Lecture 6: Alignment & Quantification

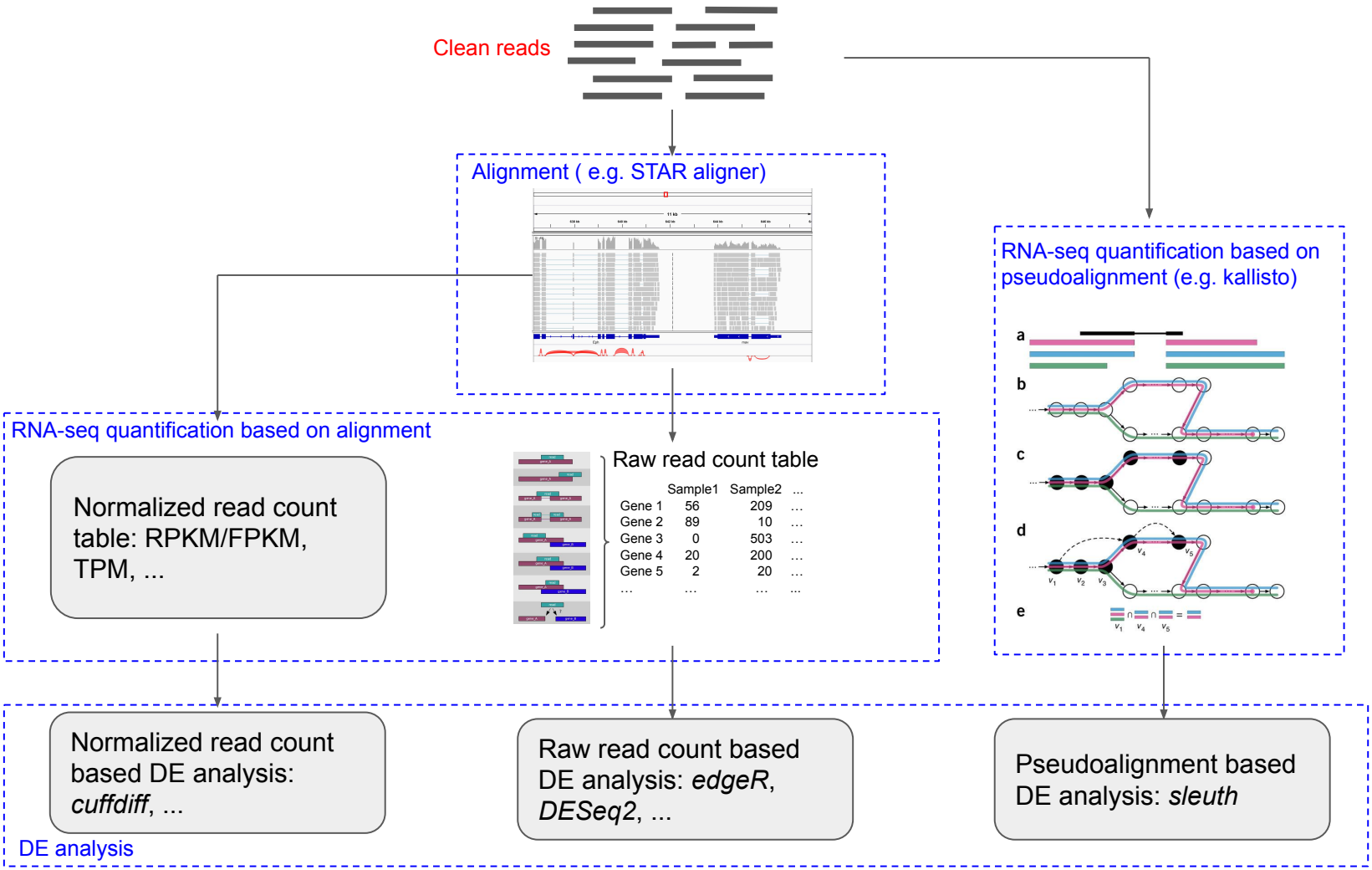
BIOINF3005/7160: Transcriptomics Applications

Zhipeng Qu

School of Biologial Sciences,
The University of Adelaide

April 27th, 2020

Alignment and quantification in RNA-Seq



Outline

- Part 1, RNA-Seq alignment
- Part 2, RNA-Seq quantification
- Part 3, Pseudoalignment

Part 1, RNA-Seq alignment

- Short read alignment
- Introduction of STAR aligner

Sequence alignment

```
G C T G G A A G - G C A T
        |   | |   | | |
        G C A G A G C A C T
```

6 matches: 6 × 5 = 30
1 mismatch: -4
1 indel: -7
Total: 19

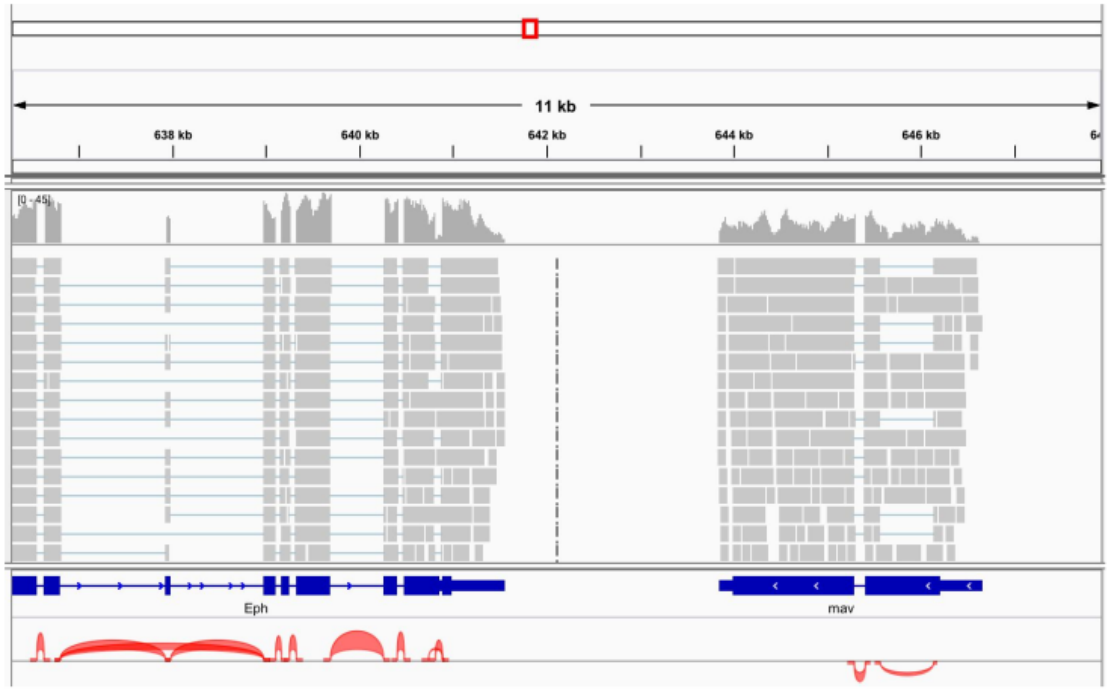
http://www.csbio.unc.edu/mcmillan/Comp555S16/Lecture14.html

Short read alignment

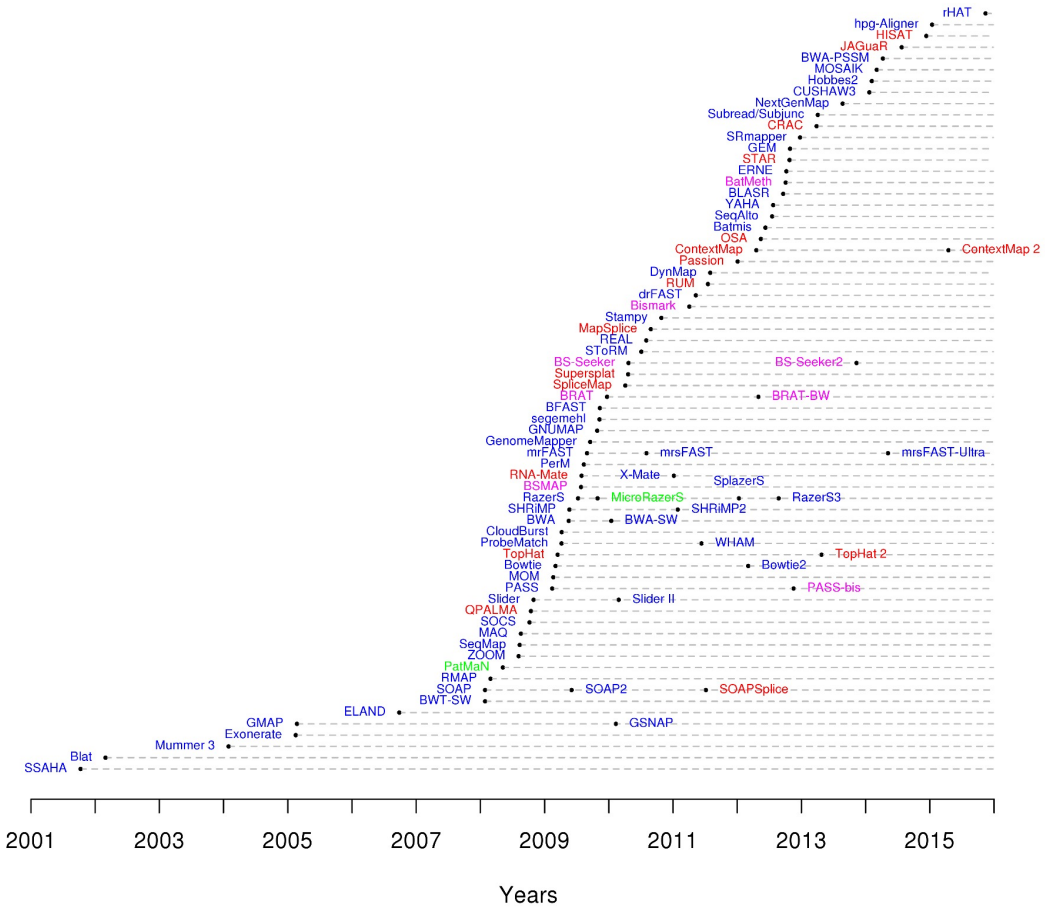
Also called read mapping, align (map) short reads from NGS to reference genome (if available, DNA-Seq/RNA-Seq) or transcriptome (RNA-Seq).

Challenges in RNA-Seq alignment:

- millions of short reads (DNA-Seq/RNA-Seq)
- RNA splicing



Short read alignment tools (short aligners)



Blue color: DNA aligner; Red color: RNA aligner; Orchid color: Bisulfite Sequencing aligner; Green color: miRNA-Seq aligner

---

https://www.ebi.ac.uk/~nf/hts_mappers/

Which short aligner should I use?

- Sequencing input type: DNA vs RNA
- Reference sequences: Genome vs Transcriptome
- Available computing resources

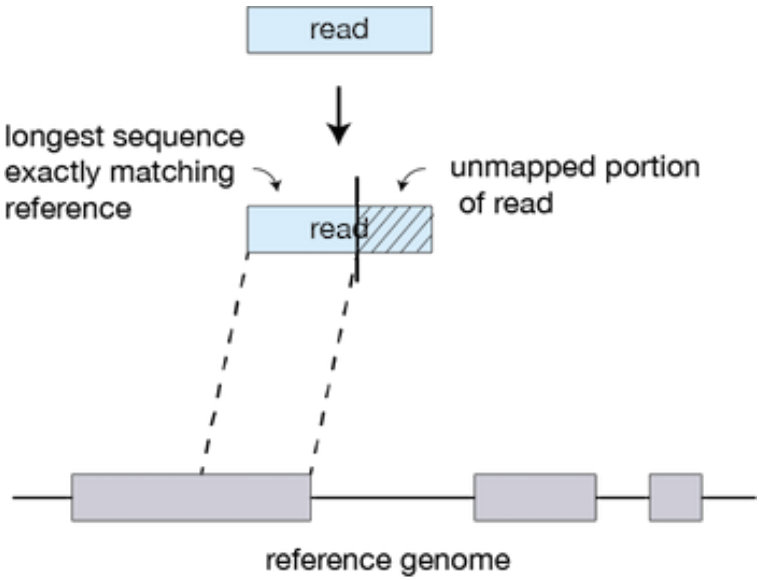
STAR aligner

STAR (Spliced Transcripts Alignment to a Reference)

- Outperforms other aligners by more than a factor of 50 in mapping speed
- Memory intensive. At least 10x Genome size (for example, ~30 Gb for human genome)
- Written in C++, only works on Linux or Mac OS
- Unbiased de novo detection of canonical junctions
- Discovers non-canonical splices and fusion transcripts

Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21.

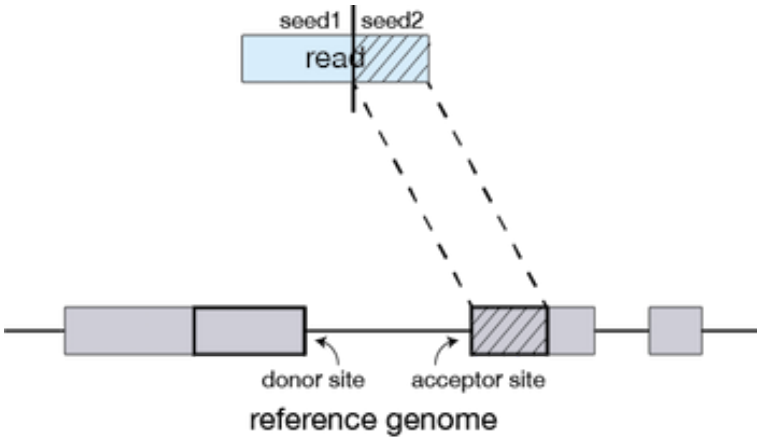
STAR alignment strategy: Seed searching



Search for the longest sequence in read exactly matching the reference genome, called the Maximal Mappable Prefixes (MMPs)

https://hbctraining.github.io/Intro-to-rnaseq-hpc-O2/lessons/03_alignment.html

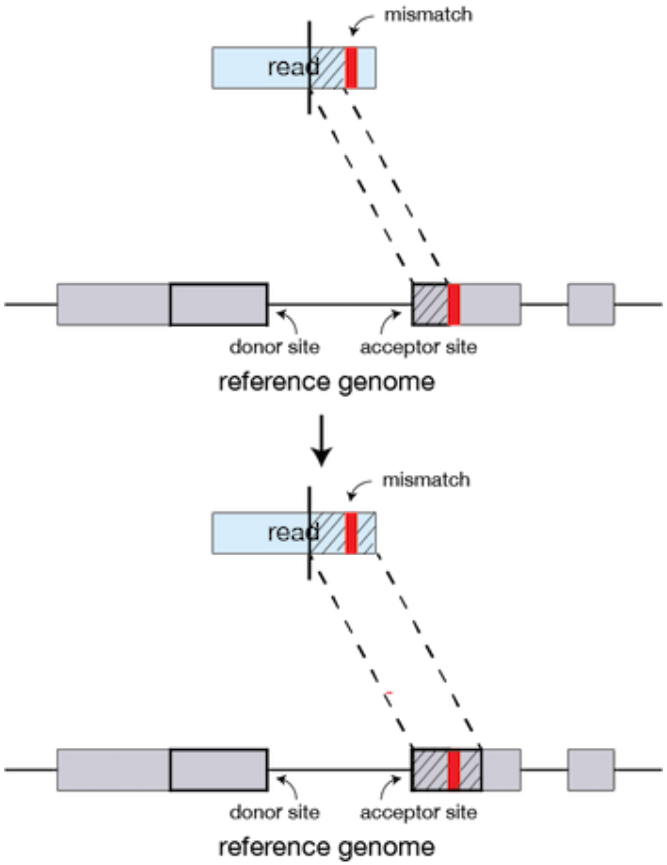
STAR alignment strategy: Seed searching



MMPs are sequentially searched and called as "seeds", e.g. seed1, seed2, ...

https://hbctraining.github.io/Intro-to-rnaseq-hpc-O2/lessons/03_alignment.html

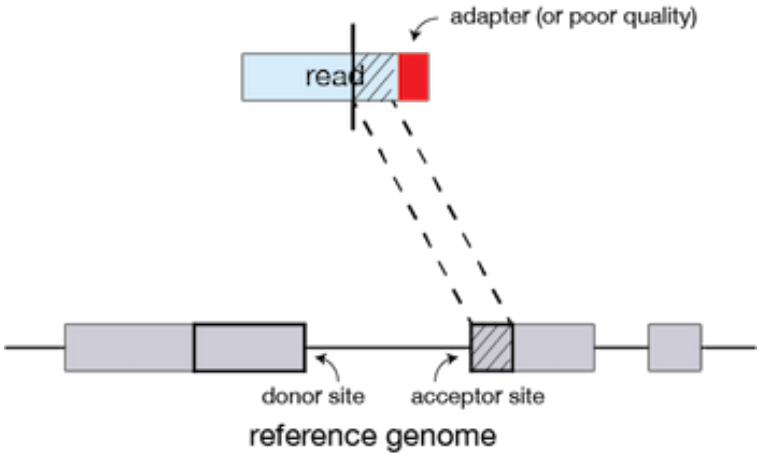
STAR alignment strategy: If STAR does not find an exact matching sequence



The previous MMPs will be extended

https://hbctraining.github.io/Intro-to-rnaseq-hpc-O2/lessons/03_alignment.html

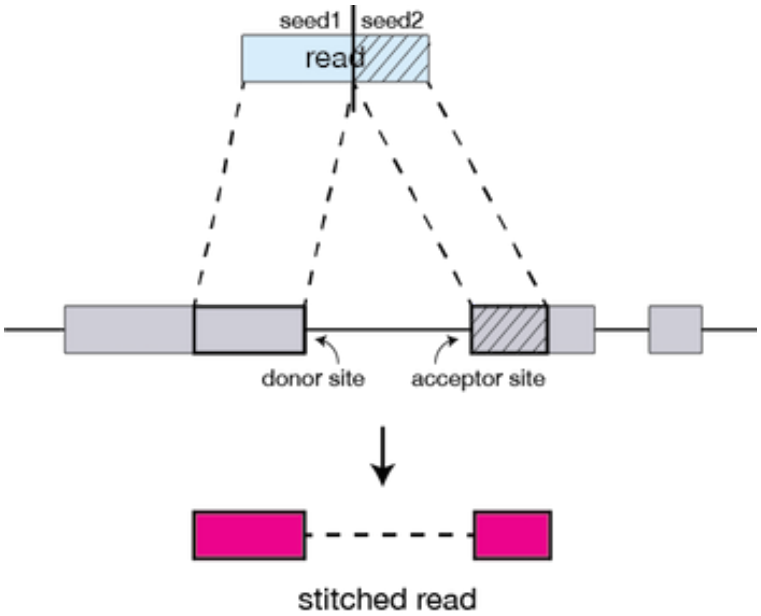
STAR alignment strategy: If extension does not give a good alignment



The poor quality or adaptor sequence (or other contaminating sequence) will be soft-clipped

https://hbctraining.github.io/Intro-to-rnaseq-hpc-O2/lessons/03_alignment.html

STAR alignment strategy: Clustering, stitching and scoring



The seperate seeds are clustered and then stitched together based on the best scoring of alignment (mismatches, indels, gaps, etc.)

https://hbctraining.github.io/Intro-to-rnaseq-hpc-O2/lessons/03_alignment.html

Adjusting alignment parameters of STAR

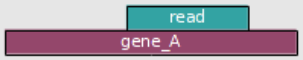
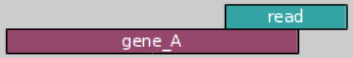
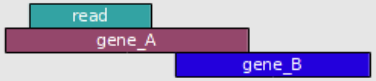
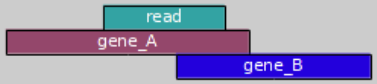
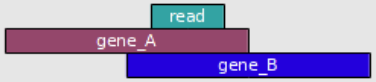
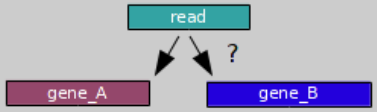
Some useful parameters:

- --outFilterMultimapNmax
- --outFilterMismatchNmax
- --outFilterMismatchNoverLmax
- --quantMode (GeneCounts)

Important: only adjust parameters when you know what you are doing!!

- Read count
- Multiple mapping
- Normalization of read count

Three read count modes

	union	intersection _strict	intersection _nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous (both genes with --nonunique all)	gene_A	gene_A
	ambiguous (both genes with --nonunique all)		
	alignment_not_unique (both genes with --nonunique all)		

https://htseq.readthedocs.io/en/release_0.11.1/count.html

Short reads can be mapped to multiple features (genes/transcripts)

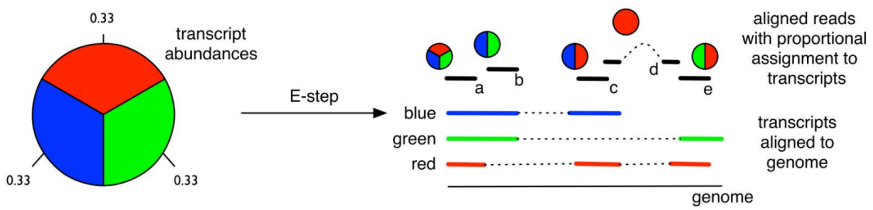
- Identical/similar sequences in different genes (e.g. gene family, repetitive elements)
- Different transcription isoforms from the same gene

Species	**Aligner**	**Read length**	**multiple mapping rate (%)**
Human	STAR	PE100	4.88
Mouse	STAR	PE100	15.72
Rat	STAR	PE75	12.07
Arabidopsis	STAR	PE150	1.41
Rice	Tophat2	PE150	43.7
Soybean	Tophat2	PE150	26.4

Strategies for handling multiple mapping

- Use uniquely mapping reads only
- Simple "rescue" method. Uniformly divide each multi-mapping read to all of the positions it maps to. In other words, a read mapping to 10 positions will count as 10% of a read at each position.
- "Rescue" method using Expectation-Maximization (EM) model
  1. E-step (Expectation) Given transcript abundances, estimate the probability of each read mapping to each transcript
  2. M-step (Maximization) Update the abundances by redistributing the reads
  3. Go to step 1 (E-step) until convergence

"Rescue" method using EM model



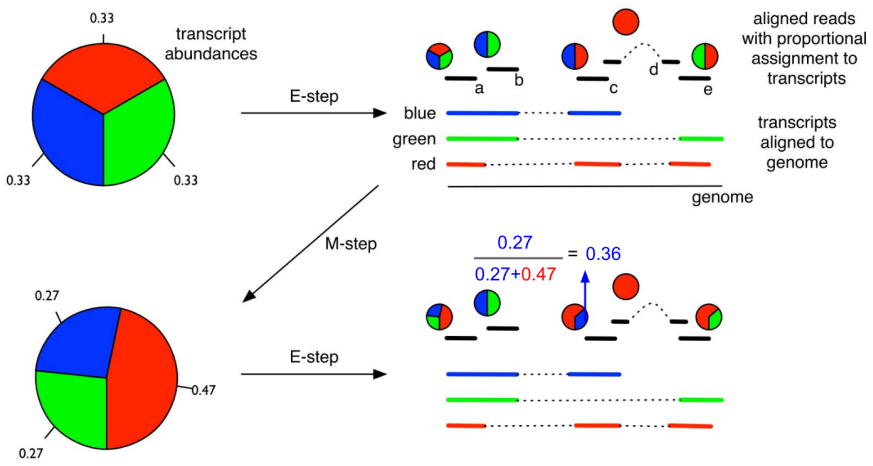
$f_{blue} = (0.33+0.5+0.5)/5 = 0.27$

$f_{green} = (0.33+0.5+0.5)/5 = 0.27$

$f_{red} = (0.33+0.5+1+0.5)/5 = 0.47$

Pachter L. Models for transcript quantification from RNA-Seq. arXiv. 2011

"Rescue" method using EM model



f_{blue} = (0.33+0.5+0.5)/5 = 0.27

f_{green} = (0.33+0.5+0.5)/5 = 0.27

f_{red} = (0.33+0.5+1+0.5)/5 = 0.47

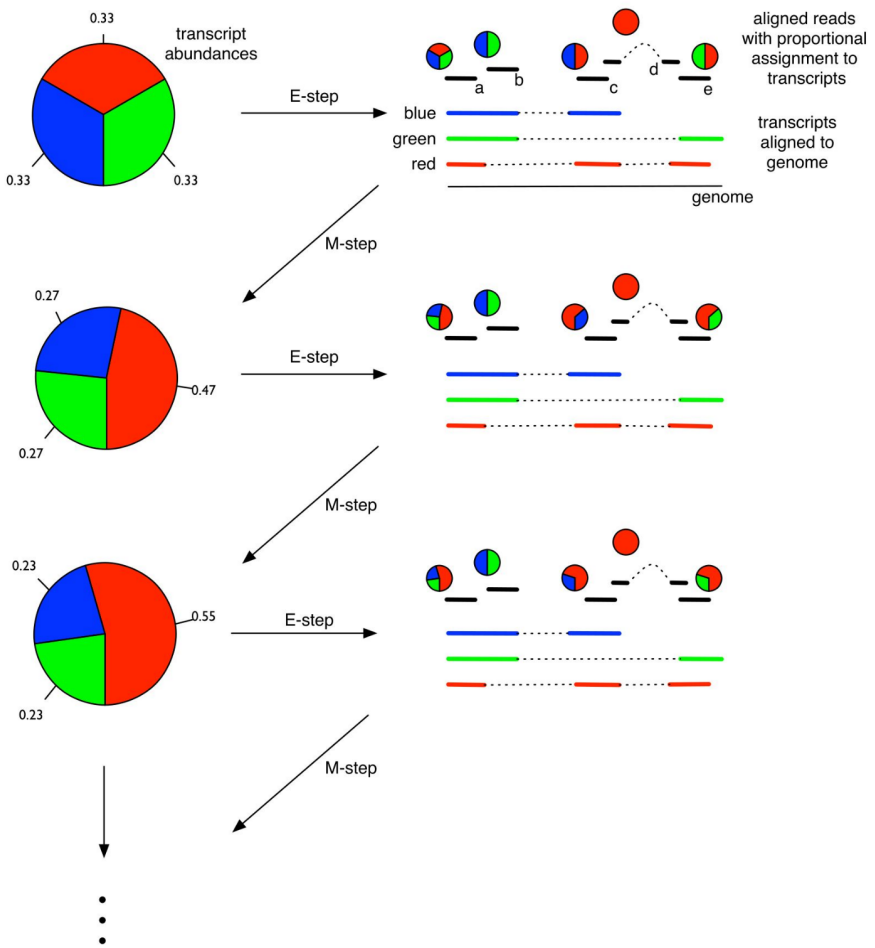
f_{blue} = (0.27+0.5+0.36)/5 = 0.23

f_{green} = (0.27+0.5+0.36)/5 = 0.23

f_{red} = (0.47+0.64+1+0.64)/5 = 0.55

Pachter L. Models for transcript quantification from RNA-Seq. arXiv. 2011

"Rescue" method using EM model



f_{blue} = (0.33+0.5+0.5)/5 = 0.27

f_{green} = (0.33+0.5+0.5)/5 = 0.27

f_{red} = (0.33+0.5+1+0.5)/5 = 0.47

f_{blue} = (0.27+0.5+0.36)/5 = 0.23

f_{green} = (0.27+0.5+0.36)/5 = 0.23

f_{red} = (0.47+0.64+1+0.64)/5 = 0.55

...

Pachter L. Models for transcript quantification from RNA-Seq. arXiv. 2011

RNA-Seq is a relative abundance measurement of RNA expression level

- Short reads are RNA fragments randomly picked and sequenced from library
- Additional information, such as levels of "spike-in" transcripts, are required for absolute measurements
- Normalization of read count is needed to compare gene/transcript abundance
  1. RPKM/FPKM (Reads/Fragments Per Kilobase Million)
  2. TPM (Transcripts Per Million)

RPKM

Gene ID	Length	Rep1	Rep2	Rep3
A	2 kb	10	12	30
B	4 kb	20	25	60
C	1 kb	5	8	15
D	10 kb	0	0	1

We assume:

1) The genome has 4 genes
2) The RNA-Seq dataset has three replicates

https://www.youtube.com/watch?v=TTUrtCY2k-w

RPKM

Gene ID	Length	Rep1	Rep2	Rep3
A	2 kb	10	12	30
B	4 kb	20	25	60
C	1 kb	5	8	15
D	10 kb	0	0	1

Replicate 3 has much more reads than the other two replicates

https://www.youtube.com/watch?v=TTUrtCY2k-w

RPKM

Gene ID	Length	Rep1	Rep2	Rep3
A	2 kb	10	12	30
B	4 kb	20	25	60
C	1 kb	5	8	15
D	10 kb	0	0	1

Gene B is twice as long as gene A, which might explain why it always gets twice as many reads

https://www.youtube.com/watch?v=TTUrtCY2k-w

RPKM

Gene ID	Length	Rep1	Rep2	Rep3
A	2 kb	10	12	30
B	4 kb	20	25	60
C	1 kb	5	8	15
D	10 kb	0	0	1
	Total reads:	35	45	106
	Tens of reads:	3.5	4.5	10.6

"Per Million" scaling factors → Tens of reads:

1) In this example, we scale the total read counts by 10 instead of 1,000,000
2) Million (1,000,000) was chosen just because it made the numbers look nice (Standard RNA-Seq datasets usually have multiple million reads)

https://www.youtube.com/watch?v=TTUrtCY2k-w

RPKM

Gene ID	Length	Rep1	Rep2	Rep3
A	2 kb	10	12	30
B	4 kb	20	25	60
C	1 kb	5	8	15
D	10 kb	0	0	1

Count table

	Total reads:	35	45	106
"Per Million" scaling factors →	Tens of reads:	3.5	4.5	10.6

Gene ID	Length	Rep1	Rep2	Rep3
A	2 kb	2.86	2.67	2.83
B	4 kb	5.71	5.56	5.66
C	1 kb	1.43	1.78	1.42
D	10 kb	0	0	0.09

RPM table

https://www.youtube.com/watch?v=TTUrtCY2k-w

RPKM

Gene ID	Length	Rep1	Rep2	Rep3
A	2 kb	2.86	2.67	2.83
B	4 kb	5.71	5.56	5.66
C	1 kb	1.43	1.78	1.42
D	10 kb	0	0	0.09

RPM table

Scale Per Kilobase

Gene ID	Length	Rep1	Rep2	Rep3
A	2 kb	1.43	1.33	1.42
B	4 kb	1.43	1.39	1.42
C	1 kb	1.43	1.78	1.42
D	10 kb	0	0	0.009

RPKM table

https://www.youtube.com/watch?v=TTUrtCY2k-w

RPKM summary

Gene ID	Length	Rep1	Rep2	Rep3
A	2 kb	10	12	30
B	4 kb	20	25	60
C	1 kb	5	8	15
D	10 kb	0	0	1

Count table

Read count was:

1) Normalized for differences in sequencing depth
2) Normalized for gene length

Gene ID	Length	Rep1	Rep2	Rep3
A	2 kb	1.43	1.33	1.42
B	4 kb	1.43	1.39	1.42
C	1 kb	1.43	1.78	1.42
D	10 kb	0	0	0.009

RPKM table

https://www.youtube.com/watch?v=TTUrtCY2k-w

TPM

Gene ID	Length	Rep1	Rep2	Rep3
A	2 kb	10	12	30
B	4 kb	20	25	60
C	1 kb	5	8	15
D	10 kb	0	0	1

Count table

Scale Per Kilobase

Gene ID	Length	Rep1	Rep2	Rep3
A	2 kb	5	6	15
B	4 kb	5	6.25	15
C	1 kb	5	8	15
D	10 kb	0	0	0.1

RPK table

https://www.youtube.com/watch?v=TTUrtCY2k-w

TPM

Gene ID	Length	Rep1	Rep2	Rep3
A	2 kb	5	6	15
B	4 kb	5	6.25	15
C	1 kb	5	8	15
D	10 kb	0	0	0.1

RPK table

Total reads:	15	20.25	45.1
Tens of reads:	1.5	2.025	4.51

"Per Million" scaling factors → Tens of reads:

In this example, we scale the total read counts by 10 instead of 1,000,000

https://www.youtube.com/watch?v=TTUrtCY2k-w

TPM

Gene ID	Length	Rep1	Rep2	Rep3
A	2 kb	5	6	15
B	4 kb	5	6.25	15
C	1 kb	5	8	15
D	10 kb	0	0	0.1

RPK table

Total reads:	15	20.25	45.1
Tens of reads:	1.5	2.025	4.51

"Per Million" scaling factors → Tens of reads:

Gene ID	Length	Rep1	Rep2	Rep3
A	2 kb	3.33	2.96	3.326
B	4 kb	3.33	3.09	3.326
C	1 kb	3.33	3.95	3.326
D	10 kb	0	0	0.02

TPM table

https://www.youtube.com/watch?v=TTUrtCY2k-w

TPM summary

Gene ID	Length	Rep1	Rep2	Rep3
A	2 kb	10	12	30
B	4 kb	20	25	60
C	1 kb	5	8	15
D	10 kb	0	0	1

Count table

Read count was:

1) Normalized for **gene length**
2) Normalized for **differences in sequencing depth**

Gene ID	Length	Rep1	Rep2	Rep3
A	2 kb	3.33	2.96	3.326
B	4 kb	3.33	3.09	3.326
C	1 kb	3.33	3.09	3.326
D	10 kb	0	0	0.02

TPM table

https://www.youtube.com/watch?v=TTUrtCY2k-w

RPKM vs TPM

Gene ID	Length	Rep1	Rep2	Rep3
A	2 kb	1.43	1.33	1.42
B	4 kb	1.43	1.39	1.42
C	1 kb	1.43	1.78	1.42
D	10 kb	0	0	0.009
	RPKM total:	4.29	4.5	4.25

RPKM table

		TPM total:	10	10	10

Gene ID	Length	Rep1	Rep2	Rep3
A	2 kb	3.33	2.96	3.326
B	4 kb	3.33	3.09	3.326
C	1 kb	3.33	3.09	3.326
D	10 kb	0	0	0.02

TPM table

https://www.youtube.com/watch?v=TTUrtCY2k-w

RPKM and TPM

- It is generally acceptable to use RPKM and TPM for within-sample transcript expression comparison
- Both RPKM and TPM are NOT suggested to be directly used for cross-sample transcript expression comparison

Zhao S., Ye Z. and Stanton R. Misuse RPKM or TPM normalization when comparing cross samples and sequencing protocols. RNA journal. 2020

Pseudoalignment tool – kallisto

- Does not require alignment to a reference genome (super fast)
- Uses gene trascripts (reference transcriptome)
- Quantification at transcript level

Nicolas L Bray, Harold Pimentel, Páll Melsted and Lior Pachter, Near-optimal probabilistic RNA-seq quantification, Nature Biotechnology 34, 525–527 (2016), doi:10.1038/nbt.3519

K-mer

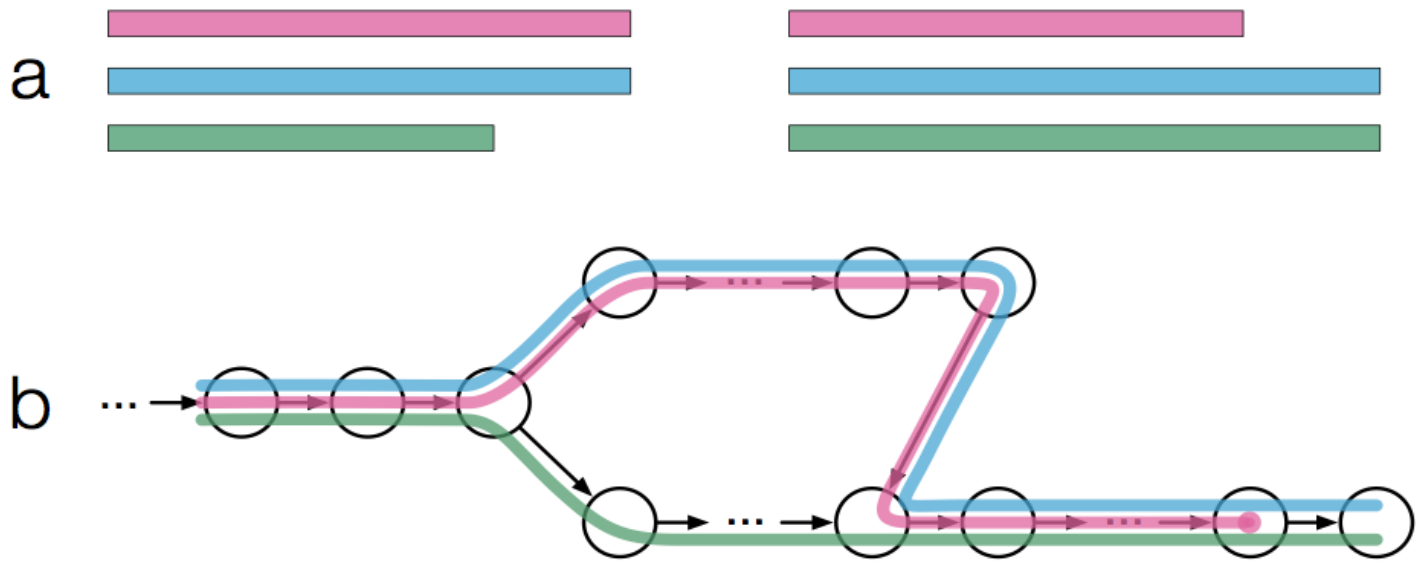
sequence **ATGGAAGTCGCGGAATC**

7mers

```
ATGGAAG
 TGGAAGT
  GGAAGTC
   GAAGTCG
    AAGTCGC
     AGTCGCG
      GTCGCGG
       TCGCGGA
        CGCGGAA
         GCGGAAT
          CGGAATC
```

Compare the compatibility of k-mers in short reads and target transcripts

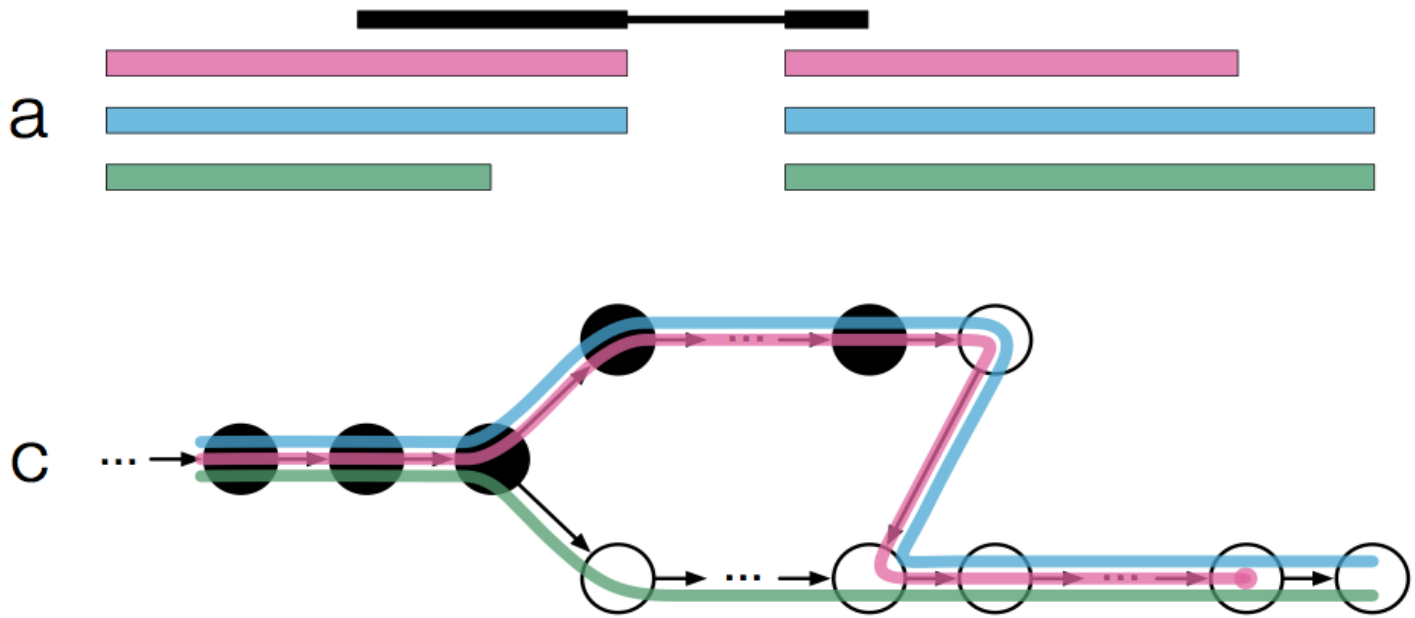
Pseudoalignment tool – kallisto



- Construct a target de-Bruijn graph (t-DBG) from the target transcripts
- Each node is a k-mer in the t-DBG and is associated with a transcript or set of transcripts, named as a k-compatibility class

https://tinyheero.github.io/2015/09/02/pseudoalignments-kallisto.html

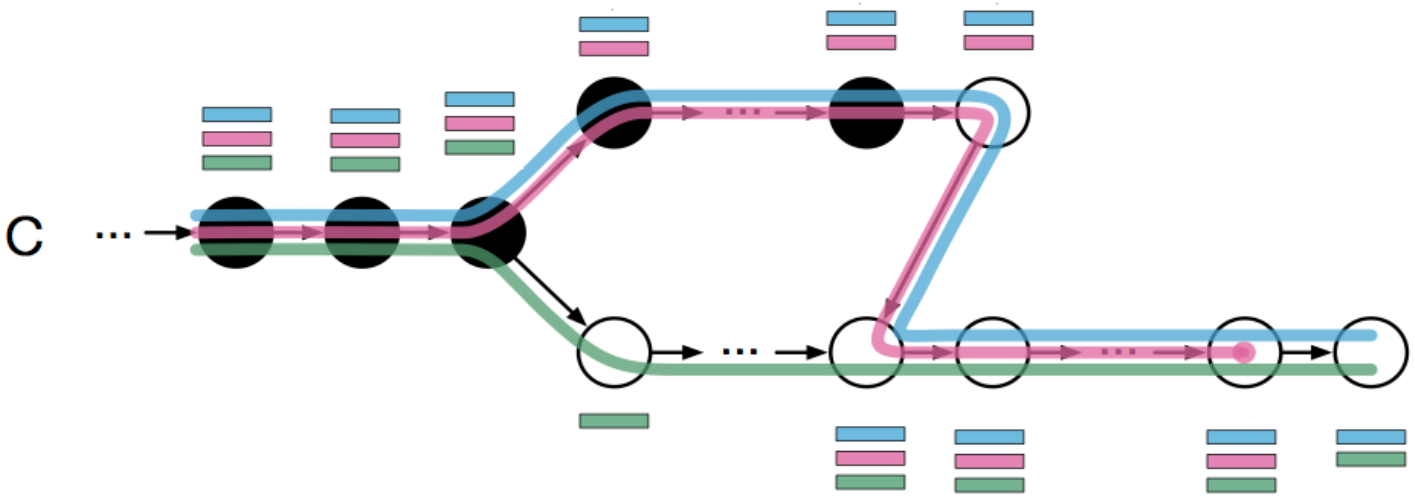
Pseudoalignment tool – kallisto



Evalute k-mers of short reads for compatibility with the t-DBG

https://tinyheero.github.io/2015/09/02/pseudoalignments-kallisto.html

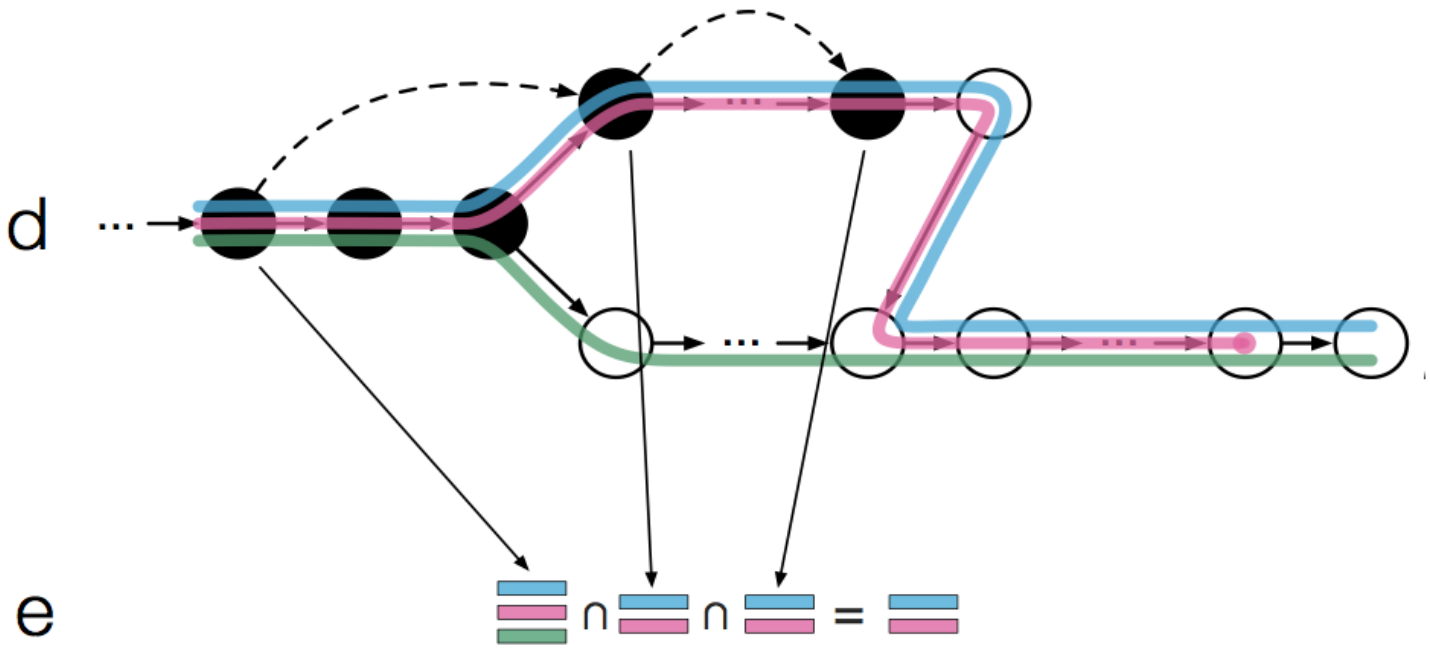
Pseudoalignment tool – kallisto



Find the transcript(s) a read is compatible with based on the k-compatibility class of a read

https://tinyheero.github.io/2015/09/02/pseudoalignments-kallisto.html

Pseudoalignment tool – kallisto



- Looks up the k-compatibility class of the node and then "skips" to the node that is after the last node in the same equivalence class
- Intersect the k-compatibility classes on the "non-skipped" nodes

https://tinyheero.github.io/2015/09/02/pseudoalignments-kallisto.html

Thank you!