Alignment tools for RNA-seq data

Irina Pulyakhina
2nd year PhD student
Department of Human Genetics
Leiden University Medical Center
DNA to RNA path

DNA → transcription → pre-mRNA → splicing → mature mRNA

reference transcript
DNA to RNA path

DNA → pre-mRNA → mature mRNA

- transcription
- splicing
**DNA to RNA path**

1. DNA
2. Transcription
3. Pre-mRNA
4. Splicing
5. Mature mRNA
   - Reference transcript
   - Exon skip
   - Partial intron retention
Alignment (1)

ATACCAGCATCCCTGGTCACGATACTACTCCCAGACTGCTACTACGACGATAC ref. seq.

TTGCTACTACTA query (read)
**Alignment (2)**

ATACCAGCTACCCATGGTCACGATACTACTACTCCCGATCTGCTACTACTACGATAC
TTGCTACTACTA

(70%)
Alignment (3)

ATACCCAGCTACCA TGGTCA CGATA CTACTACTCC CAGATTGCTACTACTAC TGCTTACTACTA (70%)

(100%)
Aligning RNA to the reference

pre-mRNA reads

standard DNA aligner
Aligning RNA to the reference

pre-mRNA reads

standard DNA aligner

mature mRNA reads

standard DNA aligner will not work:
Aligning RNA to the reference

pre-mRNA reads

mature mRNA reads

standard DNA aligner will not work:
Aligning RNA to the reference

Why genome? Why not transcriptome?
Vs genome or vs transcriptome?

Genome:
+ complete information;
+ if you only have transcriptome/transcriptome assembly;
– problem with exon-exon boundaries.
Vs genome or vs transcriptome?

Genome:
+ complete information;
+ if you only have transcriptome/transcriptome assembly;
– problem with exon-exon boundaries.

Transcriptome:
– incomplete information about only the transcriptomes that have been previously detected;
– problem of multiple mappings to alternative isoforms;
+ BUT you don't have to think about exon-exon junctions anymore.
– Map RNA reads to DNA reference.
– Find exon-exon junctions.
Main groups of RNA alignment tools

- Splitting both reference and reads (hash tables)
- Seeding only read and looking for a good match in the whole length reference
- Seed-and-extend: looking for a partial match and trying to expand
Creating index hash tables of both ref.seq. and reads:

– GMAP
– supersplat
Creating index hash tables of both ref.seq. and reads:

- GMAP
- superssplat
Creating index hash tables of both ref.seq. and reads:

- GMAP
- supersplat
Creating index hash tables of both ref.seq. and reads:

- GMAP
- supersplat
2. seeding only reads (1)

HMMSplicer
2. seeding only reads (1)

**HMMSplicer**

training HMM on part of the mappings
2. seeding only reads (1)

HMMSplicer

training HMM on part of the mappings
2. seeding only reads (1)

HMMSplicer

training HMM on part of the mappings
MapSplice

2. seeding only reads (2)

2.1 Seeding only reads (2)

Double anchored exonic

Single anchored

Double anchored

Single anchored
2. seeding only reads (3)

TopHat

potential exon area
2. seeding only reads (3)

TopHat

- potential exon area
- a microintron or a lowly covered exonic region
- coverage islands
TopHat

2. seeding only reads (3)

- potential exon area
- a microintron or a lowly covered exonic region
- coverage islands
2. seeding only reads (3)

TopHat

- potential exon area
- a microintron or a lowly covered exonic region
- coverage islands
2. seeding only reads (4)

**TopHat** first tries to map full length reads; splits only the unmapped reads.
3. seed-and-extend (1)

PASSion

__________________________________________
3. seed-and-extend (1)

PASSion
3. seed-and-extend (1)

PASSion
PASSion
PASSion first tries to map full length reads; splits only the unmapped reads.
Common problems for RNAseq aligners (1)
Low coverage regions:
Common problems for RNAseq aligners (1)

Low coverage regions:

- microintron

How it should be:

What we will have:
Low coverage regions:

- microintron

How it should be:

What we will have:

- poorly covered exonic region

How it should be:

What we will have:
Multiple mappings:

1. repetitive regions within one molecule;
Multiple mappings:

1. repetitive regions within one molecule;

2. same region on different transcripts;
Multiple mappings:

1. repetitive regions within one molecule;

2. same region on different transcripts;

3. repetitive regions on alternative transcripts.
Solutions for multiple mappings

1. «Remove» repeat maskers (but what if they are 90% of your target?).
2. Take into account «region mapability».
3. Choose unique, probabilistic or multiple mappings.
– pseudogenes (you map your reads to something that is not expressed);
– identification and quantification of alternative transcripts;
– allele-specific SNPs;
– low coverage for targeted sequencing;
– intronic «contamination»;
– again, genome or transcriptome?
– and many more …
Example: exon-intron borders (1)

```
CAAGCAG

GT

CAAGCAG

GTCCTCGA

GSNAP

TopHat
```
Example: exon-intron borders (1)

GSNAP alignment:
Example: exon-intron borders (1)

GSNAP alignment:

C A A G C A G
C A A G C A G

TopHat alignment:

C A A G C A G
C A A G C A G

GSNAP

TopHat
Example: exon-intron borders (2)
Example: exon-intron borders (2)

We filter out alignments that have an overhang < 5 nt.
Tools for RNAseq analysis differ in:
1) alignment algorithm;
2) unique/non-unique mappings;
3) canonical/non-canonical splice sites (GT-AG, GC-AG, AT-AC, anything);
4) annotation of known/unknown exons;
5) dependency on any known exon model for novel exons detection;
6) a multi-thread possibility.
# Groups of RNA seq alignment tools

<table>
<thead>
<tr>
<th></th>
<th>Non-can. sites</th>
<th>Machine-learning approach</th>
<th>Accurate for low-depth</th>
<th>Accurate with indels</th>
<th>Recommendations</th>
</tr>
</thead>
<tbody>
<tr>
<td>MapSplice</td>
<td>+</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>low-complexity data</td>
</tr>
<tr>
<td>SpliceMap</td>
<td>—</td>
<td>—</td>
<td>+</td>
<td>—</td>
<td>long RNA-seq read lengths</td>
</tr>
<tr>
<td>TopHat</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>lots of different output files; good for beginners</td>
</tr>
<tr>
<td>MapNext</td>
<td>+</td>
<td>—</td>
<td>+</td>
<td>—</td>
<td>complex splicing events</td>
</tr>
<tr>
<td>GSNAP</td>
<td>+</td>
<td>—</td>
<td>+</td>
<td>+</td>
<td>abundant intronic material</td>
</tr>
<tr>
<td>supersplat</td>
<td>+</td>
<td>—</td>
<td>+</td>
<td>—</td>
<td>complex splicing events</td>
</tr>
<tr>
<td>PASSion</td>
<td>+</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>no test</td>
</tr>
<tr>
<td>QPalma</td>
<td>+</td>
<td>+</td>
<td>—</td>
<td>—</td>
<td>gene expression levels</td>
</tr>
<tr>
<td>SOAPsplice</td>
<td>+</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>low depth of sequencing</td>
</tr>
</tbody>
</table>

https://wiki.nbic.nl/index.php/NGS_RNA-Seq
1. Different alignment tools make different choices. Think about its application to your experiment.

2. Make a proper choice between genome and transcriptome (again, think for you specific case).

3. Choose unique/non-unique mappings depending on the experiment.
Acknowledgement

Leiden University Medical Center, Department of Human Genetics:

Peter-Bram 't Hoen
Jeroen Laros
Michel Villerius