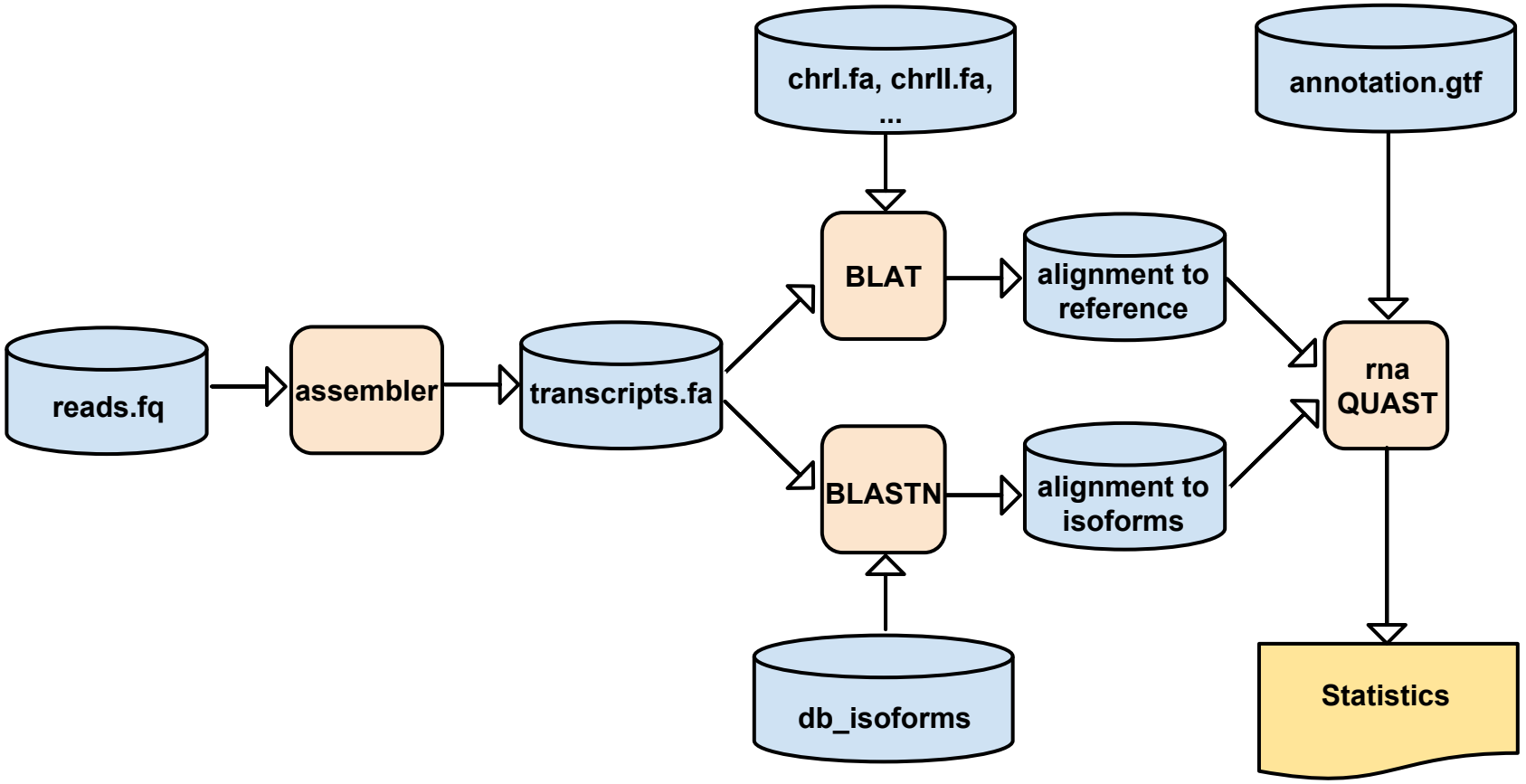

```
rnaQUAST.py -p1 --transcripts TRANSCRIPTS --database DATABASE --
  annotation ANNOTATION --alignments ALIGNMENTS --outdir OUTDIR
```

Pipeline 2



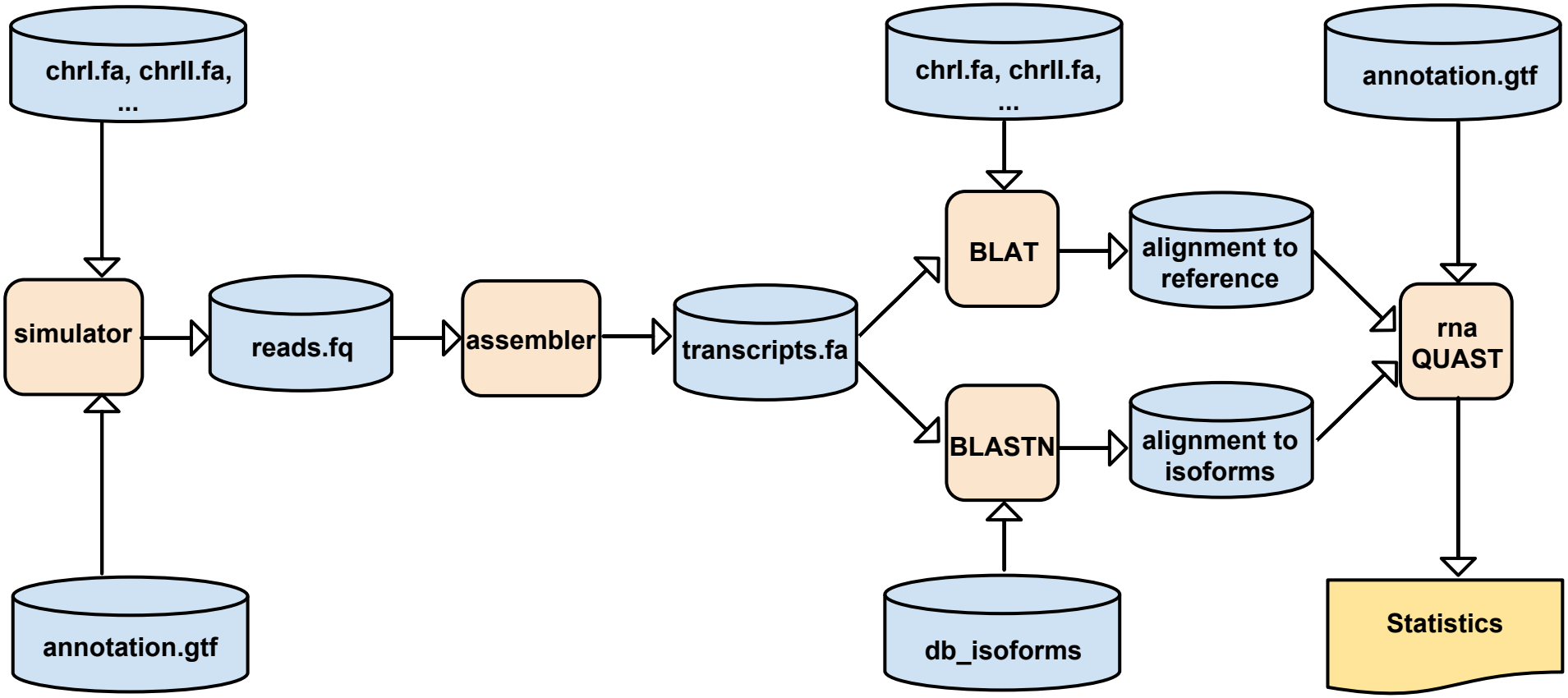

```
rnaQUAST.py -p2 --transcripts TRANSCRIPTS --database DATABASE --
annotation ANNOTATION --outdir OUTDIR
```

Pipeline 3




```
rnaQUAST.py -p3 --database DATABASE --annotation ANNOTATION --assembler
            ASSEMBLER --reads READS --outdir OUTDIR
```

Pipeline 4




```
rnaQUAST.py -p4 --database DATABASE --annotation ANNOTATION --simulator
            SIMULATOR --assembler ASSEMBLER --outdir OUTDIR
```

Already done

- **Tested BLAT on simulated transcripts**
- **Filtering short repeat alignments (greedy algorithm)**
- **Implemented pipelines**
- **Implemented various metrics**
  - Basic metrics (without alignment)
  - Simple metrics (with alignment, without annotation)
  - Mapped metrics (with alignment, with annotation)

Already done

- **Tried de novo assemblers**
  - Trinity
  - IDBA-Trans
  - Trans-ABySS
  - Oases
  - SOAPdenovo-Trans
  - SPAdes
- **Tried reference-based assemblers / aligners**
  - TopHat
  - Scripture
  - Cufflinks

Already done

- **Simulated fusion transcripts and new isoforms**
  - Isoform of two genes from one chromosome
  - Isoform of two genes from different chromosomes
  - Annotated + unannotated regions
  - New isoform of the same gene

Basic transcripts metrics

- Total number of transcripts
- Transcripts lengths distribution
- Number of transcripts with length inside isoform length range

Basic isoforms metrics

- Total number of isoforms
- Isoforms lengths distribution
- Total number of exons / introns
- Number of exons / introns per isoform distribution
- Exons / introns lengths distribution
- Estimated number of paralogous genes distribution

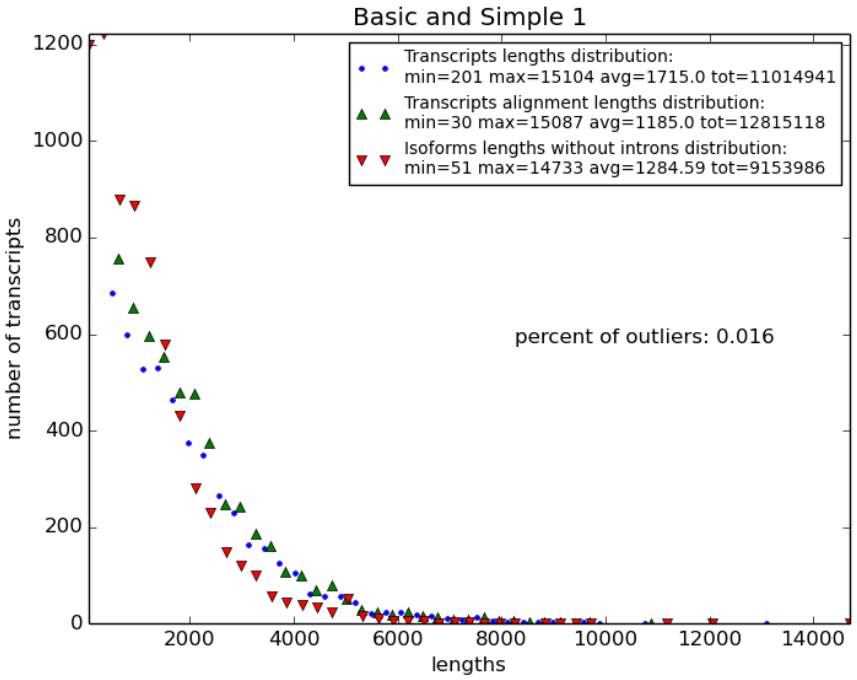
Simple transcripts alignment metrics

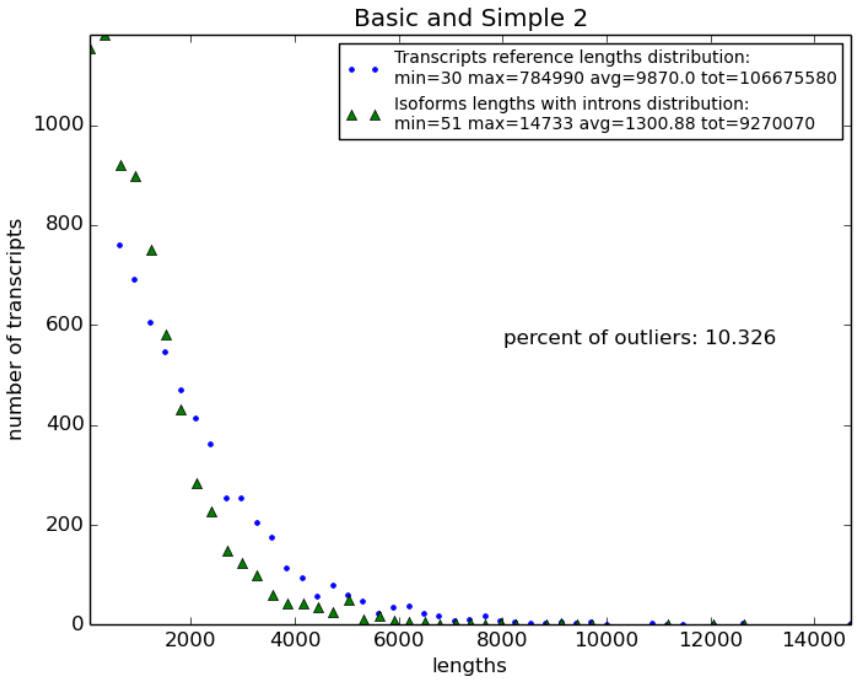
- Transcript alignment length/fraction distribution
- Total number of aligned blocks
- Number of blocks per transcript distribution
- Blocks lengths distribution

Simple transcripts alignment metrics

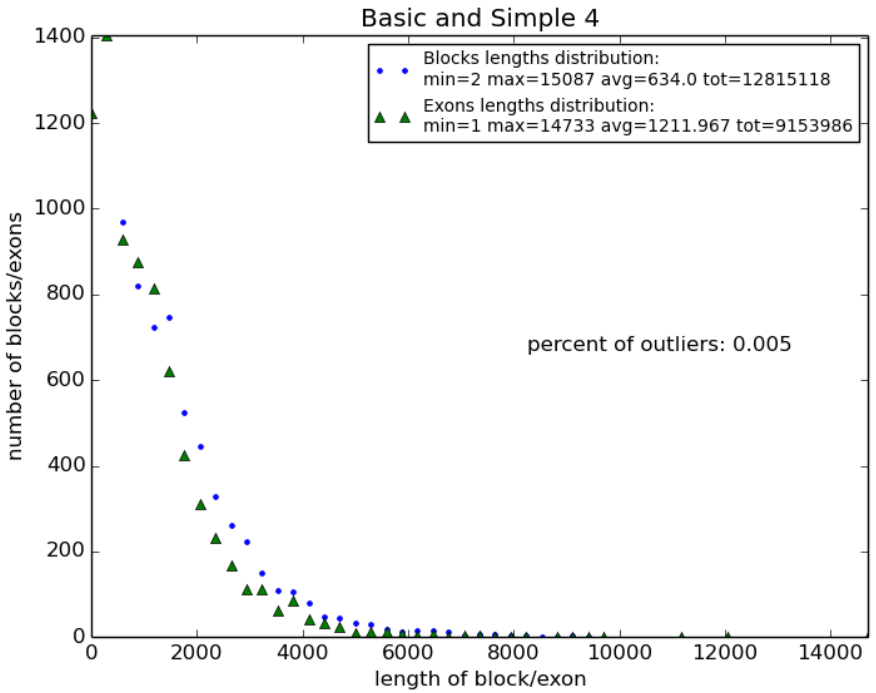
- Number of aligned/unaligned/multiple-aligned transcripts
- Multiple-aligned transcripts distribution
- Fraction of the genome mapped
- Mismatch content
- Target and query gap number/lengths distributions

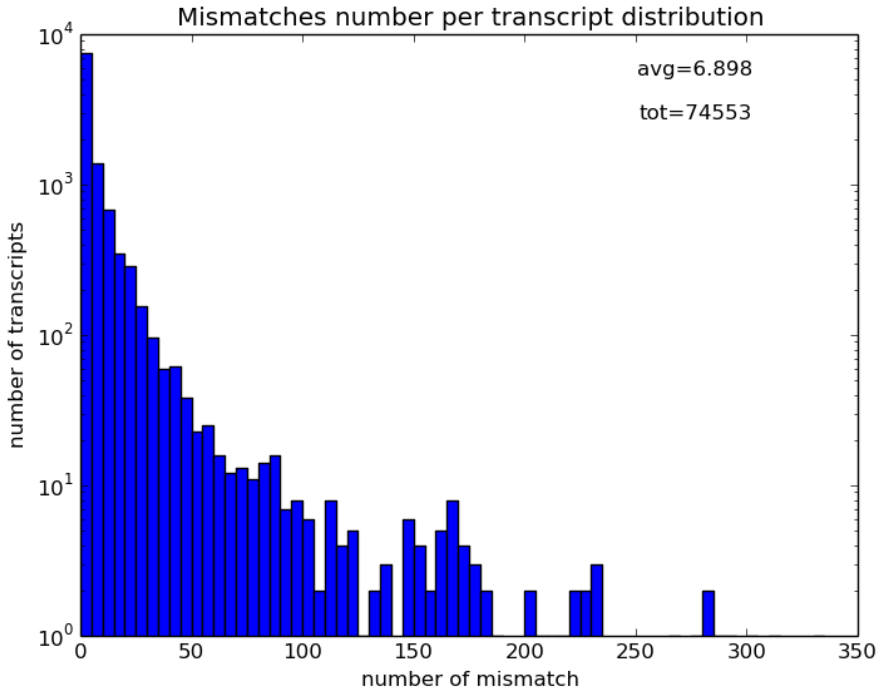
Basic and simple results 1





Basic and simple results 2

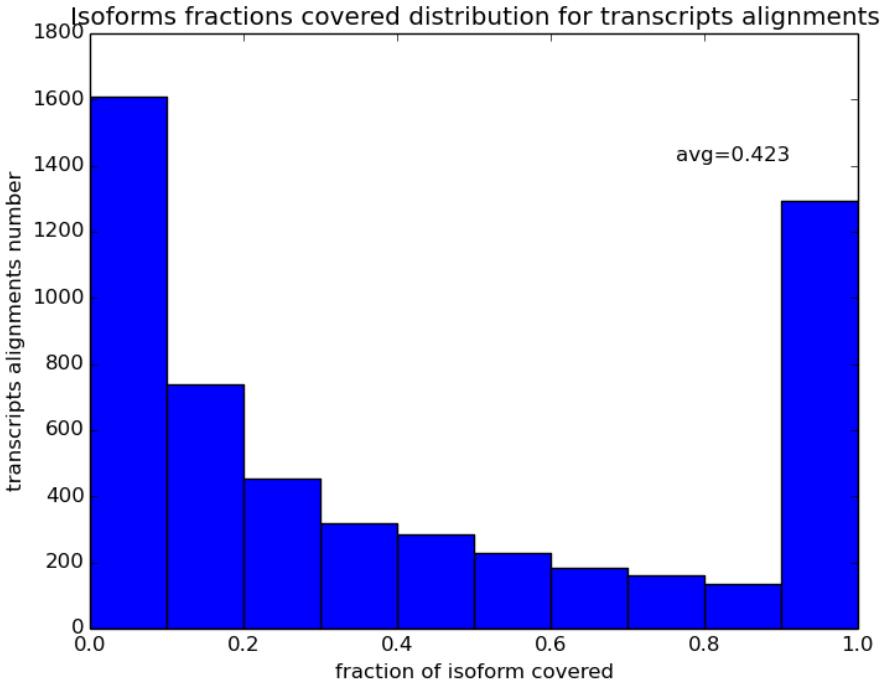


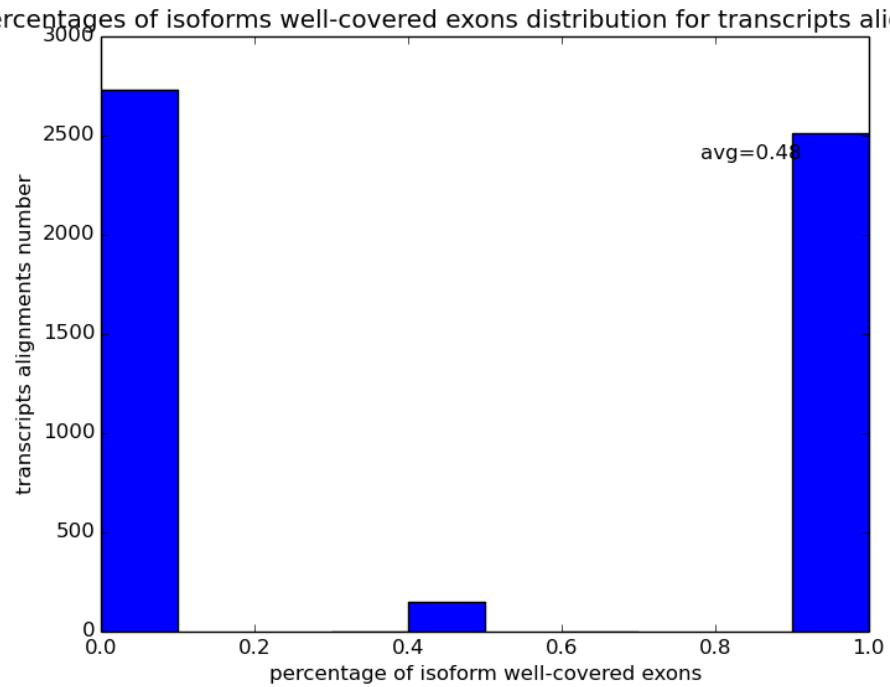


Annotation coverage by alignments

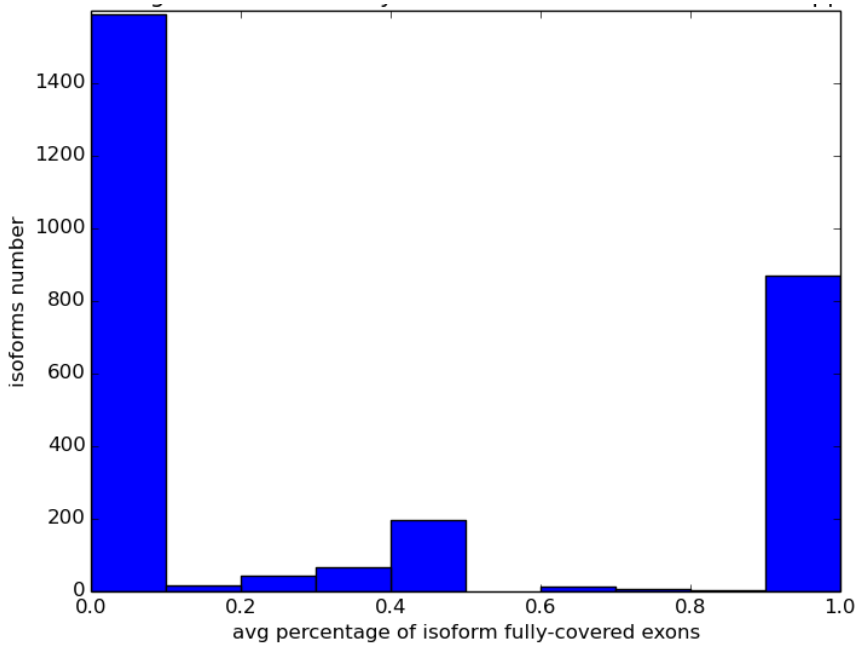
- *For alignments covering a single isoform*
  - *Distributions for transcripts alignments*
    - **Fraction of isoform covered**
    - **Average fraction of exon covered**
    - **Percentage of isoform well-covered/fully-covered exons**
  - *For all transcripts mapped to specific isoform*
    - **Average fraction covered**
    - **Average fraction of exon covered**
    - **Average percentage of well-covered/fully-covered exons**

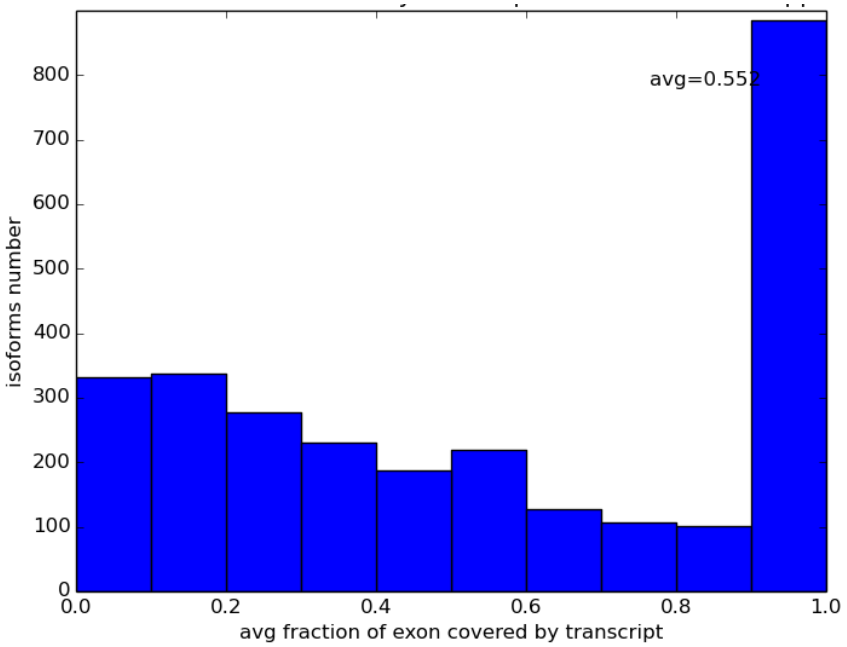
Results for transcripts alignments





Results for transcripts mapped to specific isoform

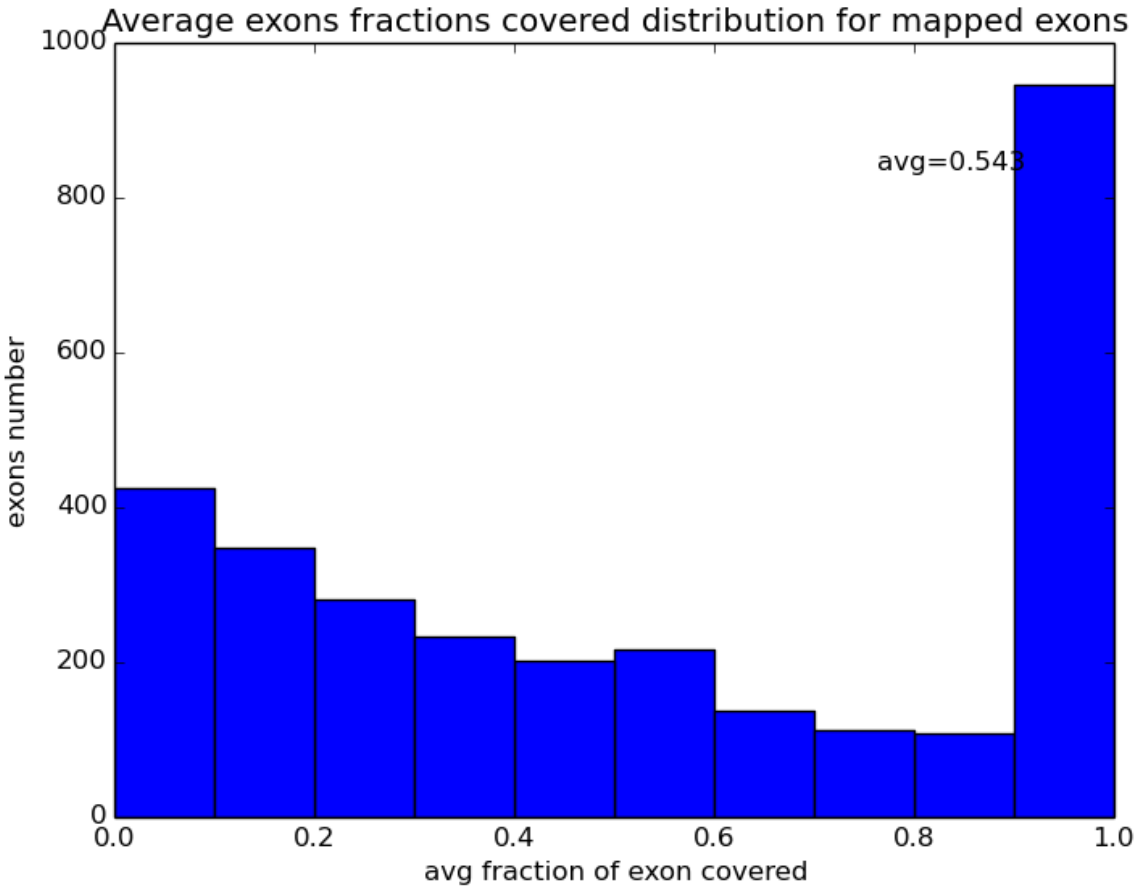




Annotation coverage by alignments

- *For alignments covering a single isoform*
  - *For all transcripts mapped to specific exon:*
    - For each exon calculate **average fraction covered**, calculate distribution
- *For all transcripts alignments*
    - **Fraction of the annotation mapped**
    - **Percentage of covered/fully-covered isoform**
    - **Percentage of well-covered/fully-covered exons**
    - **Average fraction of exon covered**

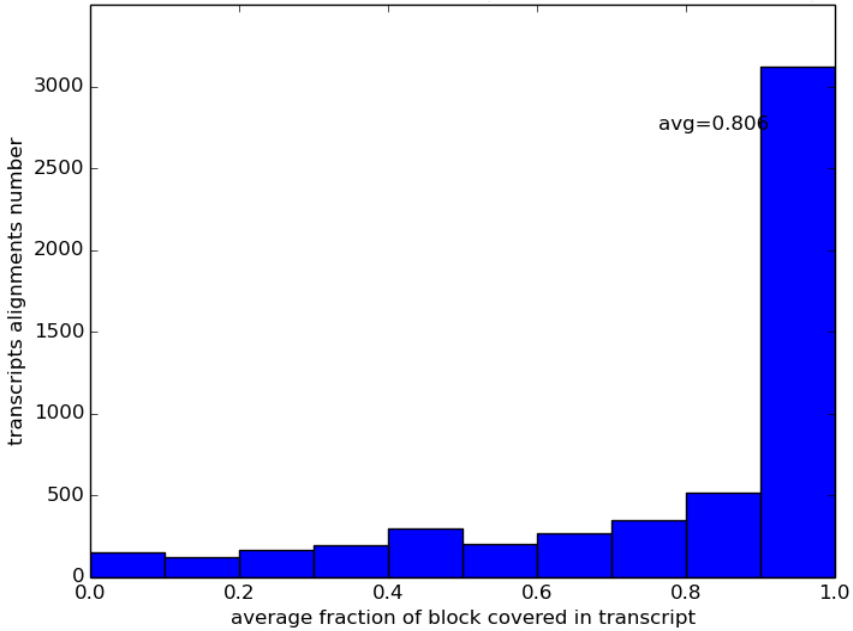
Results for transcripts mapped to specific exon

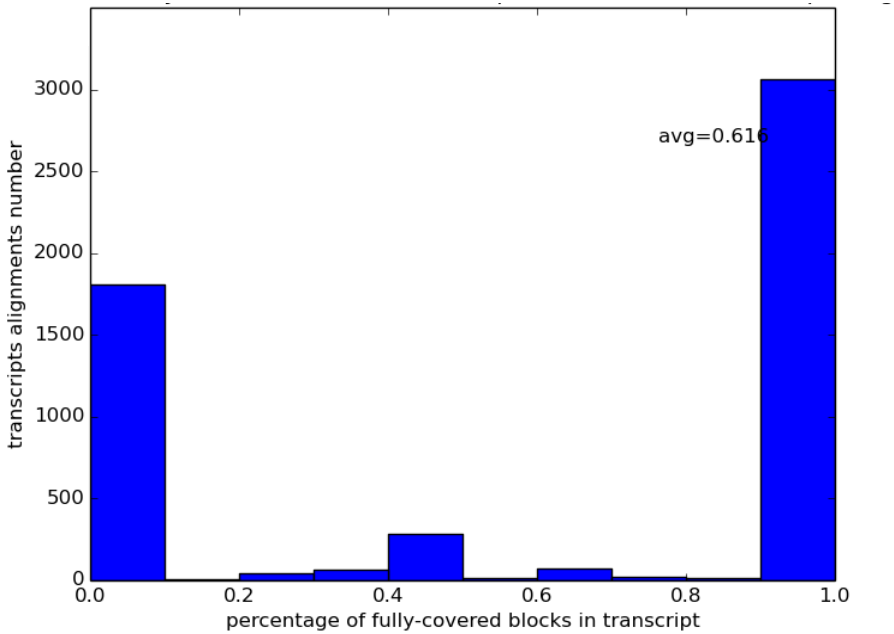


Aligned transcripts coverage by annotation

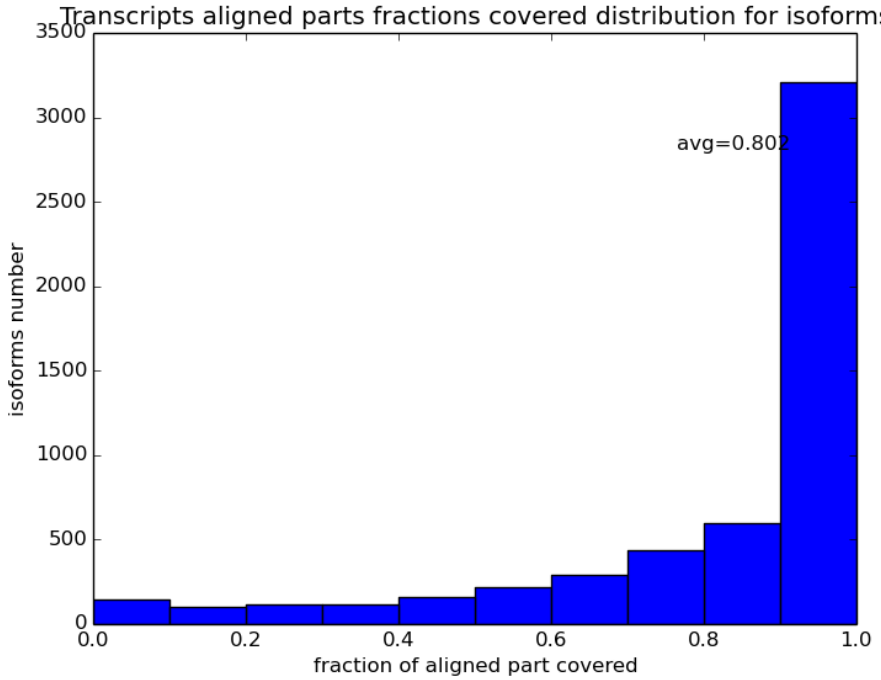
- *For alignments covering a single isoform*
  - *Distributions for transcripts alignments*
    - **Fraction of whole transcript/aligned part covered**
    - **Average fraction of block covered** in transcript
    - **Percentage of well-covered/fully-covered blocks** in transcript
  - *For all transcripts mapped to specific isoform*
    - **Fraction of whole transcript/aligned part covered**

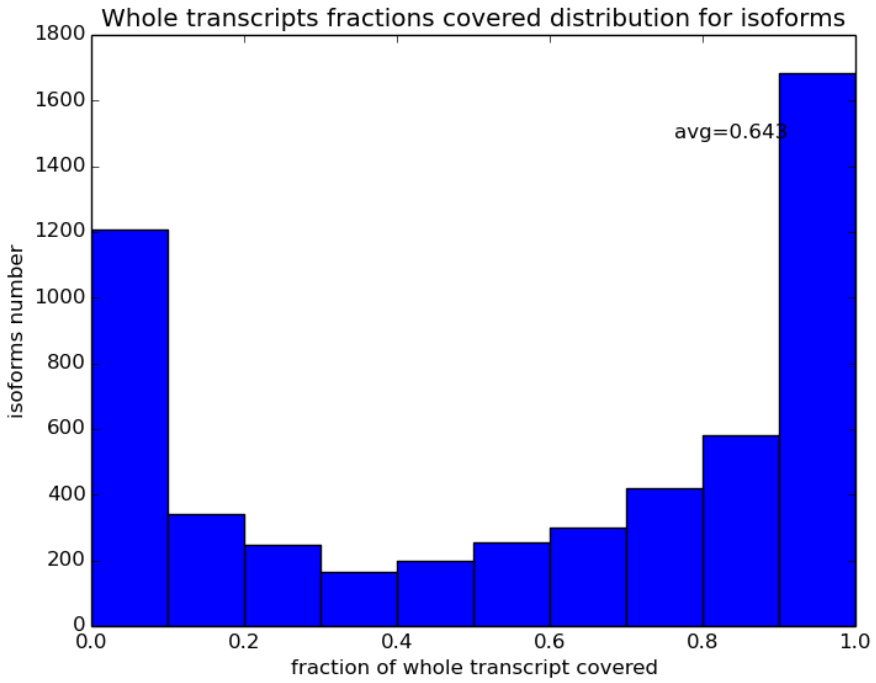
Results for transcripts alignments





Results for transcripts mapped to specific isoform





Aligned transcripts coverage by annotation

- *For all transcripts alignments*
  - **Number/percentage of unannotated transcripts alignments**
  - **Percentage of well-covered/fully-covered transcripts alignments**
  - **Percentage of well-covered/fully-covered blocks**
  - **Average fraction of block covered**

Short report for S.cerevisiae

SIMPLE TRANSCRIPTS ALIGNMENT METRICS

metrics / transcripts	ERR472933_Trinity	ERR472933_SPAdes	SRR453567_Trinity	SRR453567_SPAdes
avg alignment length	311	**347**	**1517**	1459
avg alignment fraction	**0.961**	0.917	**0.86**	0.811
avg blocks number	1.11	**1.151**	**1.543**	1.409
avg block length	280	**301**	983	**1035**
# aligned transcripts	2149	**3268**	**6315**	5321
# unaligned transcripts	22	**75**	105	**121**
# multiple aligned transcripts	164	**327**	588	**1110**

Short report for S.cerevisiae

ANNOTATION COVERAGE BY ALIGNMENTS

metrics / transcripts	ERR472933_Trinity	ERR472933_SPAdes	SRR453567_Trinity	SRR453567_SPAdes
avg fraction of isoform covered	0.229	**0.249**	**0.508**	0.432
avg fraction of exon covered	0.223	**0.244**	**0.508**	0.43
avg % of well-covered exons	0.225	**0.254**	**0.594**	0.477
avg % of fully-covered exons	0.028	**0.043**	**0.294**	0.275
% of well-covered isoforms	0.288	**0.329**	**0.782**	0.658
% of fully-covered isoform	0.033	**0.055**	**0.444**	0.419

Short report for S.cerevisiae

ALIGNED TRANSCRIPTS COVERAGE BY ANNOTATION

metrics / transcripts	ERR472933_Trinity	ERR472933_SPAdes	SRR453567_Trinity	SRR453567_SPAdes
avg fraction of whole transcript covered	**0.824**	0.799	**0.694**	0.619
avg fraction of aligned part covered	**0.852**	0.851	0.801	**0.825**
avg fraction of block covered	**0.847**	0.845	0.777	**0.802**
avg % of well-covered blocks	**0.938**	0.936	0.879	**0.89**
avg % of fully-covered blocks	**0.654**	0.647	0.538	**0.595**
% of unannotated transcripts	**0.001**	0.003	0.003	0.003
% of well-covered transcripts	0.946	**0.948**	0.923	**0.929**
% of fully-covered transcripts	**0.652**	0.641	0.517	**0.585**

Fusion genes / misassemblies detection algorithm

- Filter out short repeat alignments (greedy algorithm)
- Select transcripts with more than 1 significant alignment
- Classify multiple-aligned transcripts
  - Fusions / misassemblies
  - Novel isoforms
- Check all potential misassemblies and fusions by aligning reads and read-pairs
- Accumulating information about distance between genes, repeats and reads alignment, score all potential misassemblies and fusions

Fusion genes / misassemblies metrics

- Multiple-mapped transcripts distribution
- Number of transcripts containing consecutive genes
- Number of of transcripts covering several non consecutive genes
- Number of of transcripts covering different isoforms of the same genes
- Number of of transcripts covering both annotated and unannotated regions
- Number of transcripts covering unannotated regions only
- Number of of transcripts with strongly overlapping blocks/alignments

Thank you!

Questions?