

# Pseudo-aligners

---

RNA-seq data analysis

Johan Reimegård

# Pseudo-aligners assigns read to a transcript

- Not the actual location...
  - It does it by matching k-mers between read and transcripts

# k -mers are nucleotides of length k

- Oct4 is 1574 nt long (  $L = 1574$  )
- k 7 (k= 7)
- Oct4 will contain 1568 Kmers ( $L-k+1$ )

.....CTTGGAAACAAT.....

CTTGGAA

TTGGAAC

TGGAACA

GGAACAA

GAACAAT

# Splits up a read into the same k-mer size

Read1 = CTTGGAACAAT

Kmer Read1
CTTGGAA
TTGGAAC
TGGAACA
GGAACAA
GAACAAT
AACAATA

# Checks in which transcripts the Kmers exist and sums them up

Kmer Read1	Kmer	Oct4	Oct3	Oct2	Sox2	Sox3
CTTGGAA	CTTGGAA	TRUE	FALSE	TRUE	FALSE	FALSE
TTGGAAC	TTGGAAC	TRUE	FALSE	TRUE	FALSE	FALSE
TGGAACA	TGGAACA	TRUE	FALSE	FALSE	FALSE	FALSE
GGAACAA	GGAACAA	TRUE	TRUE	FALSE	FALSE	FALSE
GAACAAT	GAACAAT	TRUE	TRUE	FALSE	FALSE	FALSE
AACAATA	AACAATA	TRUE	FALSE	FALSE	FALSE	FALSE



Read	Oct4	Oct3	Oct2	Sox2	Sox3
Read 1	6	2	2	0	0

# Assign the read to one or many transcript

Checks which of the transcripts the number of kmers matched is least likely to happen by chance and assign it to those transcript

Read	Oct4	Oct3	Oct2	Sox2	Sox3
Read 1	6	2	2	0	0



Assign read to transcripts

Read	Oct4
Read 1	1



Add read counts to transcripts in sample

Read	Oct4	Oct3	Oct2	Sox2	Sox3
Sample 1	+1	0	0	0	0

# Redo the procedure for all reads

Read2 = GATACAGATAC  
6 kmers of length 7

Read	Oct4	Oct3	Oct2	Sox2	Sox3
Read 2	0	0	0	6	6



Assign read to transcripts

Read	Sox2	Sox3
Read 2	1	1



Add read counts to transcripts in sample

Read	Oct4	Oct3	Oct2	Sox2	Sox3
Sample 1	1	0	0	+1	+1

# But it takes time to look up so many k-mers

Real result from Kallisto

```
[quant] fragment length distribution will be estimated from the  
[index] k-mer length: 31  
[index] number of targets: 173,259  
[index] number of k-mers: 104,344,666
```

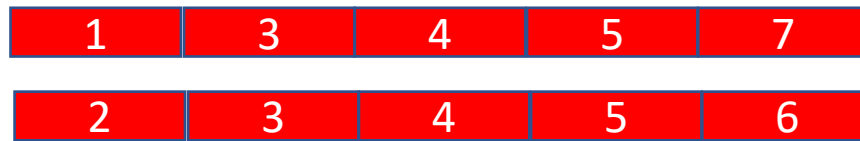
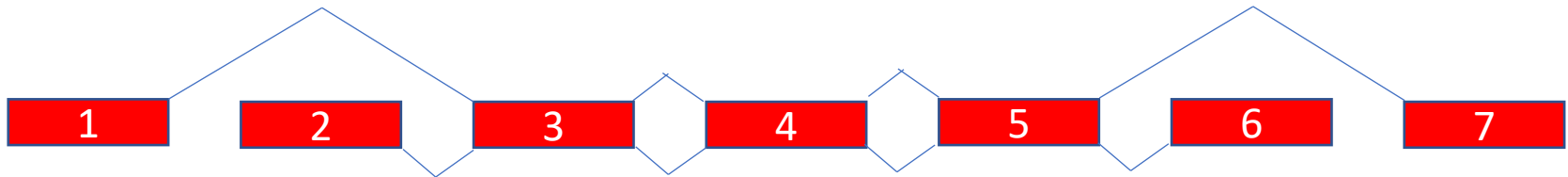


# Speed up by identifying important k-mers to separate transcripts

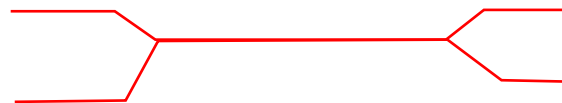
- Building a de Bruijn graph of the k-mers and identifying the important positions to separate different paths
- And using statistics assign the read to a transcript

# Kalisto builds a de Bruijn graph with all the k-mers

Red gene



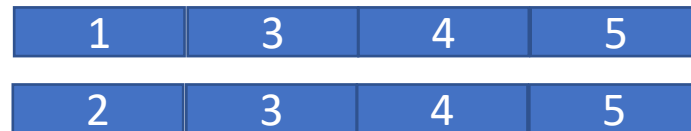
Transcripts



De bruijn graph

# Kalisto builds a graph with all the k-mers

Blue gene

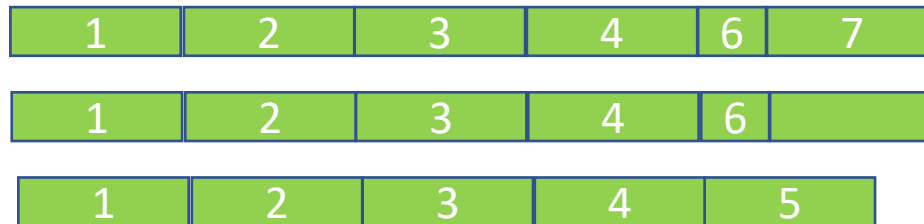


Transcripts

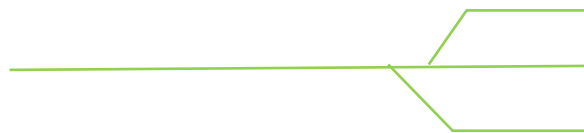


De bruijn graph

# Kalisto builds a graph with all the k-mers

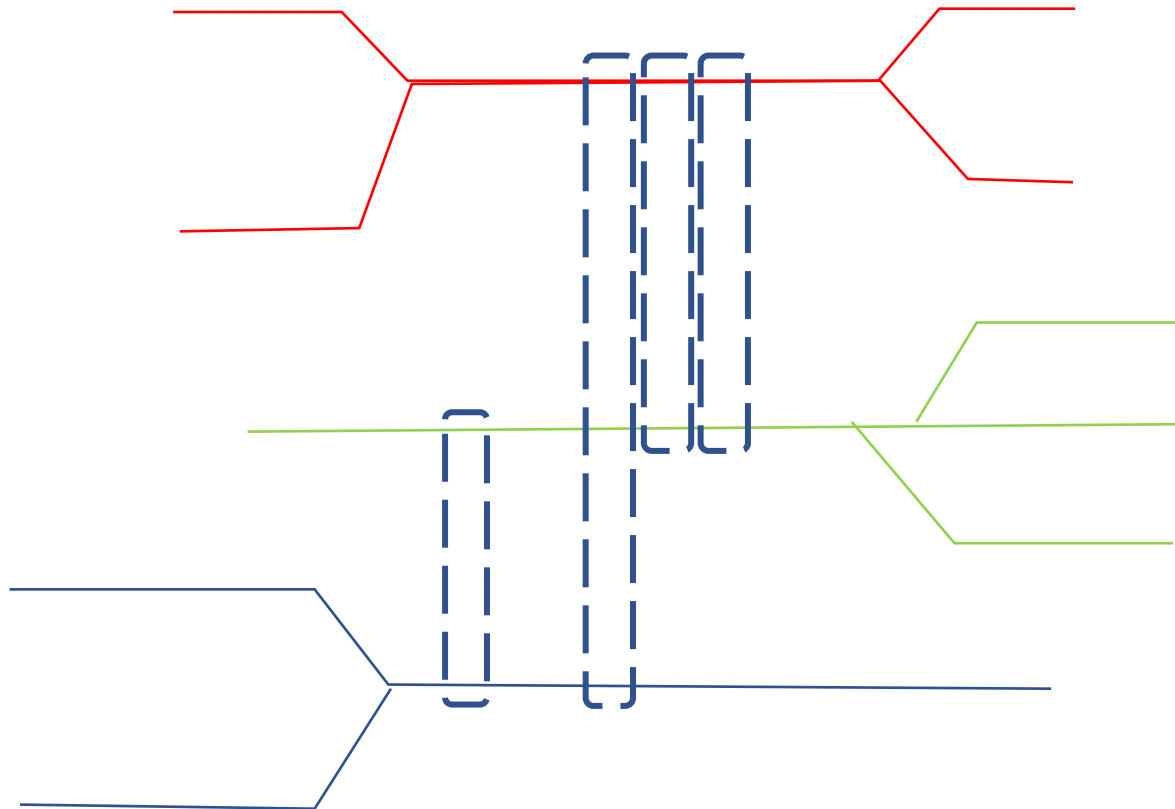


Transcripts

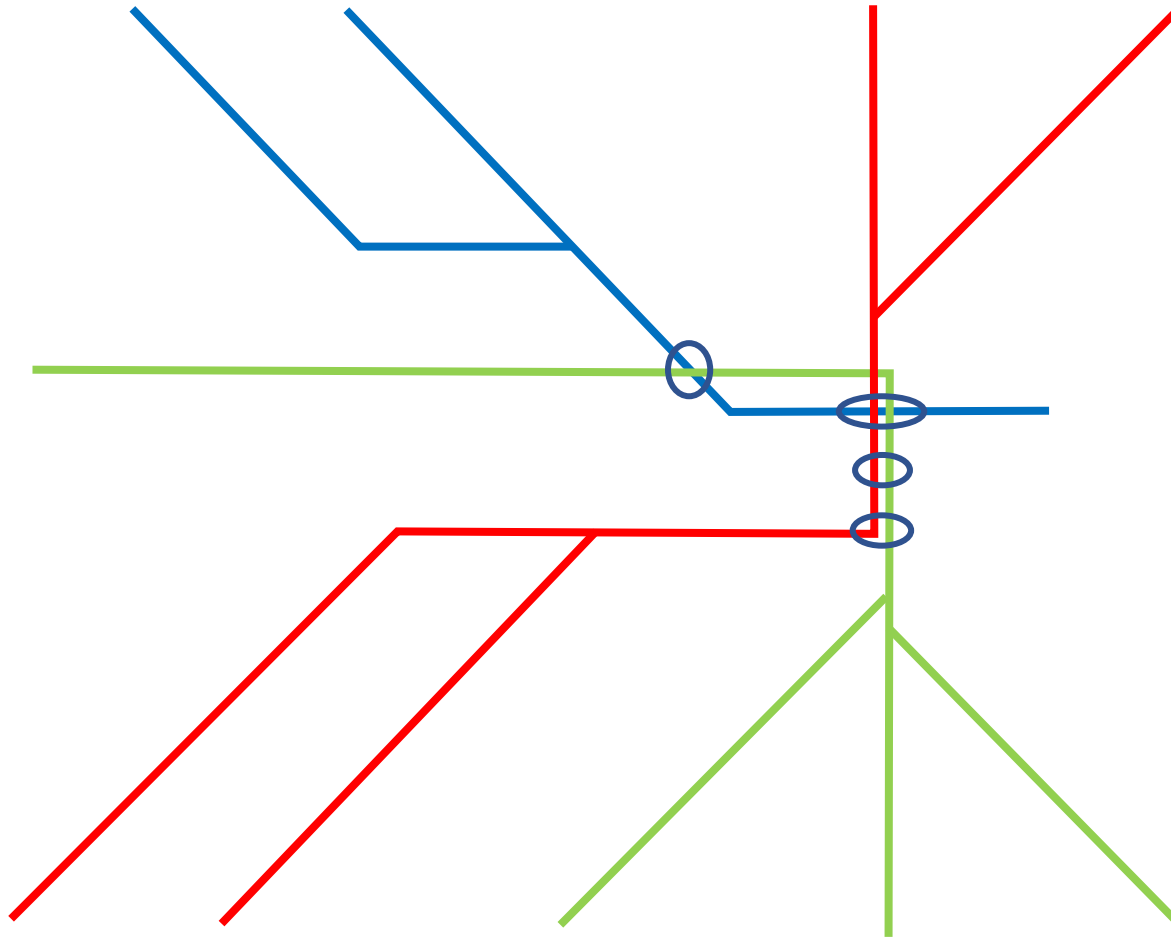


De bruijn graph

3 gene de-bruin graph with parts in common.



# 3 gene de-bruin graph



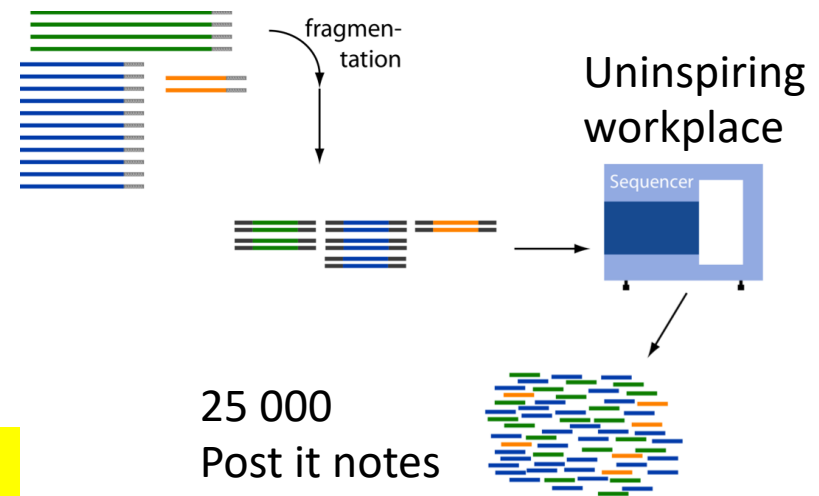
# Stockholm subway de-bruin graph



>T10 transport (transcript), Blue Line  
Hjulsta, Tensta, Rinkeby, Rissne, Duvbo, Sundbybergs centrum, ...  
..., T-Centralen, Kungsträdgården.

# What line was this person travelling

- One transport sequence per person
- Sequence so long so only one fragment is saved per travel sequence (~100 character)
- Much work to write entire sequence to post-it note so only first 25 characters are saved for each fragment.
- Total 25 000 post it notes



```
>Post it 1  
nsta,Rinkeby,Rissne,Duvbo
```



# How many persons are travelling on the different lines

- You can only remember 10 characters at the time.

```
>Post it 1  
nsta,Rinkeby,Rissne,Duvbo
```

```
Post it 1:k-mer01  
nsta,Rinke
```

```
Post it 1:k-mer02  
sta,Rinkeb
```

```
Post it 1:k-mer16  
ssne,Duvbo
```

# Align travelers transport to graph



Post-it 1: K-mer 1  
nsta, Rinke

T17	T18	T19	T10	T11	T13	T14
0	0	0	0	0	0	0

# Align travelers transport to graph



Post-it 2: K-mer 1  
a, Universi

T17	T18	T19	T10	T11	T13	T14
0	0	0	1	0	0	0

# Align travelers transport to graph



Post-it 3: K-mer 1  
T-Centrale

T17	T18	T19	T10	T11	T13	T14
0	0	0	1	0	0	1

# Align travelers transport to graph



Post-it 3: K-mer 1  
T-Centrale

Post-it 3: K-mer 2  
-Centralen

T17	T18	T19	T10	T11	T13	T14
0	0	0	1	0	0	1

# Align travelers transport to graph



Post-it 3: K-mer 1  
T-Centrale

Post-it 3: K-mer 2  
-Centralen

Post-it 3: K-mer 3  
Centralen,

T17	T18	T19	T10	T11	T13	T14
0	0	0	1	0	0	1

# Align travelers transport to graph



Post-it 3: K-mer 1  
T-Centrale

Post-it 3: K-mer 2  
-Centralen

Post-it 3: K-mer 3  
Centralen,

Post-it 3: K-mer 4  
entralen, K

T17	T18	T19	T10	T11	T13	T14
0	0	0	1	0	0	1

# Align travelers transport to graph



Post-it 4: K-mer 1  
T-Centrale

T17	T18	T19	T10	T11	T13	T14
0	0	0	2	1	0	1



# Align travelers transport to graph



Post-it 4: K-mer 1  
T-Centrale

Post-it 4: K-mer 4  
entralen, G

T17	T18	T19	T10	T11	T13	T14
0	0	0	2	1	0	1

# Align travelers transport to graph



Post-it 4: K-mer 1  
T-Centrale

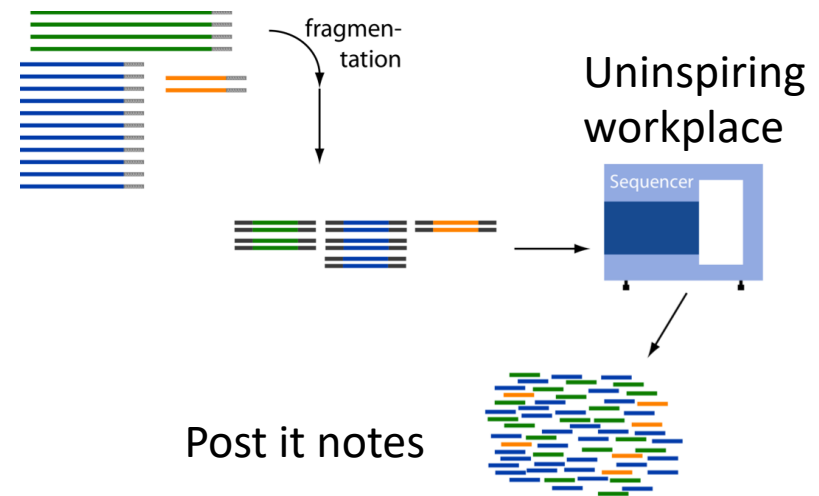
Post-it 4: K-mer 4  
entralen, G

Post-it 4: K-mer 16  
la stan, S1

T17	T18	T19	T10	T11	T13	T14
1	1	1	2	1	1	2

# What line was this person travelling

- Sequence so long so only one fragment is saved per travel sequence (~100 character)
- Much work to write entire to post-it note so only first 10 AND last 10 characters are saved for each fragment.
- Total 25 000 post it notes



```
>Post it 4  
Start:  
T-Centrale  
  
End:  
,Globen,En
```

# Align travelers transport to graph



>Post it 4  
 Start:  
 T-Centrale  
 End:  
 ,Globen,En

T17	T18	T19	T10	T11	T13	T14
1	1	1	2	1	1	2

# Align travelers transport to graph



>Post it 5  
 Start:  
 Fridhemspl  
 End:  
 -Centralen

T17	T18	T19	T10	T11	T13	T14
1	1	2	2	1	1	2

# So they divide it up to classes

## Real result from Kallisto

```
[quant] fragment length distribution will be estimated from the
[index] k-mer length: 31
[index] number of targets: 173,259
[index] number of k-mers: 104,344,666
index] number of equivalence classes: 695,212
[quant] running in paired-end mode
[quant] will process pair 1: fastq/test.1.fastq.gz fastq/test.2.fastq.gz
[quant] finding pseudoalignments for the reads ... done
[quant] learning parameters for sequence specific bias
[quant] processed 92,206,249 reads, 82,446,339 reads pseudoaligned
[quant] estimated average fragment length: 187.018
```

# Align travelers transport to graph

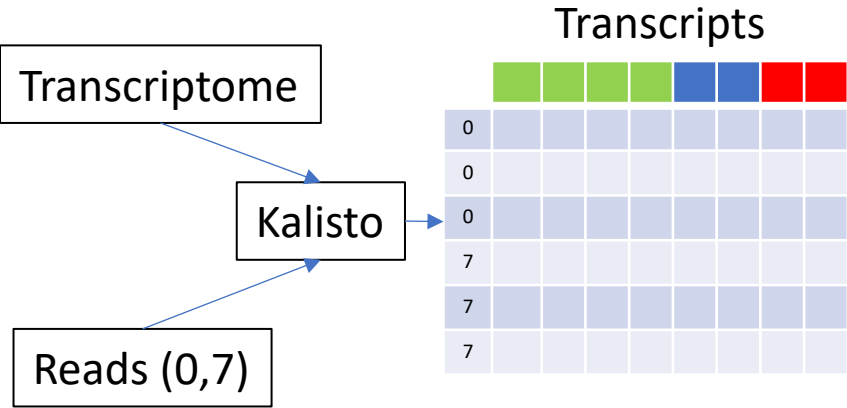


Hässl  
 Hässelb  
 For the majority of reads, kallisto ends up performing a hash lookup for only two k-mers

>Post it 5  
 Start:  
 Fridhemsp1  
  
 End:  
 -Centralen

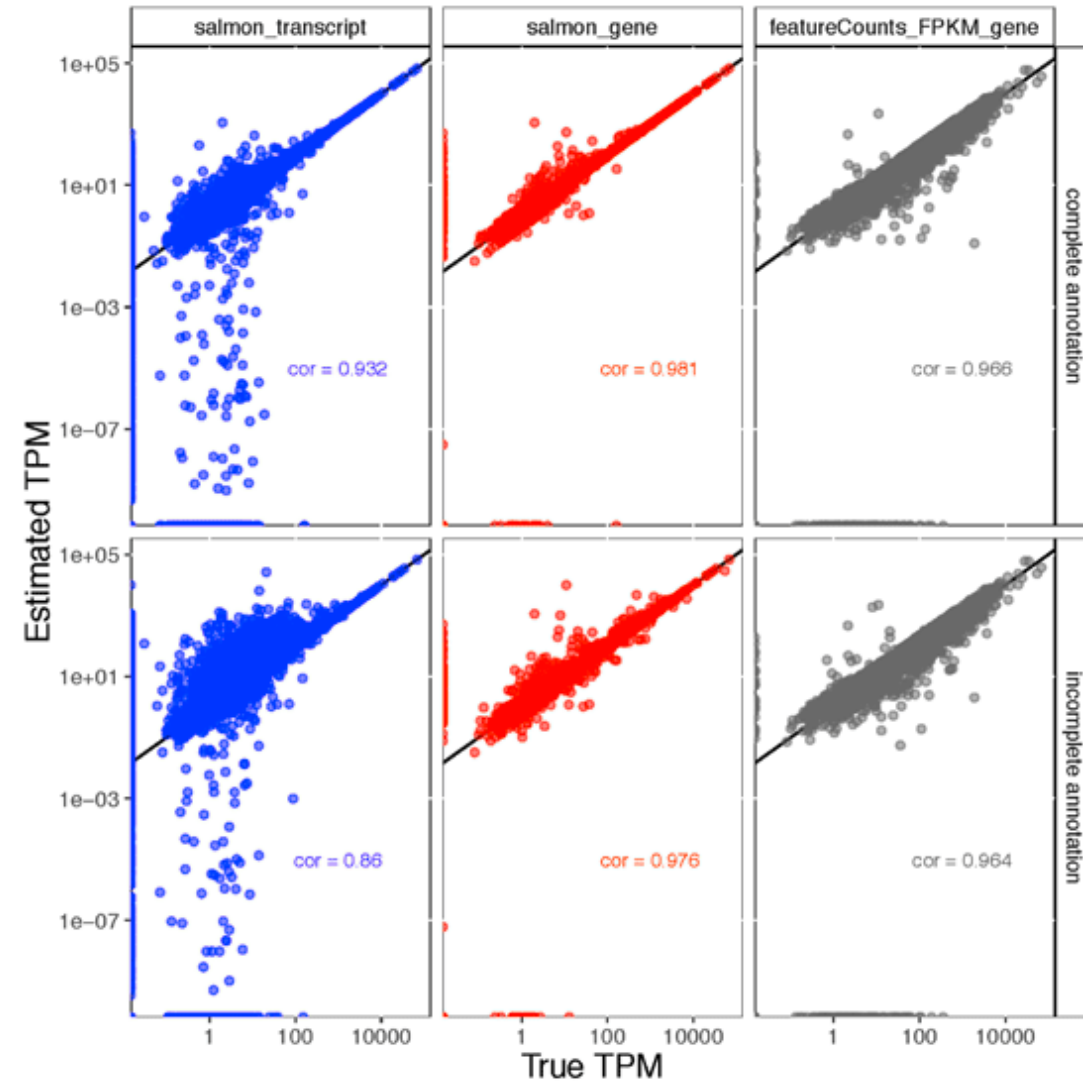
T17	T18	T19	T10	T11	T13	T14
1	1	2	2	1	1	2

# Map reads to transcriptome



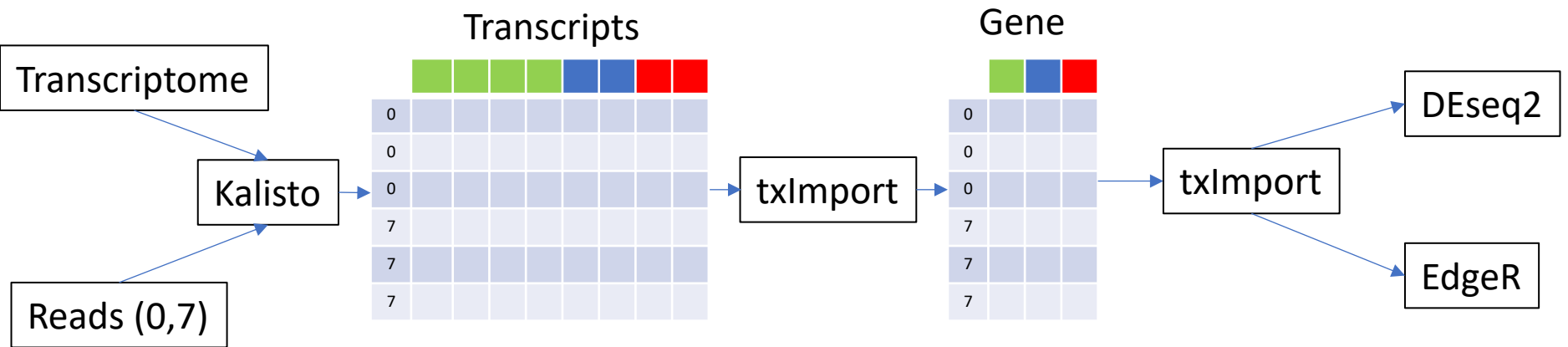


# Better estimations on gene level



In this article, we have contrasted transcript- and gene-resolution analyses in terms of both abundance estimation and statistical inference, and illustrated that gene-level results are often more accurate, powerful and interpretable than transcript-level results.

# Convert from transcript to gene using tximport





**Thank you. Questions?**

---

**Johan Reimegård**