Mapping short RNA-Seq by comparing tree

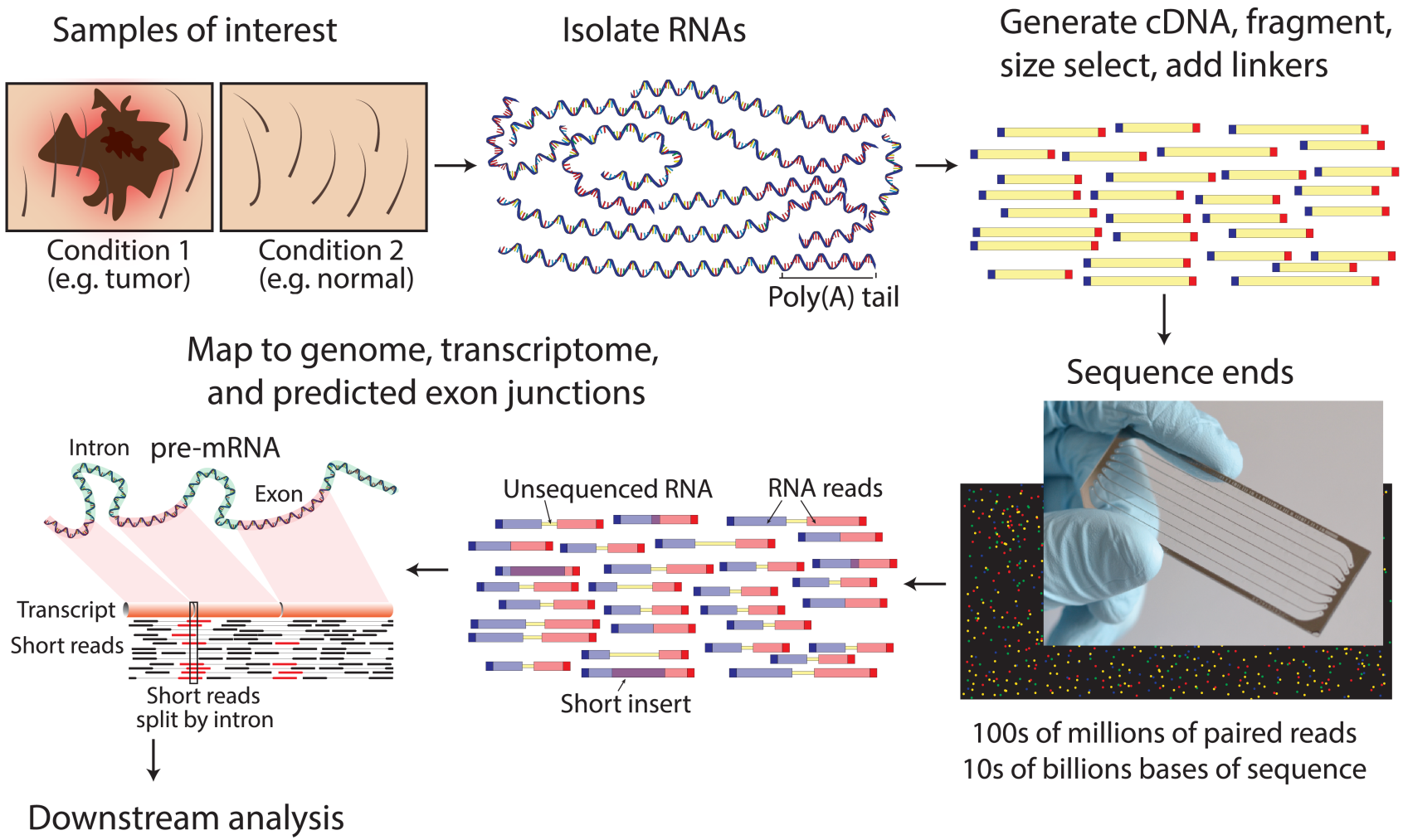
Work in progress
Possibly useless

Matthias Zytnicki

INRAE, MIAT

DSB 2020

RNA-Seq



Griffiths et al., PLOS Comp. Biol., 2015

Mapping

Definition

Prediction of the *locus* which produced the RNA *read*.

Read	Genome
ACGT	CATCAGTCTAGACGTTCACAACCA

$\Rightarrow$ chr1:12–15

Tricky situations

- Reads may be slightly different from the genome sequence.

Read	Genome
ACGT	CATCAGTCTAGACGGTCACAACCA

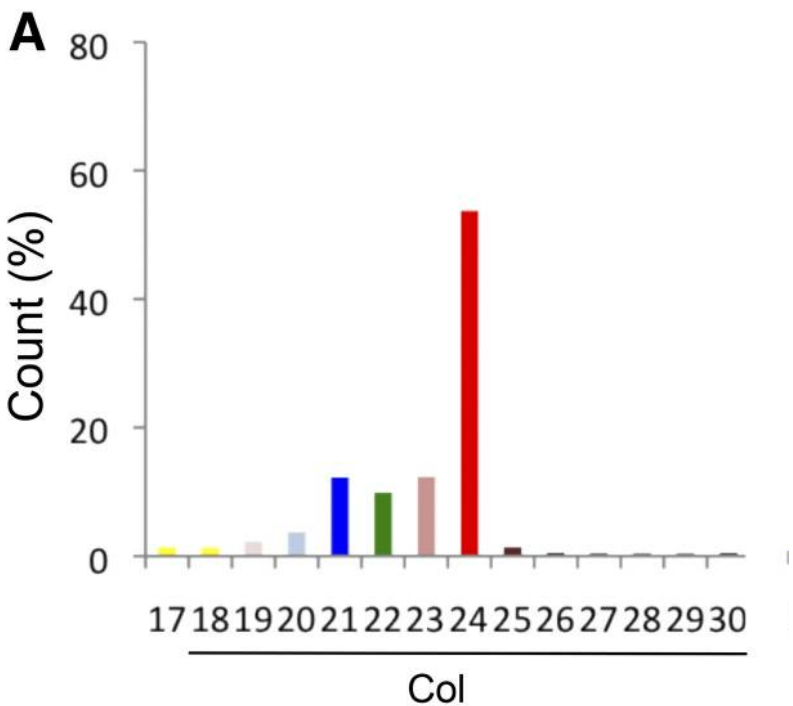
- Corresponding loci are repeated.

Read	Genome
ACGT	ACATACGTTCACACGTCGAT

Our question

Particularities of sRNA-Seq

- A population of different classes of small RNAs: miRNAs, tRFs, siRNAs, piRNAs, etc.
- They are short (about 22–24bp, after trimming).
- Sequences are highly duplicated (∼5% the exact same read).
- Most mismatches happen at the ends of the reads.



ID	Accession	RPM	Chromosome	Start	End	Strand
ath-MIR156a	MI0000178	-	chr2	10676451	10676573	-
ath-MIR156b	MI0000179	-	chr4	15074899	15075081	+
ath-MIR156c	MI0000180	-	chr4	15415418	15415521	-
ath-MIR156d	MI0000181	-	chr5	3456632	3456749	-
ath-MIR156e	MI0000182	-	chr5	3867207	3867313	+
ath-MIR156f	MI0000183	-	chr5	9136106	9136237	+
ath-MIR157a	MI0000184	-	chr1	24913202	24913299	-
ath-MIR157b	MI0000185	-	chr1	24921086	24921217	+
ath-MIR157c	MI0000186	-	chr3	6244500	6244716	-
ath-MIR157d	MI0000187	-	chr1	18026811	18027031	-

from miRBase

Our question — Cont.

Observation

- Most mapping tool developments are dedicated to long reads.
- There is no dedicated tool for sRNAs.

Usual (biological) query

For each read, get me *all* the regions with *minimum* number of mismatches n, with $n \leq k$.

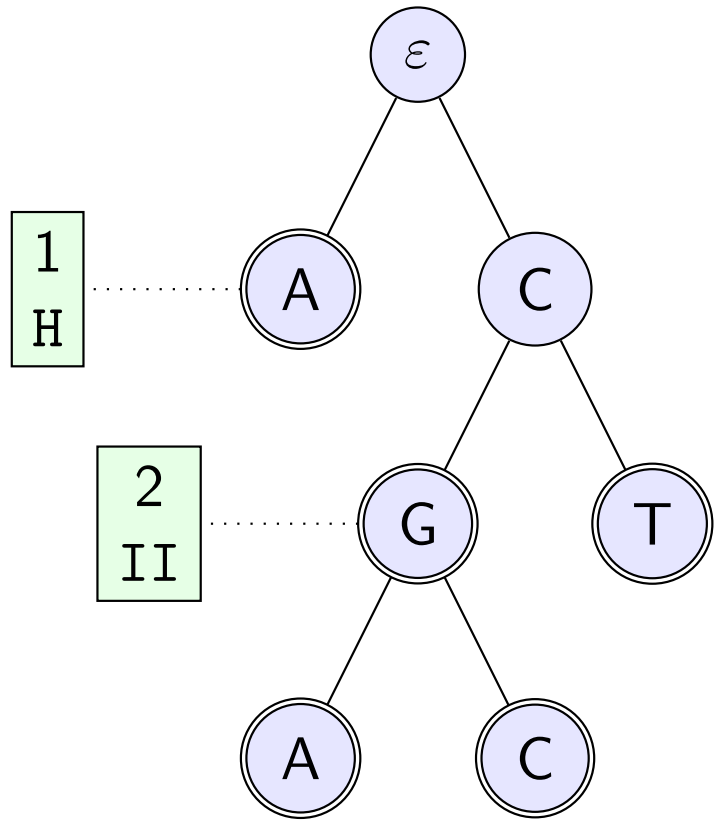
Data

Reads

- Stored in a tree.
- Counts, and best quality is kept.

```
@read1
A
+
H
@read2
CG
+
HI
@read3
CG
+
IH
```

```
@read4
CGA
+
HHI
@read5
CGC
+
IIH
@read6
CT
+
II
```



Data

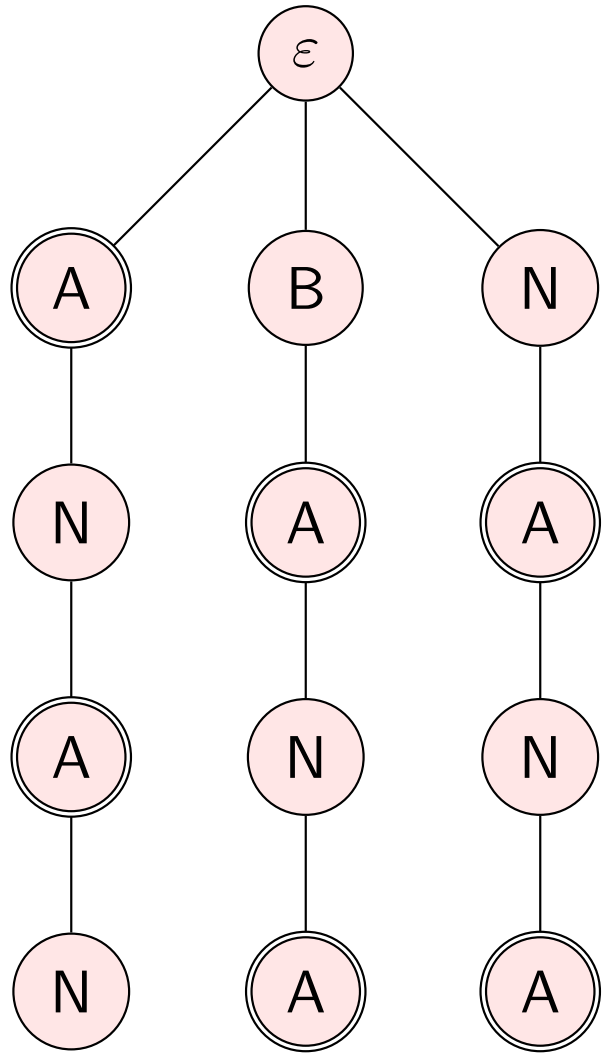
Genome

- Stored in a suffix array.
- Using BWA implementation.

Example

`BANANA`

Suffix tree



Data

Genome

- Stored in a suffix array.
- Using BWA implementation.

Example

BANANA

List of suffixes

BANANA
ANANA
NANA
ANA
NA
A

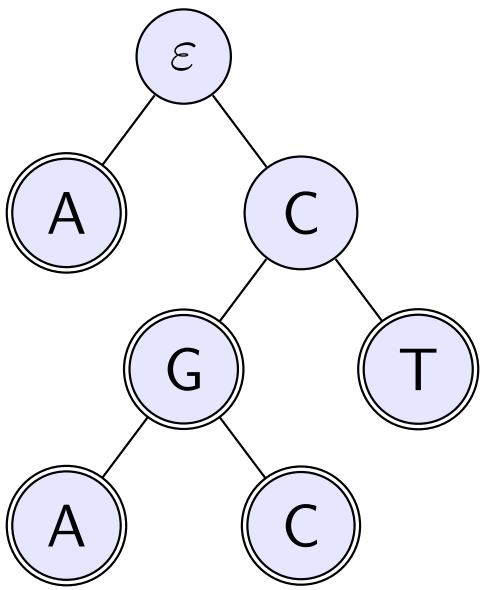
Suffix array

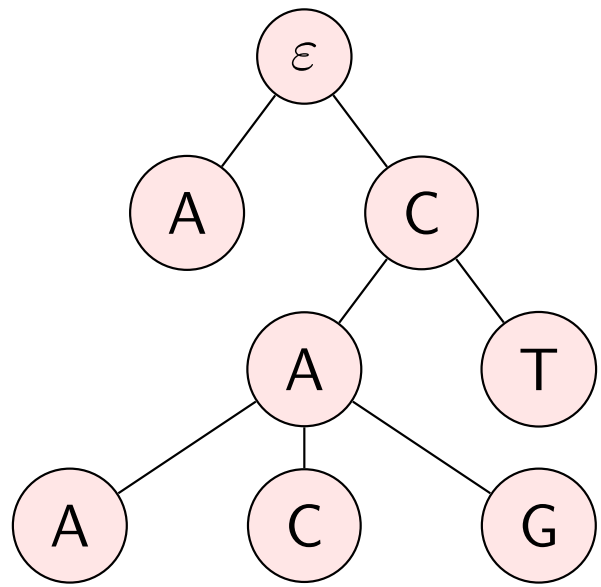
5	A
3	ANA
1	ANANA
0	BANANA
4	NA
2	NANA

Main idea

Aim

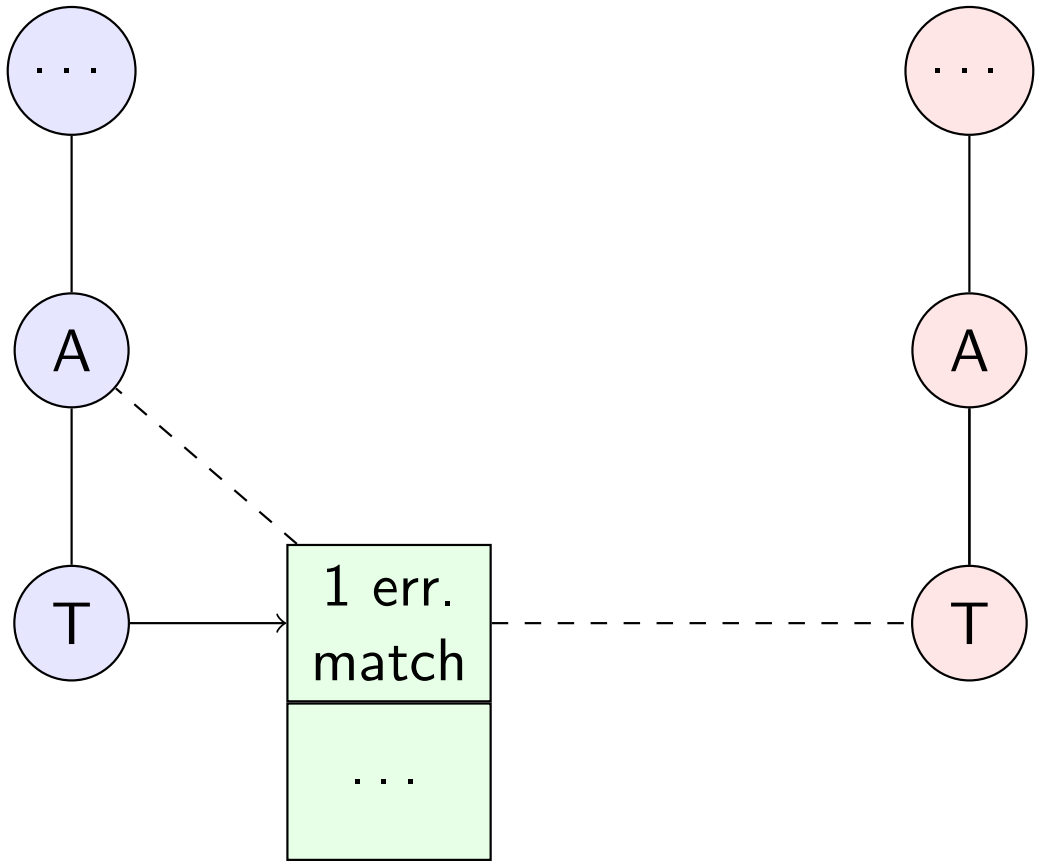
- For each accepting "read node," compute the all the "genome nodes" with minimum distance not greater than k.
- For each "reads node," compute recursively the all the "genome nodes" with distance not greater than k.





Note: The genome tree here is not an actual suffix tree. It is just presented as an illustration.

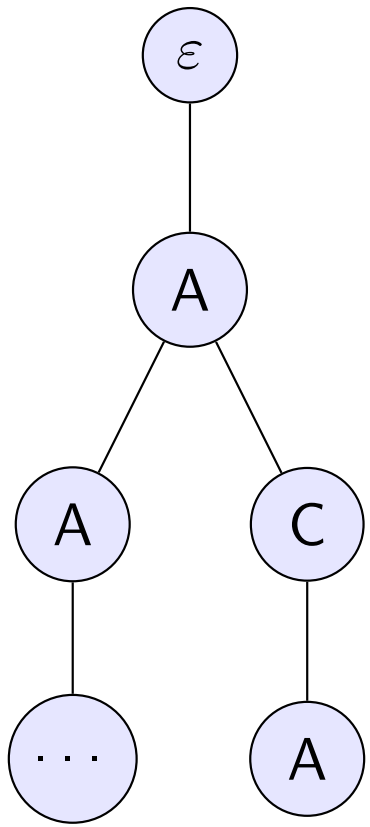
Implementation



Optimization 1

Expect a 0-error mapping first

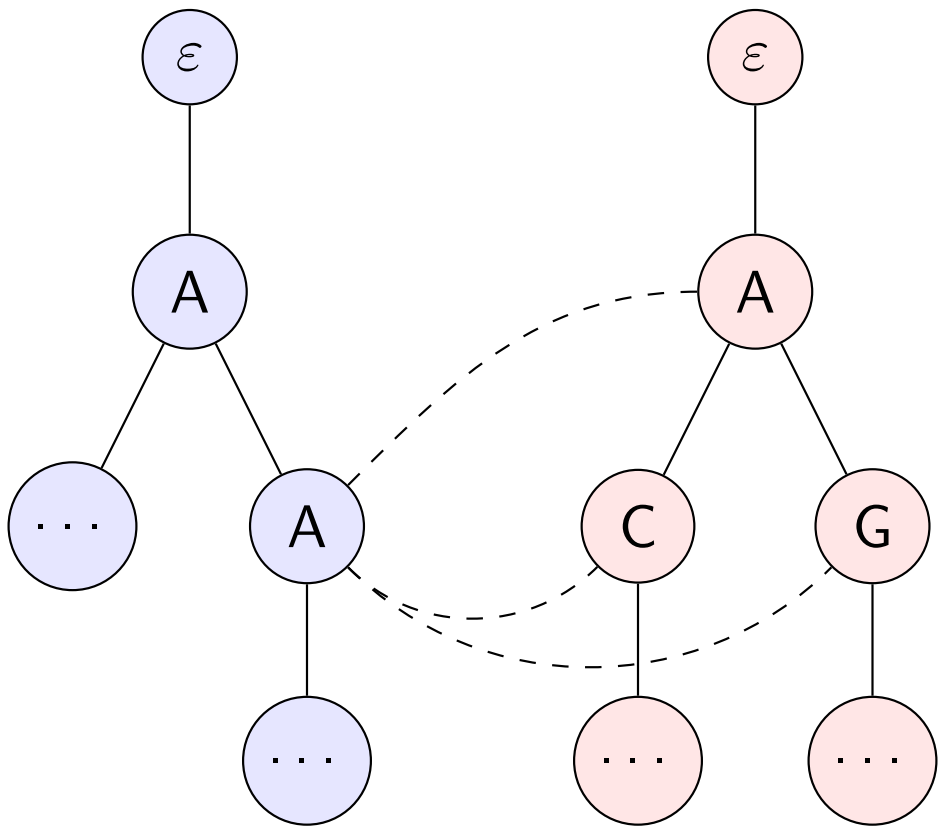
- Map with no error first.
- In case of error at depth d, add an error up to depth d.



Optimization 2

Map the unbranched regions "the usual way"

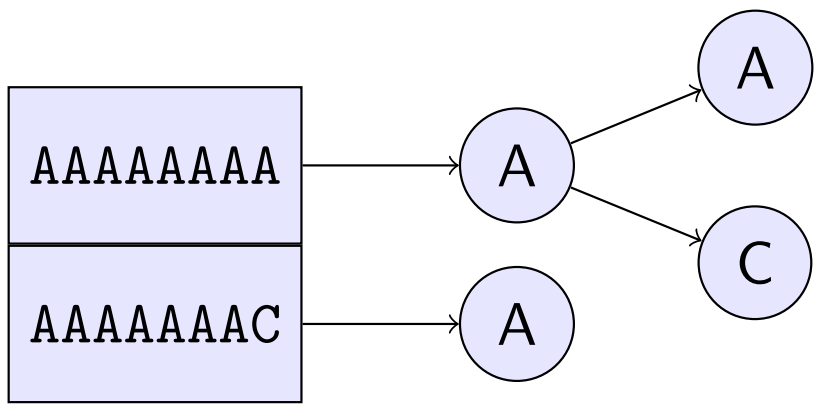
When a read unbranched terminal path is found, gather all the corresponding genome sequences, and apply a banded Smith-Waterman up to the leaves.



Optimization 3

The genome tree is a vector of 4^8 trees

- The first tree is labelled `AAAAAAAA`.
- The second tree is labelled `AAAAAAAC`.
- etc.
- Each tree starts at depth 8.



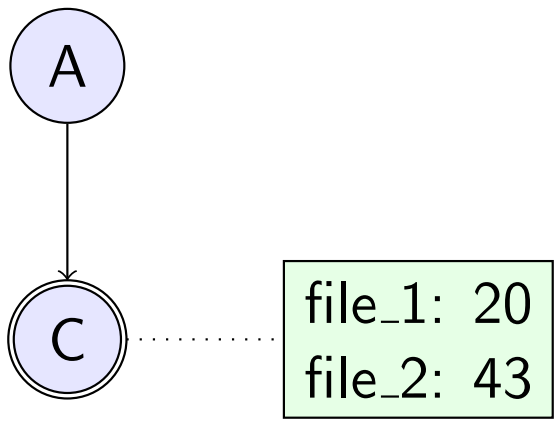
Other optimizations

Remove low complexity reads

`ACACACACACA`

Use radix tree instead of standard tree for the reads tree

Can process several reads files



Results

Test case

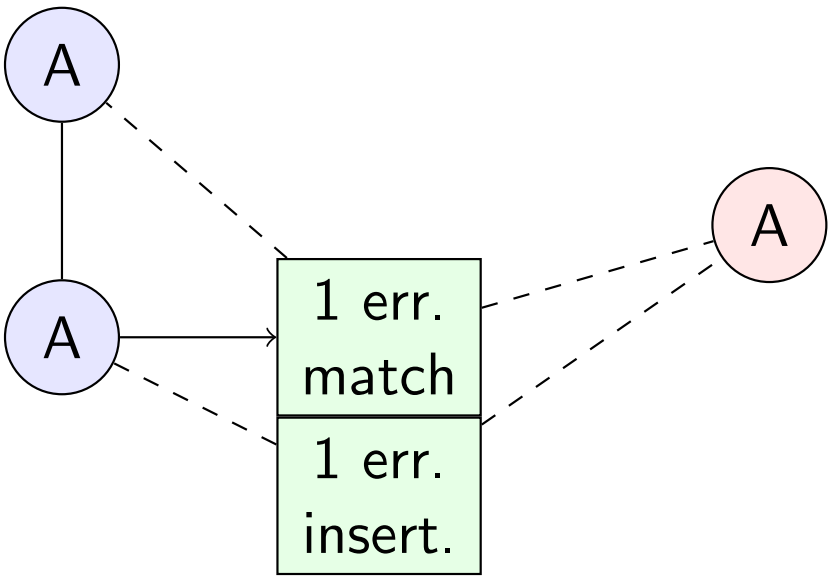
- 15,492,953 reads of size 15–101.
- Genome: *A. thaliana*.
- BWA aln: 14min, 221kB.
- srnaMapper: 6min, 1.6GB.

Bottleneck

% time	cumulative seconds	self seconds	calls	name
47.53	161.27	161.27		bwt_2occ
26.24	250.28	89.01		bwt_occ
9.92	283.93	33.65	43390524	mapWithoutError

Problem

states increase



Compare to dynamic programming

	ε		A		A
ε	0	$\rightarrow$	1	$\rightarrow$	2
	$\downarrow$	$\searrow$		$\searrow$	
A	1		0	$\rightarrow$	1

Bottom line

- You do not want *all* the mappings.
- How to implement a good # states *vs* states elimination balance?

Implementation details — Reads

First pass

- Edges contain the nucleotides (and the size), and the address to the following node.
- No predefined order.
- Each node contains 4 edges, the read counts, and the qualities.

Second pass

- Nodes are sorted in a depth-first fashion.
- Read counts and qualities are stored in a parallel vector.

Implementation details — Rest

Genome

- Tree: the BWA structure.
- Buffer: last children intervals are kept in memory.

Smith-Waterman

A (stupid) read length$\times(2k+1)$ matrix.

Next

- Clever way to reduce the number of states.
- Bug fixes (read mapping at the ends of a chromosome...).
- Other optimizations (branch sequences in an external string?).
- Use several processors.
- Available at `https://github.com/mzytnicki/srnaMapper` (branch `sw`).

That's all, folks!

Thank you for your attention!