ELECTOR: Evaluator for long reads correction methods

Camille Marchet [1], Pierre Morisse [2], Lolita Lecompte [3], Antoine Limasset [1], Arnaud Lefebvre [2], Thierry Lecroq [2], Pierre Peterlongo [3]

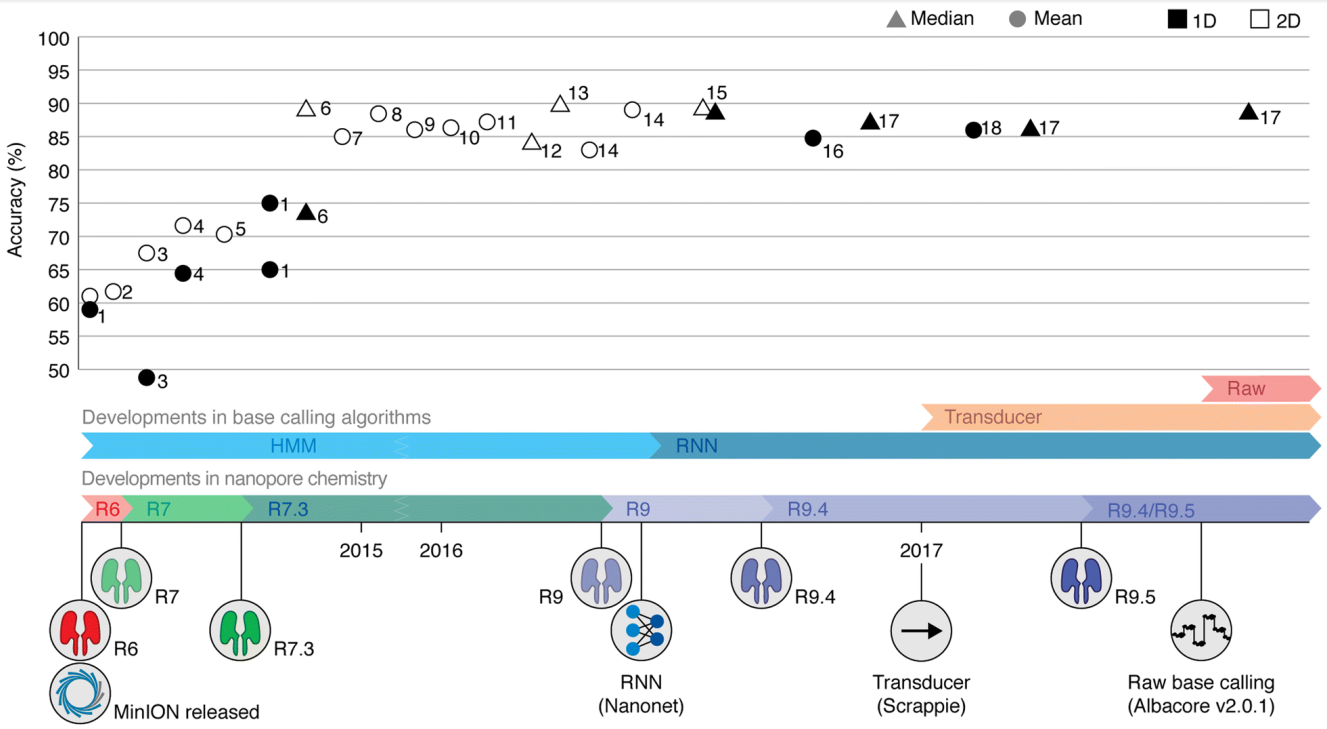
[1] Univ. Lille, CNRS, Inria, UMR 9189 - CRIStAL.
[2] Normandie Univ, UNIROUEN, LITIS, Rouen 76000, France.
[3] Univ Rennes, CNRS, Inria, IRISA - UMR 6074, F-35000 Rennes, France.

SeqBio 2018

Introduction: errors in long reads

Context

- Long reads : fast evolving field but error rates remain high
- Need quality for assembly, variant calling, ...



From [Rang et al. 2018]

Introduction: correction assessment

Ever-increasing list of correction methods:

- 2012: 3
- 2013: 1
- 2014: 3
- 2015: 2
- 2016: 4
- 2017: 7
- 2018: 3

"Which tool better performs on my problem ?"

A lost bioinformatician

"My corrector works on this ATTAGATTAC toy example so it should do the job."

Pierre M., anonymous overly confident developer

"Let's do **something**!"

C3G MASTODONS long read correction group

Introduction: correction assessment

SOTA

- Only one tool (LRCstats [La et al. 2017])
- Rather slow
- Number of metrics displayed could be increased

Correction quality assessment objectives

- Handle most of the correctors
- Quick (time $\simeq$ correction step's time)
- Scalable
- Reproducible
- Easy to include in benchmarks
- Information for users and developers

Introduction : long reads correction methods

Hybrid

- Mapping short reads/assembled short reads on long reads
- Map LR on paths of graph of short reads

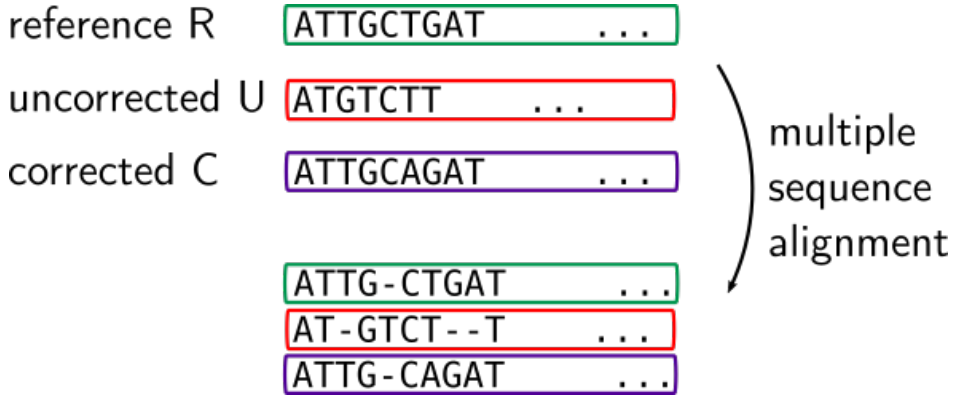
Self

- Produce consensus from LR by multiple mapping on a template LR
- Map LR on paths of graph of LR
- Produce consensus from LR using graphs built from the reads' k-mers

Corrected reads

- Can be missing
- Can be trimmed (shorter than the original)
- Can be split (separated in several corrected fragments)
- Can be elongated (longer on left or right end by bringing some context of the graph)

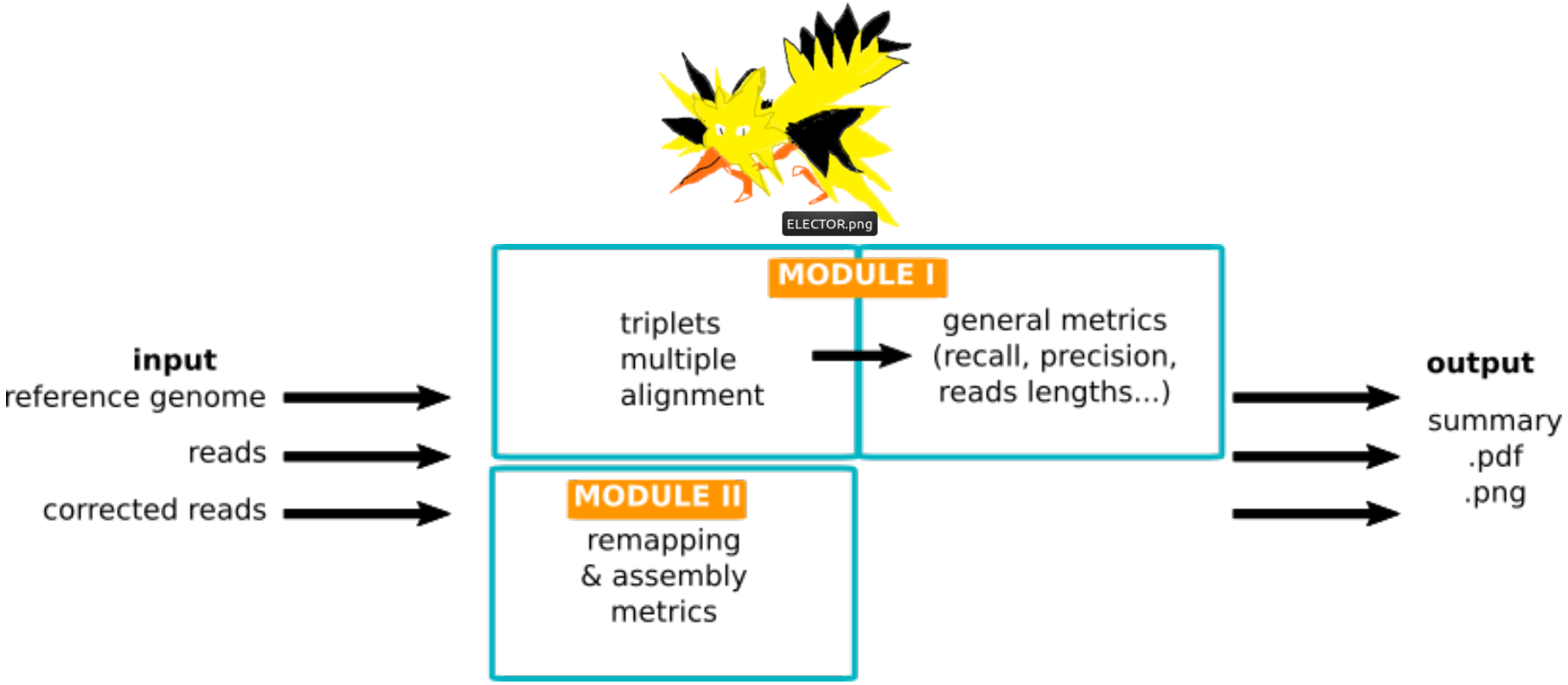
Main idea: compare different versions of a read



Multiple sequence alignment of triplets

- advantages: access recall/precision
- difficulty: scaling
- solution: MSA segmentation

ELECTOR: Overview



Main contributions of ELECTOR w.r.t. LRCstats

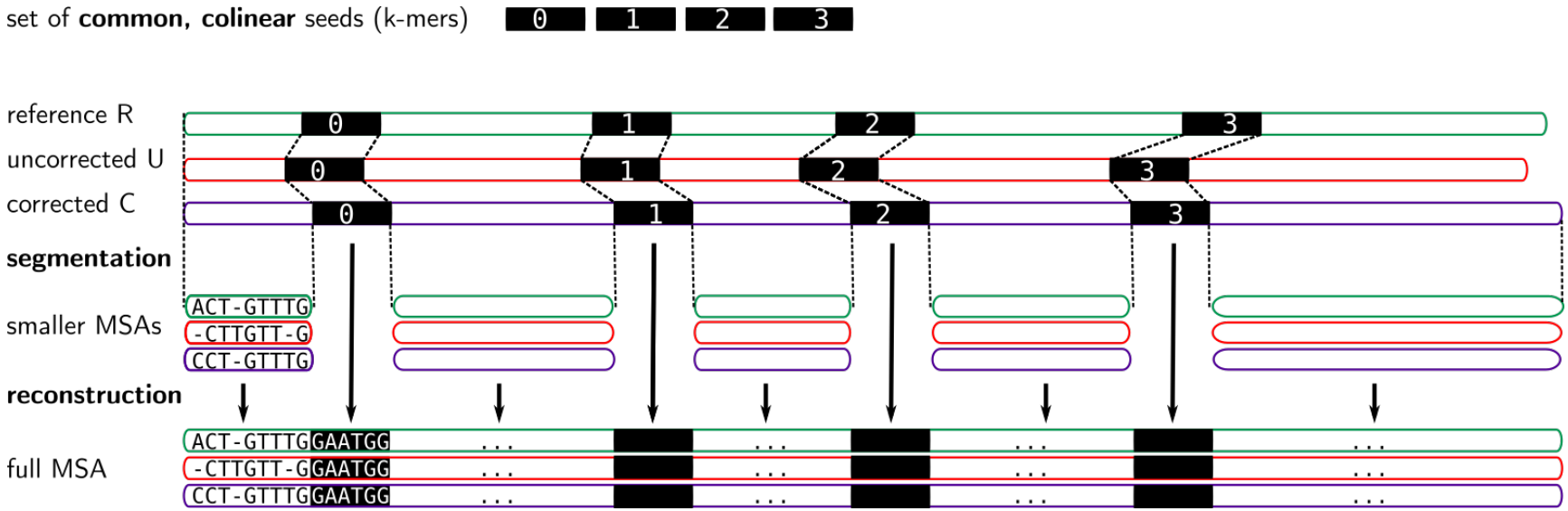
	ELECTOR	LRCstats
error rate	✔	✔
recall	✔	✖
precision	✔	✖
deletions	✔	✔
insertions	✔	✔
substitutions	✔	✔
split reads	✔	✔
mean missing size	✔	✖
%GC before/after correction	✔	✖
ratio correction in homopolymers	✔	✖
remapping stats	✔	✖
assembly stats	✔	✖

+ **decreased running time**

Segmented multiple sequence alignment

MSA segmentation

- Same idea as Pierre's talk (LoRSCo)
- For triplet of sequences
- Alignment method: POA [Lee et al. 2002]
- Added feature: handle large gaps



Issue with large gaps

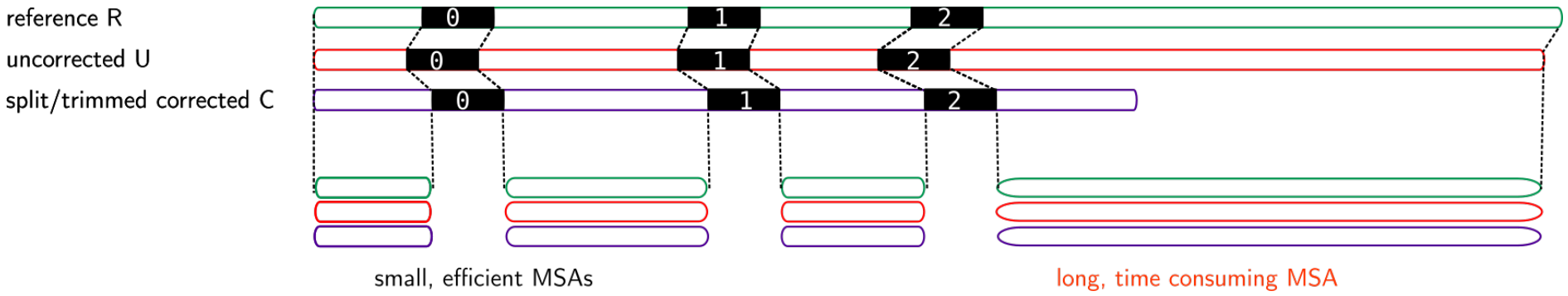
Segmentation MSA rules

Mainly for efficiency:

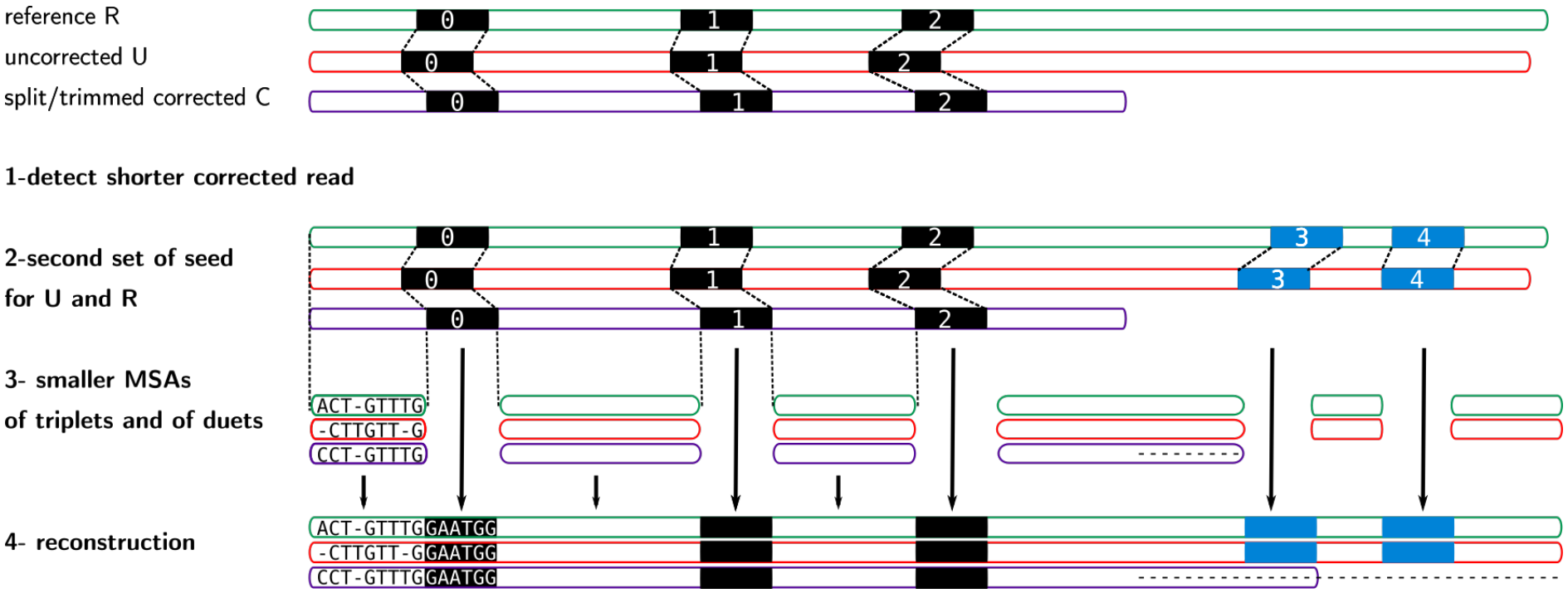
1. If a corrected read is extremely short: do not align, report
2. If the set of seeds is very small (corrected and reference are very dissimilar): do not align, report

In both cases we cannot segment and would have to perform the regular MSA: too long

Issue with trimmed/split reads



Handling large gaps



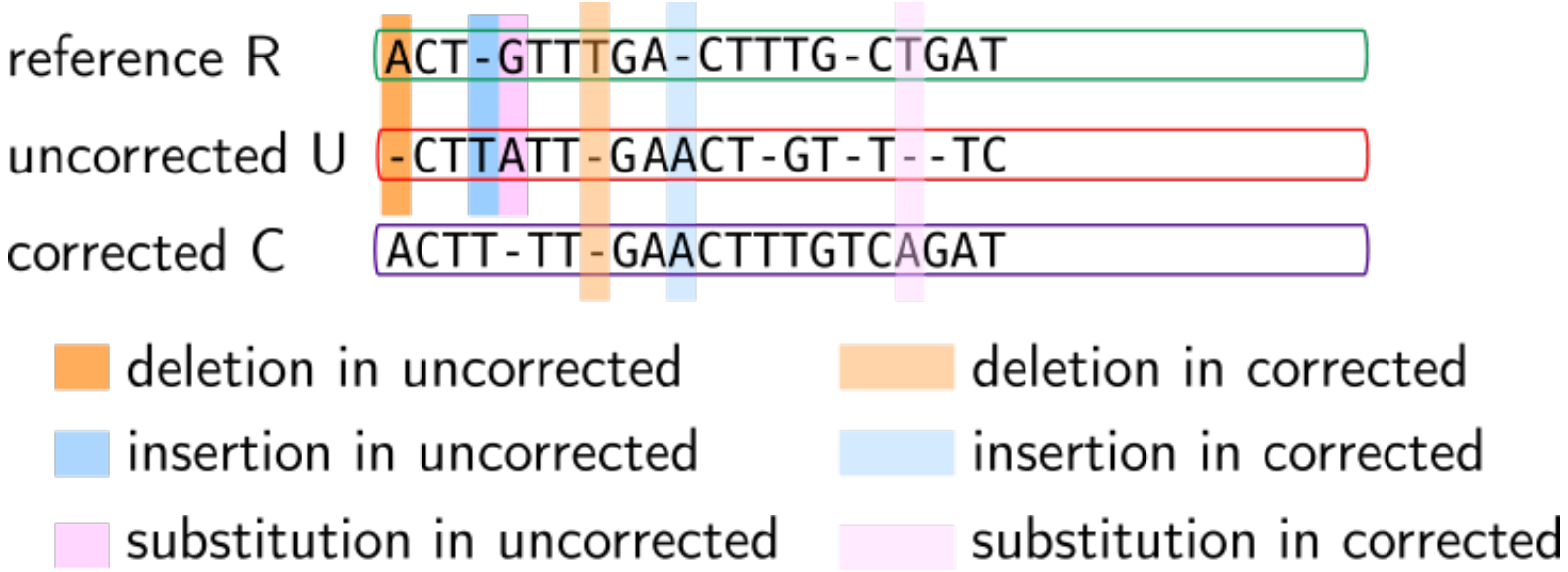
Validation of MSA segmentation

- Simulated datasets from *E. coli*
- "1k" experiment: 1k mean length, 10% error rate, coverage of 100X
- "10k" experiment: 10k mean length, 15% error rate, coverage of 100X
- Corrected with MECAT

Experiment	Recall	Precision	Correct bases	Time
"1k" MSA	**93**.96 %	**93**.48 %	**97.6**4 %	11h
"1k" segmentation + MSA	**93**.81 %	**93**.51 %	**97.6**3 %	**38min**
"10k" MSA	**84.5**1 %	**88**.35 %	**95.2**9 %	107h
"10k" segmentation + MSA	**84.5**9 %	**88**.28 %	**95.2**5 %	**42min**

- Orders of magnitude speed-up
- Similar metrics values

Metrics computation: indels



Metrics computation: split/trimmed/extended

Trimmed

```
reference R      ACT-GTTTG ...    ATTGTCTGAT     ...
uncorrected U    -CTTGTT-G ...    AT-GTCT--T     ...
corrected C      -----------------ATTGTCAGAT     ...
```

Split

```
reference R      ...    ACT_GTTTG ...            ATTGTCTGAT
uncorrected U    ...    -CTTGTT-G     ...        AT-GTCT--T
corrected C_1    ...    ACT-GTTTG
```

```
reference R      ACT-GTTTG            ...        ATTGTCTGAT
uncorrected U    -CTTGTT-G                ...    AT-GTCT--T
corrected C_2    -----------------------------------ATTGTCAGAT
```

Extended

```
reference R      ----------------ACT-GTTTG  ...          ATTGTCTGAT
uncorrected U    ----------------CTTGTT-G     ...       AT-GTCT--T
corrected C      TCTCTGGTATTAGTTAACT-TTTTG    ...       -TTGTCAGAT
```

Metrics computation: recall/precision

reference R ACT-GTTTGA-CTTTG-CTGAT

uncorrected U GCCTGT-TGGACT--GTCAG-T

corrected C ACTTGTTTGAATTTTGTCAGAT

positions to correct (nt(R) != nt(U))

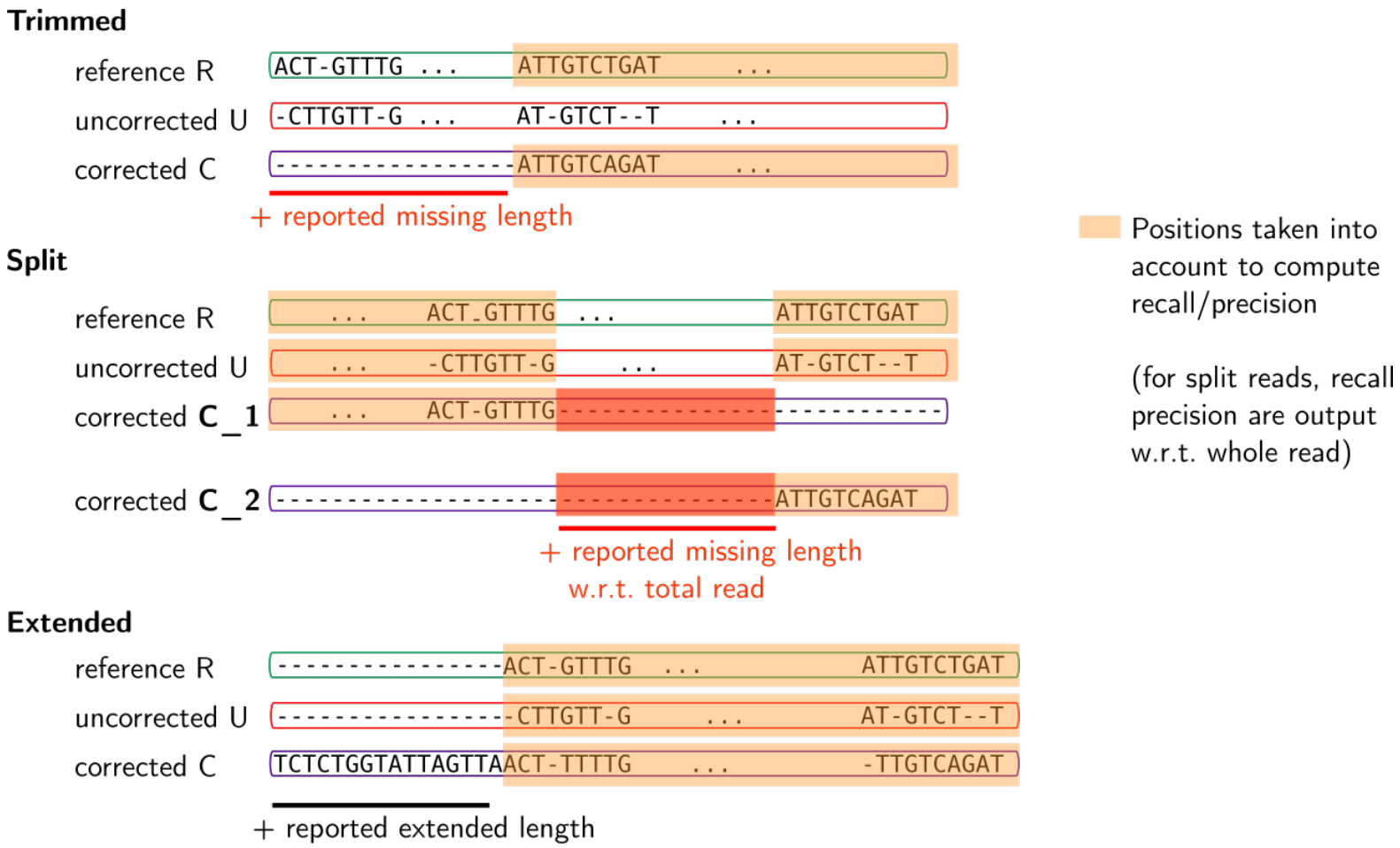
corrected positions (nt(R) != nt(U) OR nt(C) != nt(R))

corrected positions such that nt(C) == nt(R)

corrected positions such that nt(C) != nt(R)

Recall = $\frac{\quad}{\quad}$ **Precision** = $\frac{\quad}{\quad + \quad}$

Metrics computation: recall/precision in modified reads



Validation of MSA for computing metrics

Simulation for ground truth

- Data: 1X and 10X *E. coli*
- Errors: 15% and 20% errors
- Simulated correction

Compare ELECTOR results and ground truth for 10X:

metric	ELECTOR	difference (% ground truth)
recall(%)	98.99	4.0 E-2
precision(%)	99.92	1.0 E-1
error rate	9.920E-2	2.3
indels/mismatches in uncorrected	8380984	4.1
indels/mismatches in corrected	491728	3.4

Results : data sets / correctors

Dataset	*A. baylyi*	*E. coli*	*S. cerevisiae*
Reference organism			
Genome size	3.6 Mbp	4.6 Mbp	12.2 Mbp
Simulated Pacific Biosciences data			
Number of reads	8,765	11,306	30,132
Average length	8,202	8,226	8,204
Number of bases	72 Mbp	93 Mbp	247 Mbp
Coverage	20x	20x	20x
Illumina data			
Source	ERR788913	Genoscope	Genoscope
Coverage	50x	50x	50x

List of correctors

CoLoRMap, HALC, HG-CoLoR, Jabba, LoRDEC, Nanocorr, NaS, Canu, Daccord and LoRMA

Results: running time

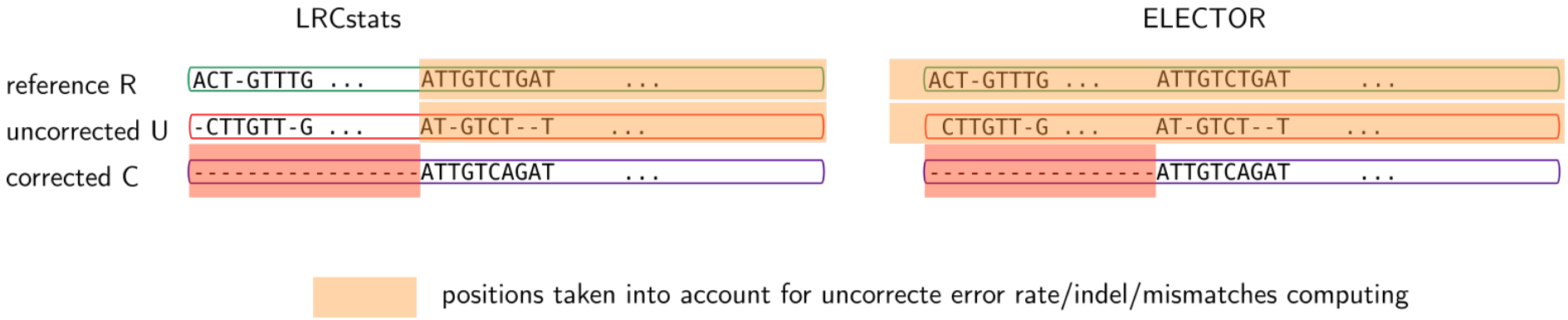
Method	CoLoRMap	Nanocorr	Daccord	Jabba
A. baylyi				
Corrector	57min	2h52min	20min	2min
LRCstats	3h59min	3h44min	3h58min	4h02min
ELECTOR	**1h07min**	**11min**	**5min**	**1h19min**
E. Coli				
Corrector	1h25min	3h17min	27min	2min
LRCstats	4h57min	3h56min	4h20min	5h12min
ELECTOR	**1h21min**	**14min**	**15min**	**32min**
S. cerevisiae				
Corrector	-	-	-	5min
LRCstats	-	-	-	12h01min
ELECTOR	-	-	-	**2h15min**

High speed-up in comparison to LRCstats

Results: comparison to LRCstats

	Nanocorr		daccord	
	ELECTOR	LRCstats	ELECTOR	LRCstats
Error rate	0.339	0.3983	0.422	0.4498
Recall	0.98503	-	0.98836	-
Precision	0.99424	-	0.98468	-
Deletions	46,596	56,708	58,110	72,547
Insertions	237,798	279,970	306,930	336,686
Substitutions	143,605	45,783	72,265	25,643
Trimmed / split reads	1,612	-	123	-
Mean missing size	341	-	3,026	-
Time	**14min**	3h52	**15min**	3h50

Results: comparison to LRCstats



Conclusion & Perspectives

Conclusion

- Fast assessing of a corrector's results
- Many metrics: recall/precision/indels/trimmed/split reads/assembly/remapping...
- A limitation: a reference genome is required
- Innovative developments in segmentation for fast MSA computing

Perspectives

- Results on larger genomes & real data to come
- Support RNA-seq (`https://gitlab.com/leoisl/LR_EC_analyser`)
- Assess variant calling

Availability: `https://github.com/kamimrcht/ELECTOR`

Acknowledgements

- SeqBio committees
- GenScale team
- BONSAI team
- TIBS Team
- C3G MASTODONS

Homopolymer detection