Impact of the dataset characteristics on the quality of long read error correction

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.

SeqBIM 2019
Marne-La-Vallée
December 17th

Plan

1. Introduction
2. Hybrid correction
3. Self-correction
4. Available methods
5. Experiments
6. Conclusion

1 Introduction

2 Hybrid correction

3 Self-correction

4 Available methods

5 Experiments

6 Conclusion

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
- But also very noisy (10-30% error rates), most errors being indels

Error correction

- Correction: efficient way to handle these errors
- Two approaches:
  - Hybrid correction (makes use of complementary short reads)
  - Self-correction (corrects the long reads solely based on the information they contain)

Hybrid correction

- Long reads + short reads, sequenced for the same individual
- Use the short reads to correct the long reads
- SOTA ⇒ 4 approaches:

1. Short reads alignment
2. Contigs alignement
3. De Bruijn graphs
4. Hidden Markov models

Self-correction

- Only uses the information contained in the long reads
- Recent developments
- Third generation sequencing technologies evolve fast:
  - Decrease of the error rates (10-12%)
  - Increase of the read length (ultra-long reads ONT $>$ 1 Mbp)

Self-correction

- SOTA:
  - Overlap the long reads
  - Compute consensus from the overlaps
- Two approaches:
  1. Pseudo multiple sequence alignment (MSA)
  2. De Bruin graphs

1 Introduction

2 **Hybrid correction**

3 Self-correction

4 Available methods

5 Experiments

6 Conclusion

Hybrid correction

1. Short reads alignment
2. Contigs alignement
3. De Bruijn graphs
4. Hidden Markov models

Hybrid correction

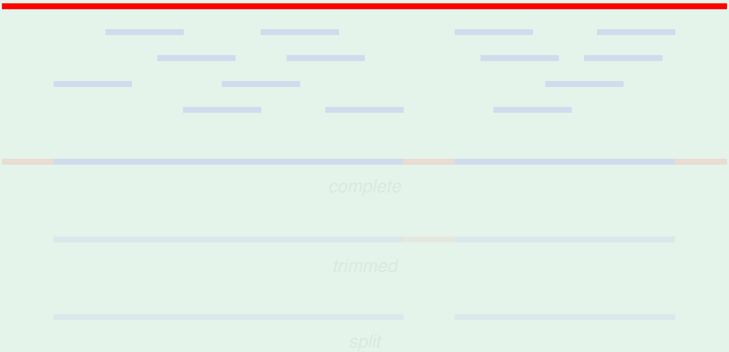
1. Short reads alignment
2. Contigs alignement
3. De Bruijn graphs

Short reads alignments

- First hybrid correction approach
- Align the short reads to the long reads
- Define MSA from the shorts reads
- Use the MSA to compute consensus

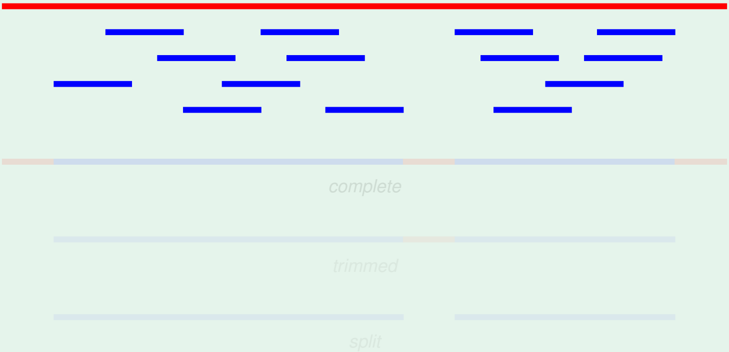
Short reads alignments

Example



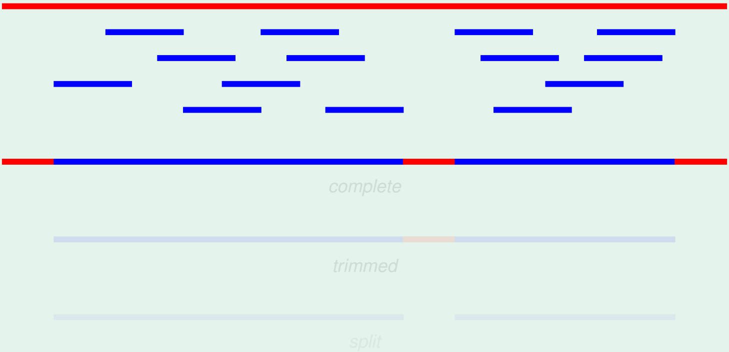
Short reads alignments

Example



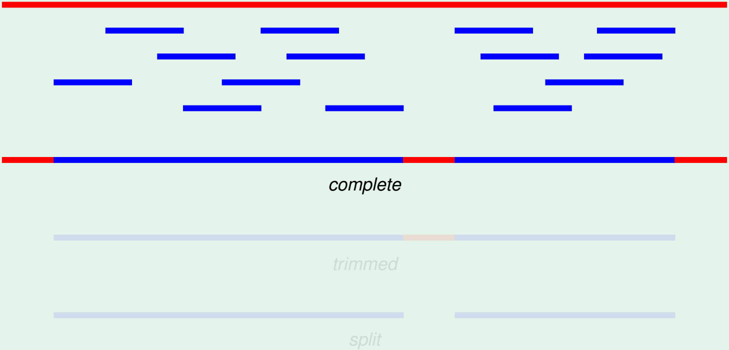
Short reads alignments

Example



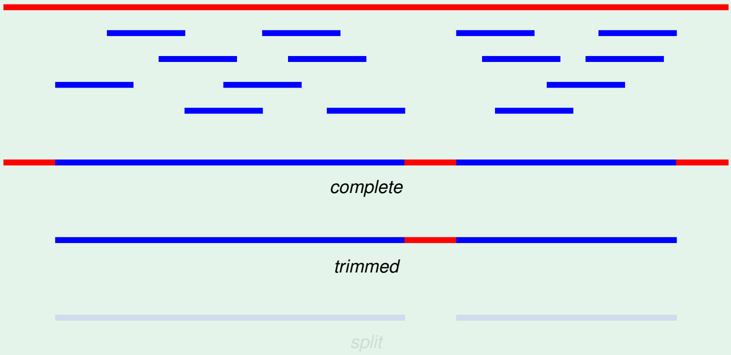
Short reads alignments

Example



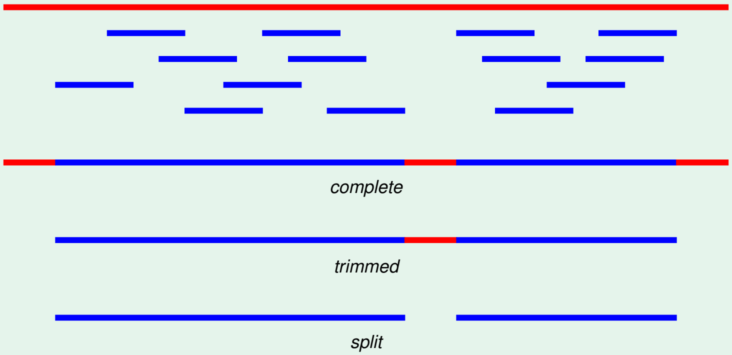
Short reads alignments

Example



Short reads alignments

Example



Hybrid correction

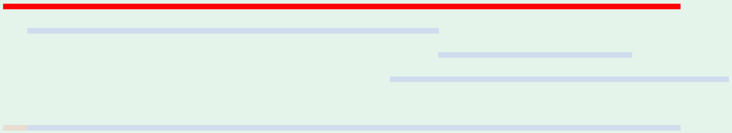
1. Short reads alignment
2. Contigs alignement
3. De Bruijn graphs

Contigs alignment

- Contigs are much longer than short reads
- Easier to cover highly noisy regions of the long reads
- Build contigs from the short reads
- Align the contigs and the long reads
- Define MSA and compute consensus

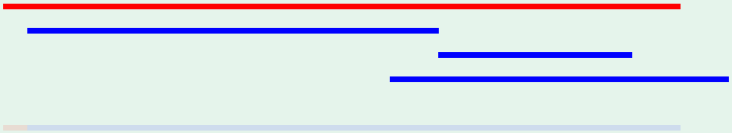
Contigs alignment

Example



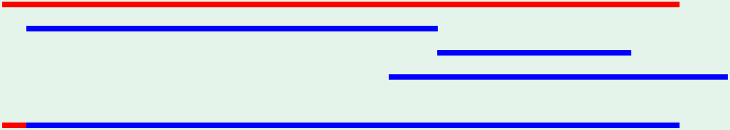
Contigs alignment

Example



Contigs alignment

Example



Hybrid correction

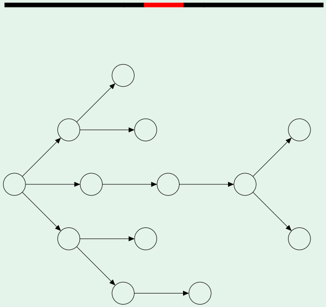
1. Short reads alignment
2. Contigs alignement
3. De Bruijn graphs

De Bruijn graphs

- Build the graph from the short reads solid *k*-mers
- Anchor the long reads to the graph
- Correct weak *k*-mer regions of the long reads with the graph

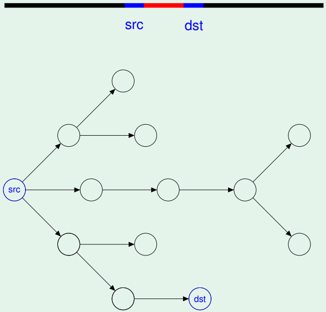
De Bruijn graphs

Example



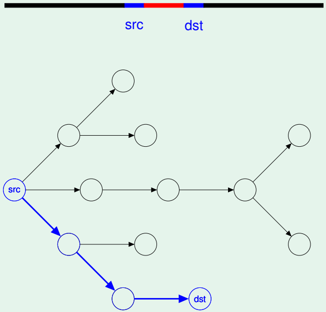
De Bruijn graphs

Example



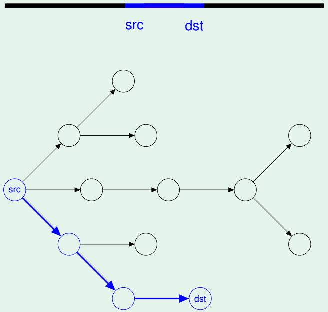
De Bruijn graphs

Example



De Bruijn graphs

Example



1. Introduction
2. Hybrid correction
3. **Self-correction**
4. Available methods
5. Experiments
6. Conclusion

Self-correction

1. Pseudo MSA
2. De Bruijn graphs

Self-correction

1. Pseudo MSA
2. De Bruijn graphs

Pseudo MSA

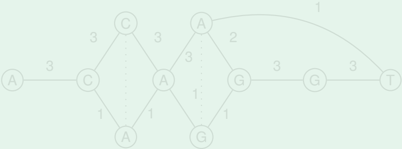
- Overlap the long reads
- Build a directed acyclic graph (DAG) to summarize the overlaps
- The DAG represents a pseudo MSA
- Compute consensus by extracting the highest weighted path

Pseudo MSA

Example

AC**C**A**A**GGT R_1
AC**A**A**G**GGT R_2

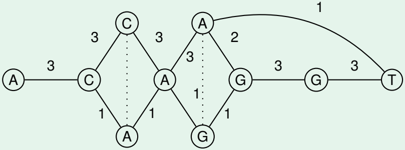
ACCAA**GG**T R_1
ACCAA..T R_3



Pseudo MSA

Example

AC**C**A**A**GGT	R_1
AC**A**A**G**GGT	R_2
ACCAA**GG**T	R_1
ACCAA..T	R_3

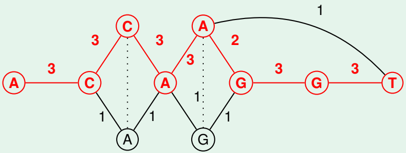


Pseudo MSA

Example

AC**C**A**A**GGT R_1
AC**A**A**G**GGT R_2

ACCAA**GG**T R_1
ACCAA..T R_3



Self-correction

1. Pseudo MSA
2. De Bruijn graphs

De Bruijn graphs

- Overlap the long reads
- Divide the overlaps into small windows
- Build a DBG for each window
- Correct the windows with the DBGs

De Bruijn graphs

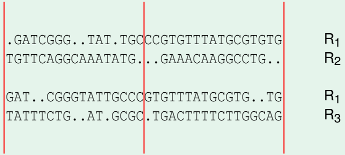
Example

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

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

De Bruijn graphs

Example



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

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

1 Introduction

2 Hybrid correction

3 Self-correction

4 **Available methods**

5 Experiments

6 Conclusion

Hybrid correction

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
MiRCA	Contigs alignment	2018
HG-CoLoR	SR alignment + DBG	2018

16 methods

Self-correction

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

12 methods

Summary

- Today: 28 available methods
- Each of them claims to be the best...
- ... But what is the **truth**?

Summary

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

1. Introduction
2. Hybrid correction
3. Self-correction
4. Available methods
5. **Experiments**
6. Conclusion

Datasets

We gathered a wide variety of datasets having varying:

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

Datasets

Dataset	Number of reads	Error rate	Coverage	Number of bases
Simulated PacBio data				
E. coli 20x	11,306	18.60	20x	93 Mbp
S. cerevisiae 20x	30,132	18.60	20x	247 Mbp
C. elegans 20x	244,277	18.60	20x	2,004 Mbp
E. coli 30x	16,959	12.28	30x	140 Mbp
S. cerevisiae 30x	45,198	12.28	30x	371 Mbp
C. elegans 30x	366,416	12.28	30x	3,006 Mbp
E. coli 60x	33,918	12.28	60x	279 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	205,923	44.51	95x	1,173 Mbp
H. sapiens[1]	1,075,867	17.60	29x	7,256 Mbp

[1] contains ONT ultra-long reads (up to 340 kbp)

Scenarios

- Low error rate and low coverage
- Low error rate and medium coverage
- Medium error rate and low coverage
- High error rate and high coverage
- Ultra-long reads (medium error rate)

Aim

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

Assessed tools

To lighten the presentation, we only assess:

Hybrid correction:

- CoLoRMap
- HG-CoLoR
- LoRDEC

Self-correction:

- CONSENT
- Daccord
- MECAT

Scenarios

- Low error rate and low coverage
- Low error rate and medium coverage
- Medium error rate and low coverage
- High error rate and high coverage
- Ultra-long reads (medium error rate)

Low error rate and low coverage

	Metric	CoLoRMap	HG-CoLoR	LoRDEC	CONSENT	Daccord	MECAT
E. coli 30x	Number of bases (Mbp)	**134**	131	131	130	131	**107**
	Error rate (%)	0.1137	**0.0726**	**0.0695**	**0.3350**	**0.0248**	**0.2569**
	Recall (%)	99.9881	**99.9986**	99.9831	**99.9419**	**99.9965**	**99.9302**
	Precision (%)	99.8880	99.9279	99.9328	**99.6701**	**99.9757**	**99.7533**
	Runtime	**1 h 33 min**	**1 h 20 min**	12 min	17 min	14 min	**2 min**
	Memory (MB)	**13,097**	1,538	**460**	2,212	**6,813**	1,600
S. cerevisiae 30x	Number of bases (Mbp)	343	347	**348**	344	**348**	**285**
	Error rate (%)	0.3183	**0.5115**	0.3990	**0.4258**	**0.1259**	0.3040
	Recall (%)	99.9135	99.9592	**99.8123**	99.9296	**99.9874**	99.9160
	Precision (%)	99.6860	**99.4937**	99.6093	99.5807	**99.8762**	99.7072
	Runtime	**4 h 36 min**	**7 h 20 min**	35 min	47 min	1 h 19 min	**5 min**
	Memory (MB)	**14,243**	3,656	**799**	5,514	**31,798**	2,907
C. elegans 30x	Number of bases (Mbp)	**1,198**	2,795	**2,824**	2,787	-	2,084
	Error rate (%)	0.8955	**1.1664**	**1.2710**	**0.6720**	-	**0.3908**
	Recall (%)	**99.9165**	99.9104	**99.4191**	99.8970	-	99.8903
	Precision (%)	99.1230	**98.4889**	98.7441	99.3378	-	**99.6212**
	Runtime	**150 h 21 min**	**108 h 26 min**	11 h 30 min	7 h 54 min	-	**48 min**
	Memory (MB)	**32,267**	**27,212**	**2,320**	16,772	**> 250,000**	10,535

Summary

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

Scenarios

- Low error rate and low coverage
- Low error rate and medium coverage
- Medium error rate and low coverage
- High error rate and high coverage
- Ultra-long reads (medium error rate)

Low error rate and medium coverage

	Metric	CoLoRMap	HG-CoLoR	LoRDEC	CONSENT	Daccord	MECAT
E. coli 60x	Number of bases (Mbp)	**266**	262	261	259	261	**233**
	Error rate (%)	0.1621	**0.0771**	**0.0684**	**0.1799**	**0.0214**	0.1714
	Recall (%)	**99.9631**	**99.9987**	99.9832	99.9801	**99.9971**	**99.9547**
	Precision (%)	99.8400	99.9234	99.9339	**99.8229**	**99.9790**	**99.8362**
	Runtime	**3 h 01 min**	**2 h 03 min**	20 min	37 min	54 min	**5 min**
	Memory (MB)	**19,898**	2,744	**457**	4,913	**18,450**	2,387
S. cerevisiae 60x	Number of bases (Mbp)	664	690	**696**	688	695	**616**
	Error rate (%)	**0.6143**	0.5995	0.3984	0.2812	**0.0400**	0.2088
	Recall (%)	**99.7755**	99.9433	99.8136	99.9582	**99.9928**	99.9428
	Precision (%)	**99.3917**	99.4059	99.6100	99.7231	99.6906	**99.7996**
	Runtime	**8 h 08 min**	**12 h 23 min**	1 h 09 min	1 h 49 min	2 h 26 min	**16 min**
	Memory (MB)	**24,375**	7,297	**794**	11,335	**23,190**	4,954
C. elegans 60x	Number of bases (Mbp)	-	-	**5,657**	5,586	-	**4,938**
	Error rate (%)	-	-	**1.2731**	0.3806	-	**0.2675**
	Recall (%)	-	-	**99.4201**	**99.9489**	-	99.9258
	Precision (%)	-	-	**98.7420**	99.6254	-	**99.7415**
	Runtime	**> 250 h**	**> 200 h**	23 h 30 min	19 h 13 min	-	**2 h 43 min**
	Memory (MB)	-	-	2,332	15,607	**> 250,000**	10,563

Summary

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

Scenarios

- Low error rate and low coverage
- Low error rate and medium coverage
- Medium error rate and low coverage
- High error rate and high coverage
- Ultra-long reads (medium error rate)

Medium error rate and low coverage

	Metric	CoLoRMap	HG-CoLoR	LoRDEC	CONSENT	Daccord	MECAT
E. coli 20x	Number of bases (Mbp)	81	**84**	78	61	84	**59**
	Error rate (%)	0.1946	**0.0691**	0.1474	**8.5423**	0.3965	0.5243
	Recall (%)	99.9890	**99.9982**	99.9890	**97.9155**	99.8817	99.8317
	Precision (%)	99.8118	**99.9315**	99.8570	**91.5687**	99.6077	99.4915
	Runtime	**1 h 25 min**	**51 min**	8 min	8 min	24 min	**26 sec**
	Memory (MB)	**6,659**	1,517	**455**	1,552	**4,538**	1,322
S. cerevisiae 20x	Number of bases (Mbp)	211	220	188	166	**222**	**162**
	Error rate (%)	**0.2655**	0.2959	0.5400	**8.2652**	0.5447	0.6555
	Recall (%)	99.9805	**99.9900**	99.9483	**98.0349**	99.8591	99.8015
	Precision (%)	**99.7413**	99.7071	99.4730	**91.8483**	99.4630	99.3636
	Runtime	**4 h 42 min**	**4 h 55 min**	28 min	22 min	1 h 10 min	**1 min**
	Memory (MB)	**13,544**	3,237	**799**	4,514	**14,111**	2,207
C. elegans 20x	Number of bases (Mbp)	**517**	**1,726**	1,155	1,359	-	871
	Error rate (%)	2.6255	**0.6524**	1.2643	**9.5548**	-	**0.6540**
	Recall (%)	99.8445	**99.9682**	99.8871	**97.9553**	-	99.8196
	Precision (%)	**99.4526**	99.3554	98.7542	**90.5794**	-	99.3597
	Runtime	**125 h 44 min**	**88 h 10 min**	6 h 01 min	3 h 49 min	-	**18 min**
	Memory (MB)	**32,188**	19,730	**2,238**	14,522	-	10,340

Summary

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

Scenarios

- Low error rate and low coverage
- Low error rate and medium coverage
- Medium error rate and low coverage
- High error rate and high coverage
- Ultra-long reads (medium error rate)

High error rate and high coverage

	Metric	CoLoRMap	HG-CoLoR	LoRDEC	CONSENT	Daccord	MECAT
A. baylyi real	Number of bases (Mbp)	**141**	**285**	175	183	175	154
	Mean length (bp)	3,882	**11,156**	3,449	**10,815**	**3,244**	9,186
	Error rate (%)	0.4921	**0.0240**	**0.0552**	**8.0530**	**6.7454**	**8.5324**
	Genome overage (%)	**100.0000**	**100.0000**	**100.0000**	**100.0000**	**100.0000**	**100.0000**
	Runtime	**3 h 41 min**	**1 h 34 min**	16 min	48 min	43 min	**23 min**
	Memory (MB)	**13,028**	3,750	**436**	5,150	**25,801**	9,978
S. cerevisiae real	Number of bases (Mbp)	165	**512**	221	179	-	**84**
	Mean length (bp)	**2,294**	**6,725**	**1,125**	7,186	-	5,668
	Error rate (%)	**0.3042**	**0.2824**	1.1832	**23.2735**	-	**19.9237**
	Genome coverage (%)	99.1528	**99.5341**	98.8934	98.1075	-	**92.6533**
	Runtime	**10 h 44 min**	**8 h 51 min**	1 h 09 min	40 min	-	**14 min**
	Memory (MB)	**18,241**	11,575	**797**	14,663	**> 250,000**	7,374

Summary

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

Scenarios

- Low error rate and low coverage
- Low error rate and medium coverage
- Medium error rate and low coverage
- High error rate and high coverage
- Ultra-long reads (medium error rate)

	Metric	CoLoRMap	HG-CoLoR	LoRDEC	CONSENT	Daccord	MECAT
H. sapiens	Number of bases (Mbp)	**1,511**	6,553	**6,851**	6,349	-	-
	Mean length (bp)	**3,603**	6,754	6,368	**7,301**	-	-
	Error rate (%)	3.5498	**1.1958**	**8.2795**	**6.9996**	-	-
	Genome coverage (%)	**91.9475**	**92.4523**	**92.4693**	**92.3993**	-	-
	Runtime	**304 h 10 min**	**167 h 47 min**	12 h 52 min	**8 h 29 min**	-	-
	Memory (MB)	**80,613**	**50,898**	**7,902**	17,350	-	-

Summary

	Bacterial	Small eukaryotic	Larger eukaryotic
Low error rate, low coverage	Both, **Daccord**	Both, **Daccord**	Self, **MECAT**
Low error rate, medium coverage	Both, **Daccord**	Self, **Daccord**	Self, **MECAT**
Medium error rate, low coverage	Hybrid, **HG-CoLoR**	Hybrid, **CoLoRMap**	Hybrid, **HG-CoLoR**
High error rate, high coverage	Hybrid, **HG-CoLoR**	Hybrid, **HG-CoLoR**	-
Ultra-long reads	Most self-correction methods do not scale... Hybrid, or **CONSENT**		

1 Introduction

2 Hybrid correction

3 Self-correction

4 Available methods

5 Experiments

6 Conclusion

Take home messages

- Lots of error correction methods
- Each of them *can* be the best... ... on a particular dataset
- We provide a few guidelines:
  - Low coverage: self-correction performs quite well
  - Complex organism: self-correction (Daccord is quickly limited ⇒ CONSENT? MECAT?)
  - High error rate: hybrid correction (HG-CoLoR)
  - Fast: self-correction ⇒ MECAT (but LoRDEC is not so slow)
  - Ultra-long reads: hybrid correction or CONSENT

Future work

- Add new datasets:
  - Medium error rate with higher coverage: does self-correction perform better?
  - Low error rate and extremely low coverage (10x): can self-correction still work?