Long-read error correction: a survey and qualitative comparison

Pierre Morisse [1], Arnaud Lefebvre [2], Thierry Lecroq [2]

[1] Normandie Université, UNIROUEN, INSA Rouen, LITIS, 76000 Rouen, France.
[2] Normandie Université, UNIROUEN, LITIS, Rouen 76000, France.

Montpellier JOBIM 2020

UNIVERSITÉ DE ROUEN NORMANDIE

litis

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES ROUEN NORMANDIE

Context

- 2011: Inception of third generation sequencing technologies
- Two main actors: Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT)
- Sequencing of much longer reads, tens of kbps on average
- Expected to solve various problem in the genome assembly field
- Very noisy (10-30% error rates), most errors being indels

Error correction

- Correction: efficient way to handle these errors
- Two approaches:
  1. Hybrid correction (makes use of complementary short reads)
  2. Self-correction (only relies on long reads)

Hybrid correction

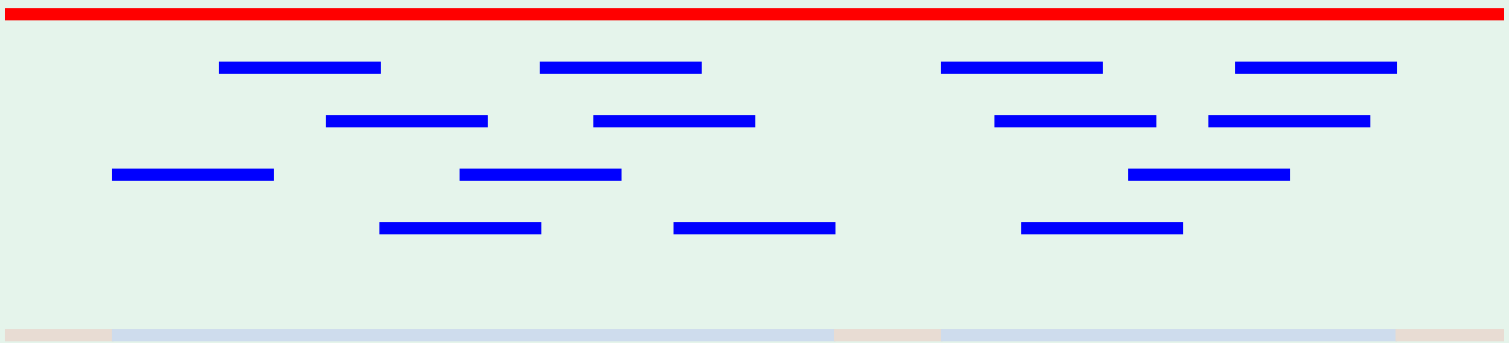
- Long reads + short reads, sequenced for the same individual
- Use the short reads to correct the long reads
- 3 main approaches:
  1. Short reads alignment
  2. Contigs alignement
  3. De Bruijn graphs (DBG)

1) Short reads alignment

Overview

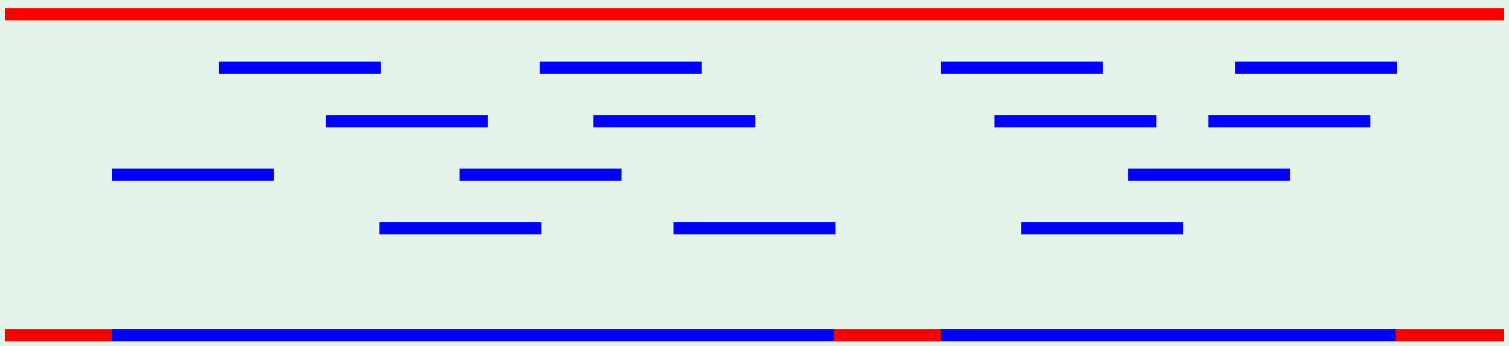
1) Short reads alignment

Overview



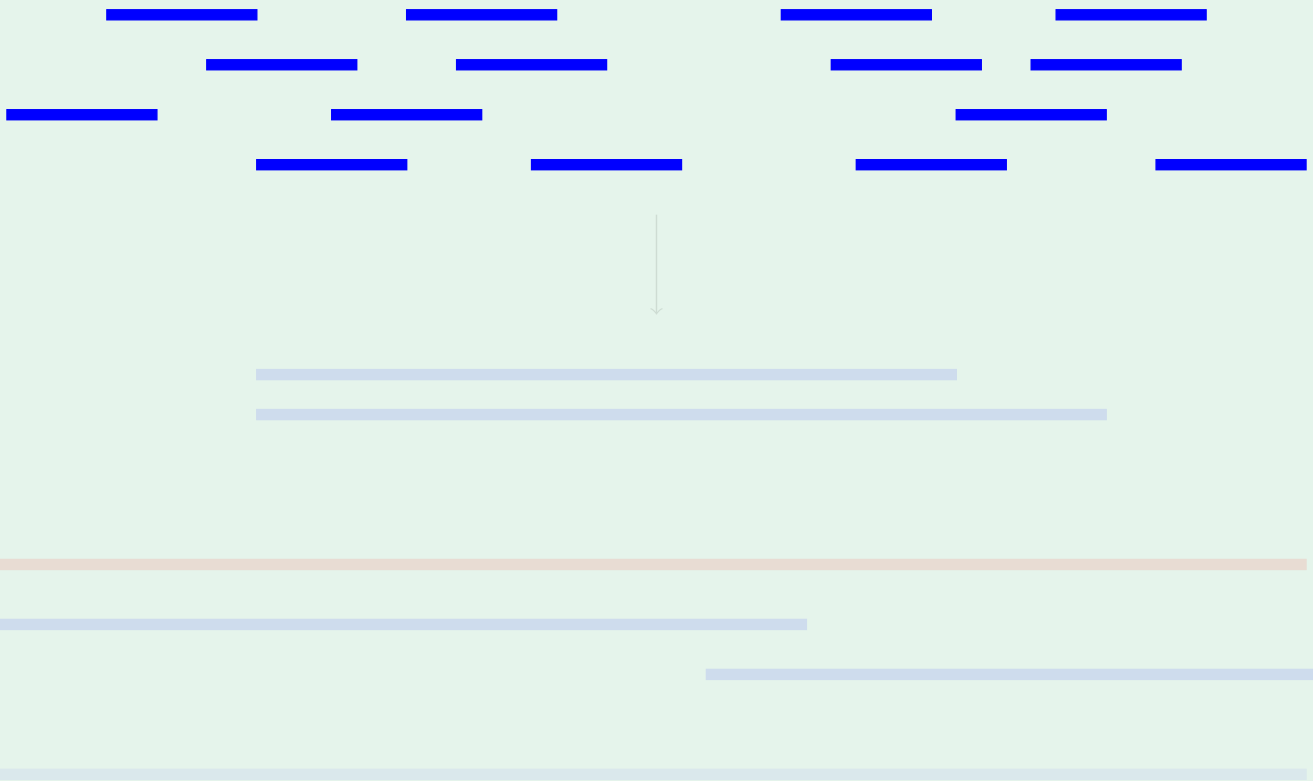
1) Short reads alignment

Overview



2) Contigs alignment

Overview

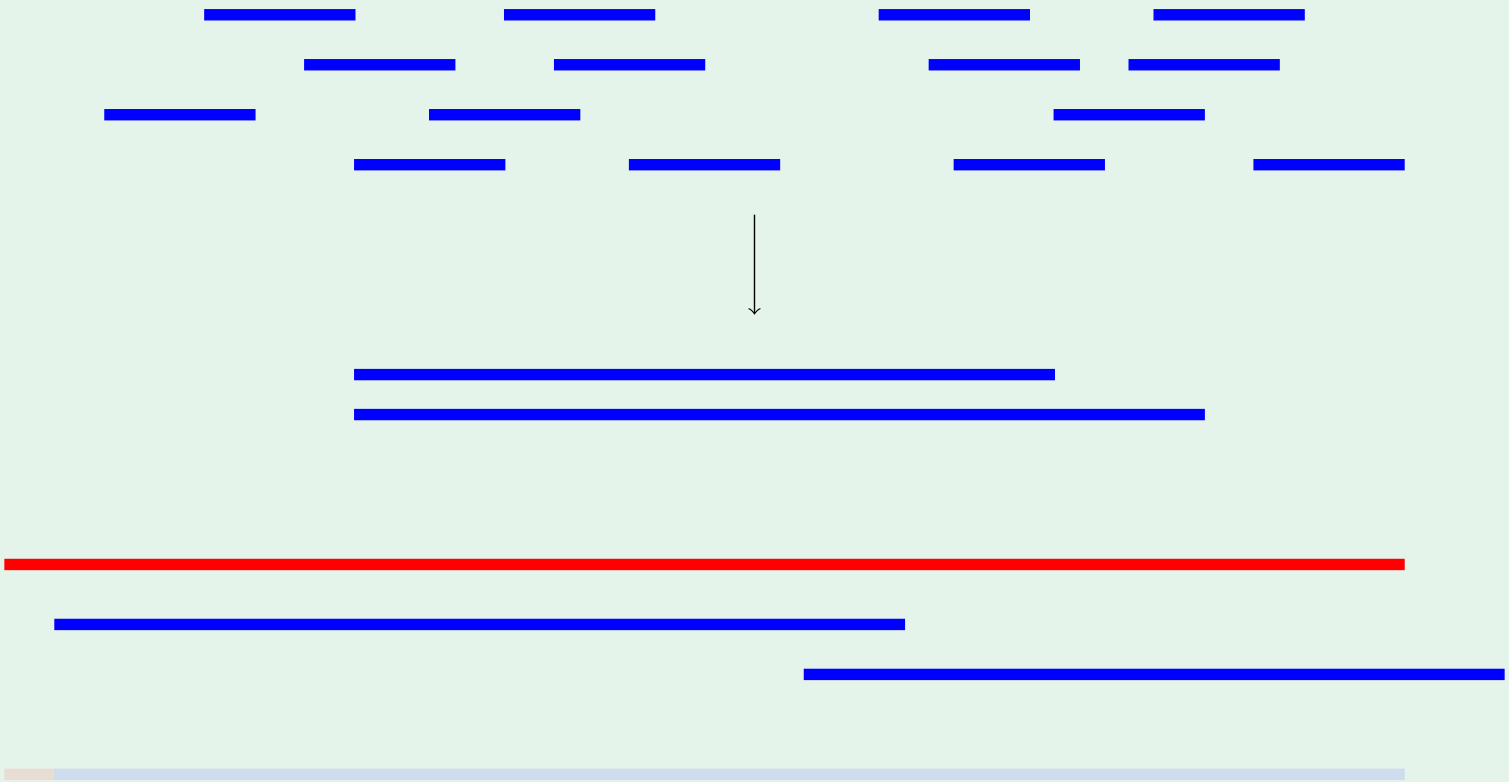


2) Contigs alignment

Overview

2) Contigs alignment

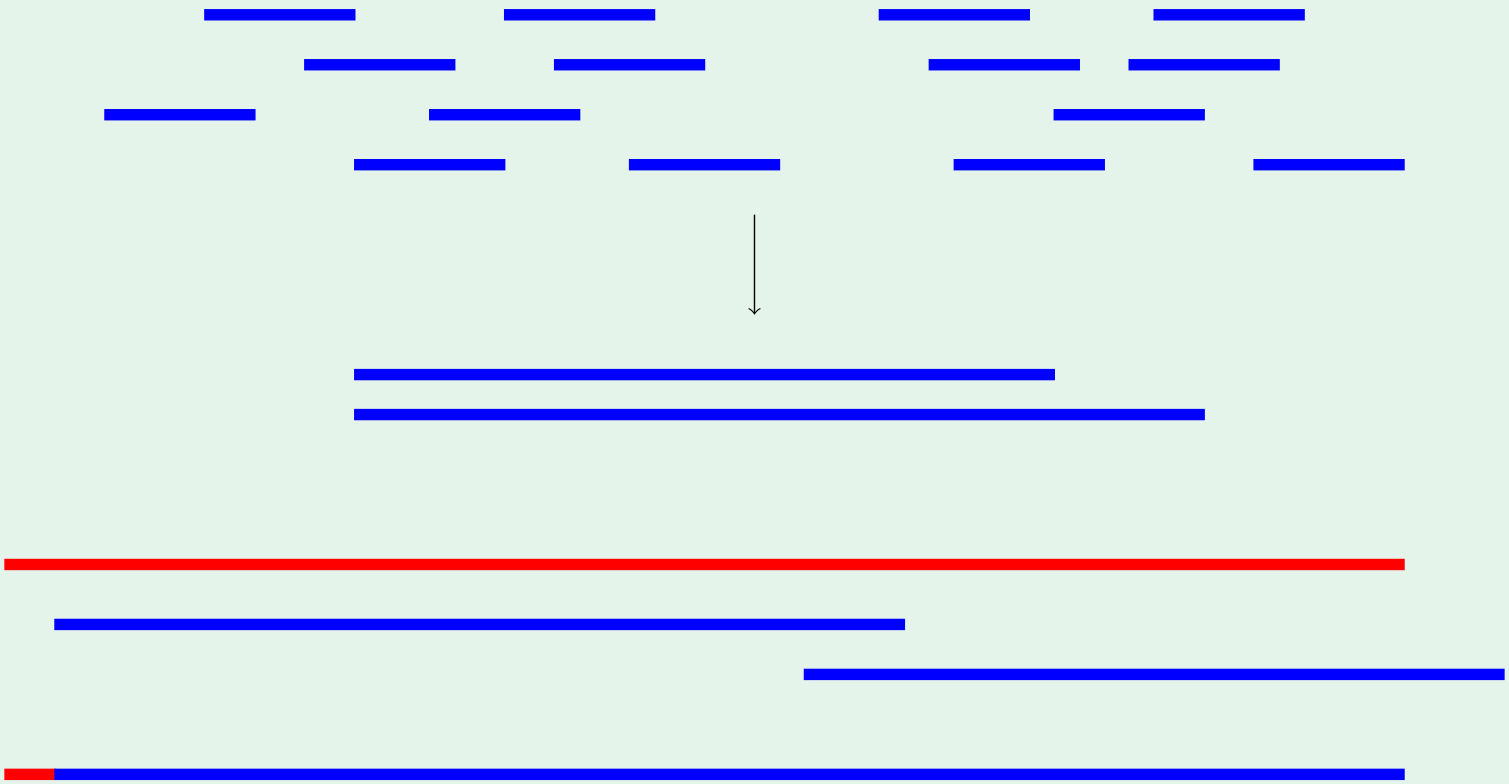
Overview



Olitis

2) Contigs alignment

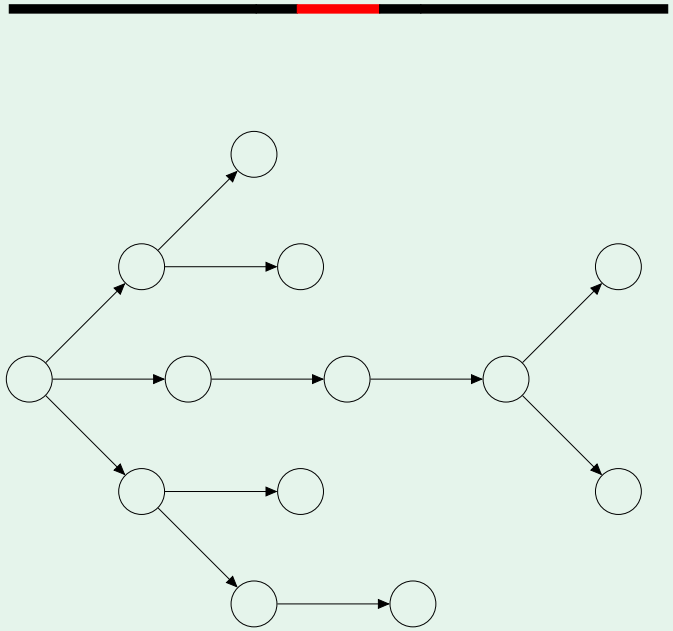
Overview



UNIVERSITÉ DE ROUEN NORMANDIE

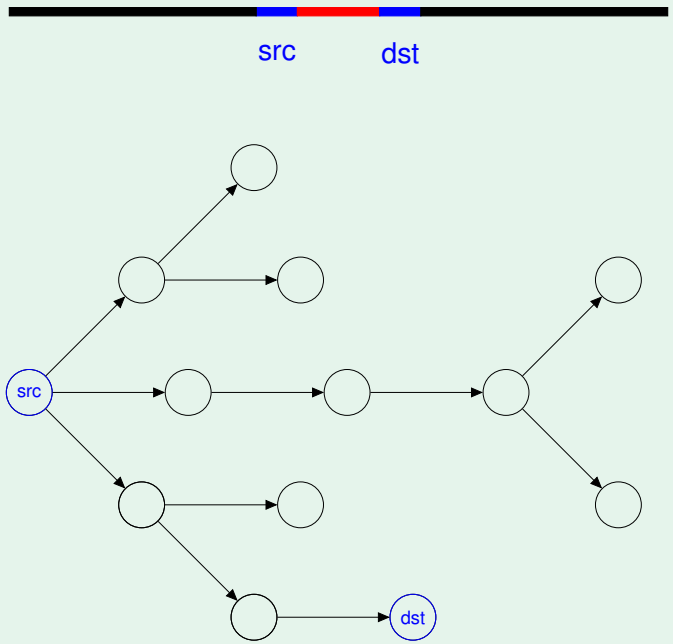
3) De Bruijn graphs

Overview



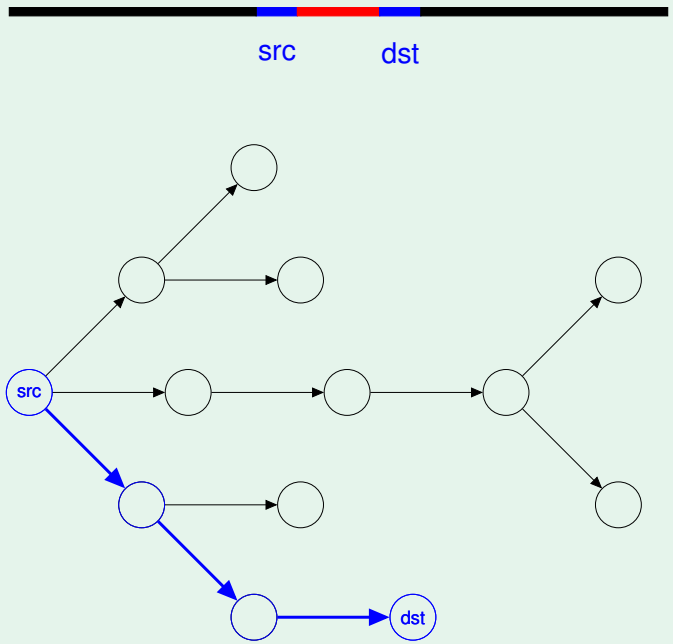
3) De Bruijn graphs

Overview



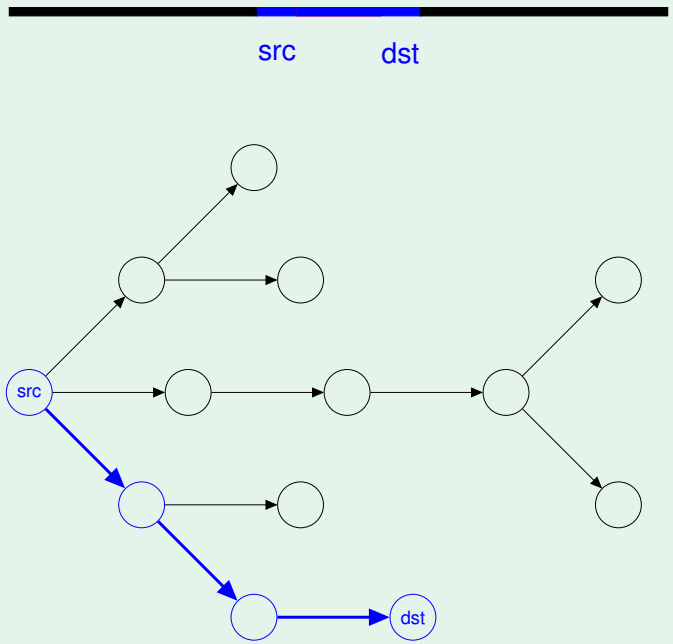
3) De Bruijn graphs

Overview



3) De Bruijn graphs

Overview



Olitis

17 Available methods

Method	Approach	Release
PBcR	SR alignment	2012
LSC	SR alignment	2012
ECTools	Contigs alignment	2014
LoRDEC	DBG	2014
Proovread	SR alignment	2014
Nanocorr	SR alignment	2015
NaS	SR alignment	2015
CoLoRMap	SR alignment	2016
Jabba	DBG	2016
LSCplus	SR alignment	2016
HALC	Contigs alignment	2017
HECIL	SR alignment	2017
Hercules	Hidden Markov models	2017
FMLRC	DBG	2018
HG-CoLoR	SR alignment + DBG	2018
MiRCA	Contigs alignment	2018
ParLECH	DBG	2019

Self-correction

- Only uses the information contained in the long reads
- State-of-the-art:
  1. Overlap the long reads
  2. Compute consensus from the overlaps
- Two approaches:
  1. Pseudo multiple sequence alignment (MSA)
  2. De Bruin graphs

1) Pseudo MSA

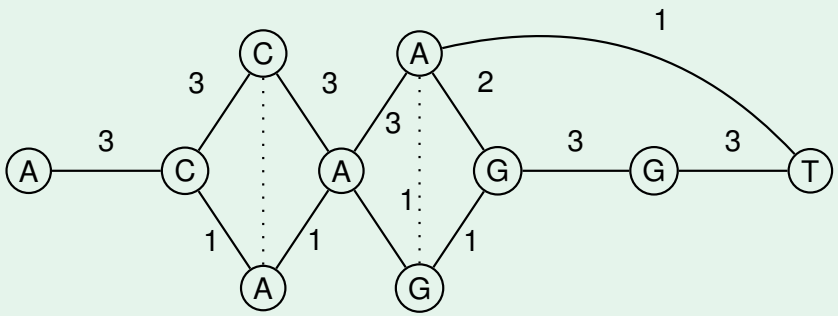
Overview

```
ACCAAGGT    R1
ACAAGGGT    R2

ACCAAGGT    R1
ACCAA..T    R3
```

1) Pseudo MSA

Overview

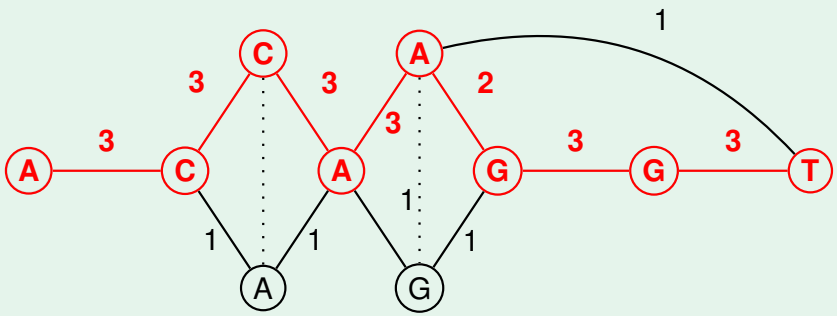
```
ACCAAGGT    R1
ACAAGGGT    R2

ACCAAGGT    R1
ACCAA..T    R3
```



1) Pseudo MSA

Overview

```
ACCAAGGT    R1
ACAAGGGT    R2

ACCAAGGT    R1
ACCAA..T    R3
```



2) De Bruijn graphs

Overview

```
.GATCGGG..TAT.TGCCCGTGTTTATGCGTGTG      R1
TGTTCAGGCAAATATG...GAAACAAGGCCTG..      R2

GAT..CGGGTATTGCCCGTGTTTATGCGTG..TG      R1
TATTTCTG..AT.GCGC.TGACTTTTCTTGGCAG      R3
```

2) De Bruijn graphs

Overview

Olitis

12 Available methods

Method	Approach	Release
PBcR-BLASR	Pseudo MSA	2013
PBDAGCon	Pseudo MSA	2013
Sprai	Pseudo MSA	2014
PBcR-MHAP	Pseudo MSA	2015
FalconSense	Pseudo MSA	2016
Sparc	Pseudo MSA	2016
Canu	Pseudo MSA	2017
Daccord	DBG	2017
LoRMA	DBG	2017
MECAT	Pseudo MSA	2017
FLAS	Pseudo MSA	2018
CONSENT	Pseudo MSA + DBG	2019

Problem

- Today: **29** tools are available
- Each of them claims to be the best...
- ... But what is the **truth**?

A truth

- Datasets charasteristics have huge impacts on correction:
  - Read length
  - Error rate
  - Sequencing depth
  - Organism complexity

Datasets

We gathered a total of 20 datasets having varying:

- Complexity (from bacteria to human)
- Sequencing technologies (PacBio and ONT)
- Error rates (12 to 44%)
- Sequencing depths (20x to 100x)
- Read length (few kbps to few hundreds of kbps)

Minimalist benchmark

- To lighten the presentation, we only study

Dataset	Number of reads	Error rate	Coverage	Number of bases
Simulated PacBio data				
S. cerevisiae 30x	45,198	12.28	30x	371 Mbp
C. elegans 30x	366,416	12.28	30x	3,006 Mbp
S. cerevisiae 60x	90,397	12.28	60x	742 Mbp
C. elegans 60x	732,832	12.28	60x	6,011 Mbp
Real ONT data				
A. baylyi	89,011	29.91	106x	381 Mbp
S. cerevisiae real	205,923	44.51	95x	1,173 Mbp

Hybrid correction:

- CoLoRMap
- LoRDEC
- HG-CoLoR

Self-correction:

- MECAT
- Daccord
- CONSENT

Scenarios

1. Low error rate, low coverage (30x *S. cerevisiae*, *C. elegans*)
2. Low error rate, medium coverage (60x *S. cerevisiae*, *C. elegans*)
3. High error rate, high coverage (real *A. baylyi*, *S. cerevisiae*)

Aim

- For each scenario, identify:
  - Is hybrid correction or self-correction more suited?
  - Which method does perform the best?

Low error rate and low coverage

		Hybrid correction			Self-correction		
Dataset	Metric	CoLoRMap	HG-CoLoR	LoRDEC	CONSENT	Daccord	MECAT
S. cerevisiae 30x	Number of bases (Mbp)	**343**	347	**348**	344	**348**	**285**
	Error rate (%)	0.3183	**0.5115**	0.3990	**0.4101**	**0.1259**	**0.3040**
	Runtime	**4 h 36 min**	**7 h 20 min**	35 min	**30 min**	1 h 19 min	**5 min**
	Memory (MB)	**14,243**	3,656	**799**	5,527	**31,798**	**2,907**
C. elegans 30x	Number of bases (Mbp)	**1,198**	**2,795**	**2,824**	2,789	-	**2,084**
	Error rate (%)	0.8955	**1.1664**	**1.2710**	**0.6495**	-	**0.3908**
	Runtime	**150 h 21 min**	**108 h 26 min**	11 h 30 min	**5 h 30 min**	-	**48 min**
	Memory (MB)	**32,267**	**27,212**	**2,320**	17,332	$>$ **250,000**	**10,535**

Summary

	Bacterial	Small eukaryotic	Larger eukaryotic
Low error rate, low coverage	-	Both, **Daccord**	Self, **MECAT**

Low error rate and medium coverage

		Hybrid correction			Self-correction		
Dataset	Metric	CoLoRMap	HG-CoLoR	LoRDEC	CONSENT	Daccord	MECAT
S. cerevisiae 60x	Number of bases (Mbp)	**664**	690	**696**	688	**695**	**616**
	Error rate (%)	**0.6143**	**0.5995**	0.3984	0.2897	**0.0400**	**0.2088**
	Runtime	**8 h 08 min**	**12 h 23 min**	**1 h 09 min**	1 h 31 min	2 h 26 min	**16 min**
	Memory (MB)	**24,375**	7,297	**794**	11,391	**23,190**	**4,954**
C. elegans 60x	Number of bases (Mbp)	-	-	**5,657**	**5,587**	-	**4,938**
	Error rate (%)	-	-	**1.2731**	**0.3858**	-	**0.2675**
	Runtime	**> 250 h**	**> 200 h**	23 h 30 min	16 h 43 min	-	**2 h 43 min**
	Memory (MB)	-	-	**2,332**	15,529	**> 250,000**	10,563

Summary

	Bacterial	Small eukaryotic	Larger eukaryotic
Low error rate, low coverage	-	Both, **Daccord**	Self, **MECAT**
Low error rate, medium coverage	-	Self, **Daccord**	Self, **MECAT**

High error rate and high coverage

		Hybrid correction			Self-correction		
	Metric	CoLoRMap	HG-CoLoR	LoRDEC	CONSENT	Daccord	MECAT
A. baylyi real	Number of bases (Mbp)	**141**	**285**	175	**185**	175	**154**
	Error rate (%)	0.4921	**0.0240**	**0.0552**	**5.7841**	**6.7454**	**8.5324**
	Runtime	**3 h 41 min**	**1 h 34 min**	**16 min**	26 min	43 min	**23 min**
	Memory (MB)	**13,028**	3,750	**436**	5,370	**25,801**	9,978
S. cerevisiae real	Number of bases (Mbp)	**165**	**512**	221	215	-	**84**
	Error rate (%)	**0.3042**	**0.2824**	1.1832	**13.3623**	-	**19.9237**
	Runtime	**10 h 44 min**	**8 h 51 min**	1 h 09 min	**12 min**	-	**14 min**
	Memory (MB)	**18,241**	11,575	**797**	13,697	$>$ **250,000**	7,374

Summary

	Bacterial	Small eukaryotic	Larger eukaryotic
Low error rate, low coverage	-	Both, **Daccord**	Self, **MECAT**
Low error rate, medium coverage	-	Self, **Daccord**	Self, **MECAT**
High error rate, high coverage	Hybrid, **HG-CoLoR**	Hybrid, **HG-CoLoR**	-

Olitis

Stay home messages

- Lots of error correction methods
- Each of them *can* be the best... ...on a particular dataset
- We provide a few guidelines:
  - Low coverages: self-correction performs quite well
  - Complex organisms: self-correction (Daccord, but quickly limited)
  - High error rates: hybrid correction (HG-CoLoR)
  - Speed: self-correction (MECAT), but LoRDEC is not so slow

Stay home messages

- Only a subset of results presented here
- Extended pre-print on bioRxiv:
  `https://doi.org/10.1101/2020.03.06.977975`
- Covers:
  - Algorithmic specifities
  - In-depth benchmark of all available tools on 20 datasets