Assessing efficiency of third generation sequencing read correction

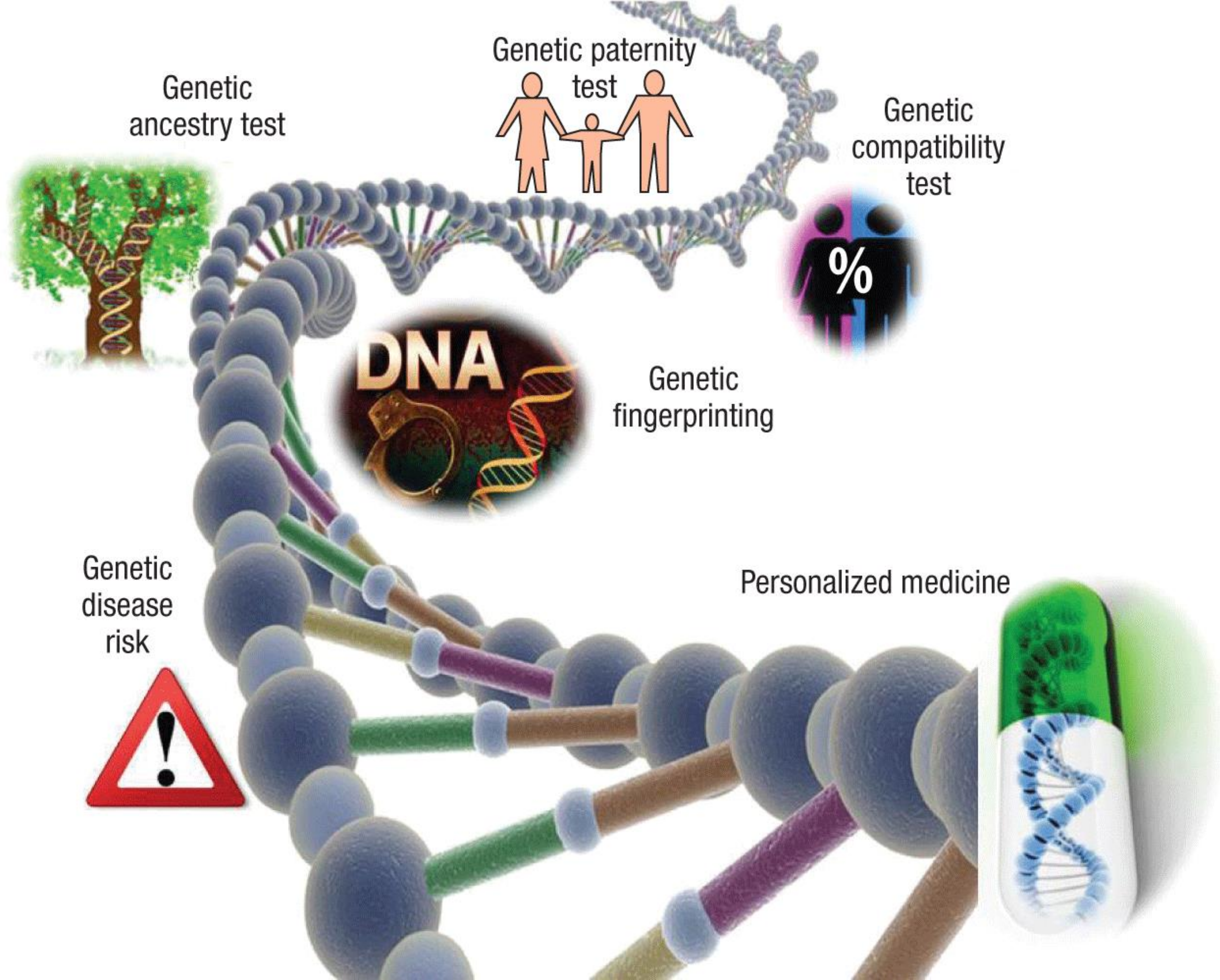
Medhat Mahmoud
Department of Protein Biosynthesis (IBCh)
&
Department of Computational Biology (AMU)

Poznan 24-05-2017

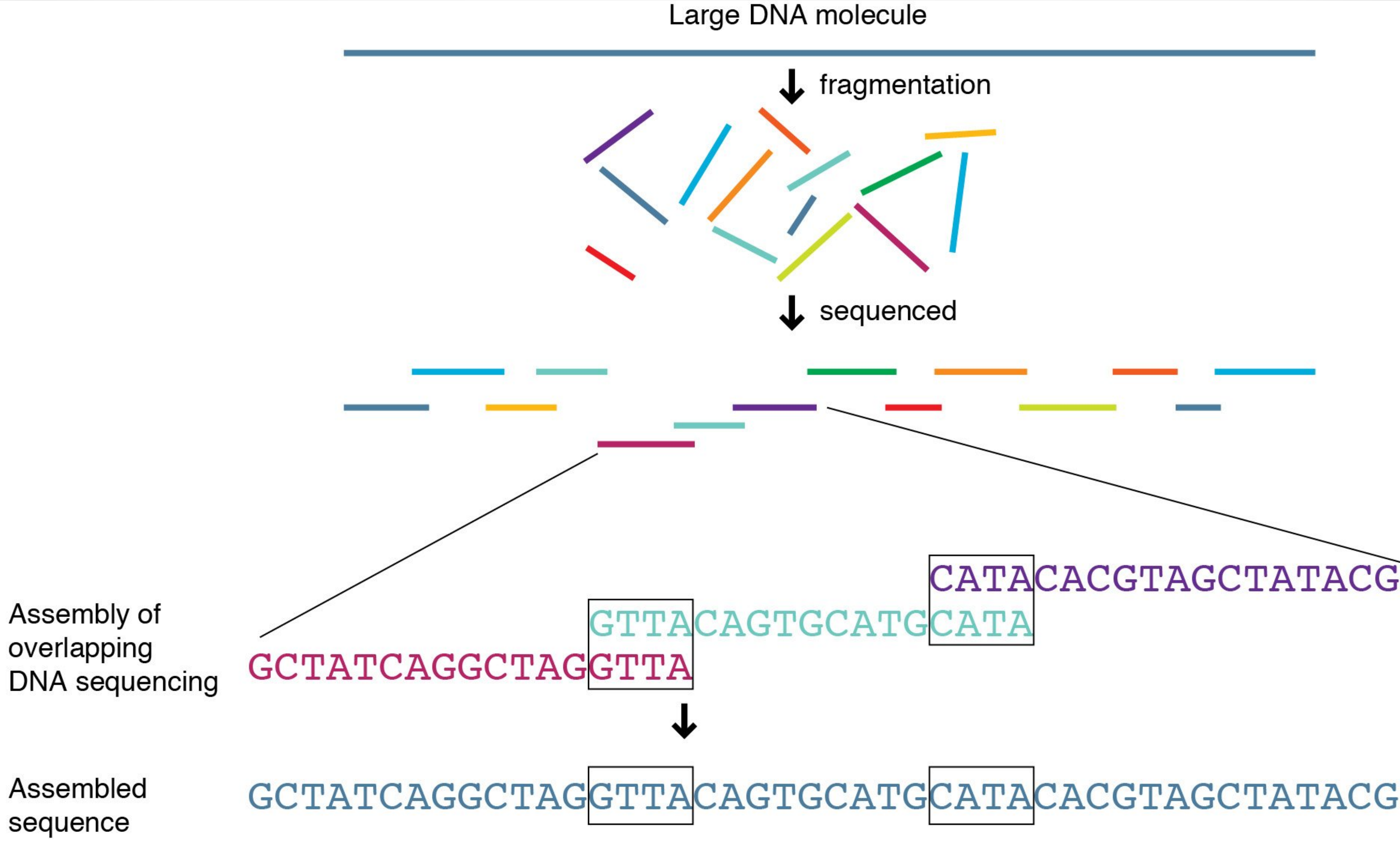
The goal

Compare sensitivity, specificity and performance of Single molecule real-time (SMRT) PacBio reads correction tools

…


Genetic paternity test
Genetic ancestry test
Genetic compatibility test
%
DNA
Genetic fingerprinting
Genetic disease risk
Personalized medicine

Nice...
We have so many options!
- Deep whole genome
- Low pass whole genome
- Deep whole exome
-Genomewide array
- Exome array
How would you like to be sequenced?
MENU
Genomics Core Facility
www.biocomicals.com



Large DNA molecule

fragmentation

sequenced

Assembly of overlapping DNA sequencing

CATACACGTAGCTATACG

GTTACAGTGCATGCATA

GCTATCAGGCTAGGTTA

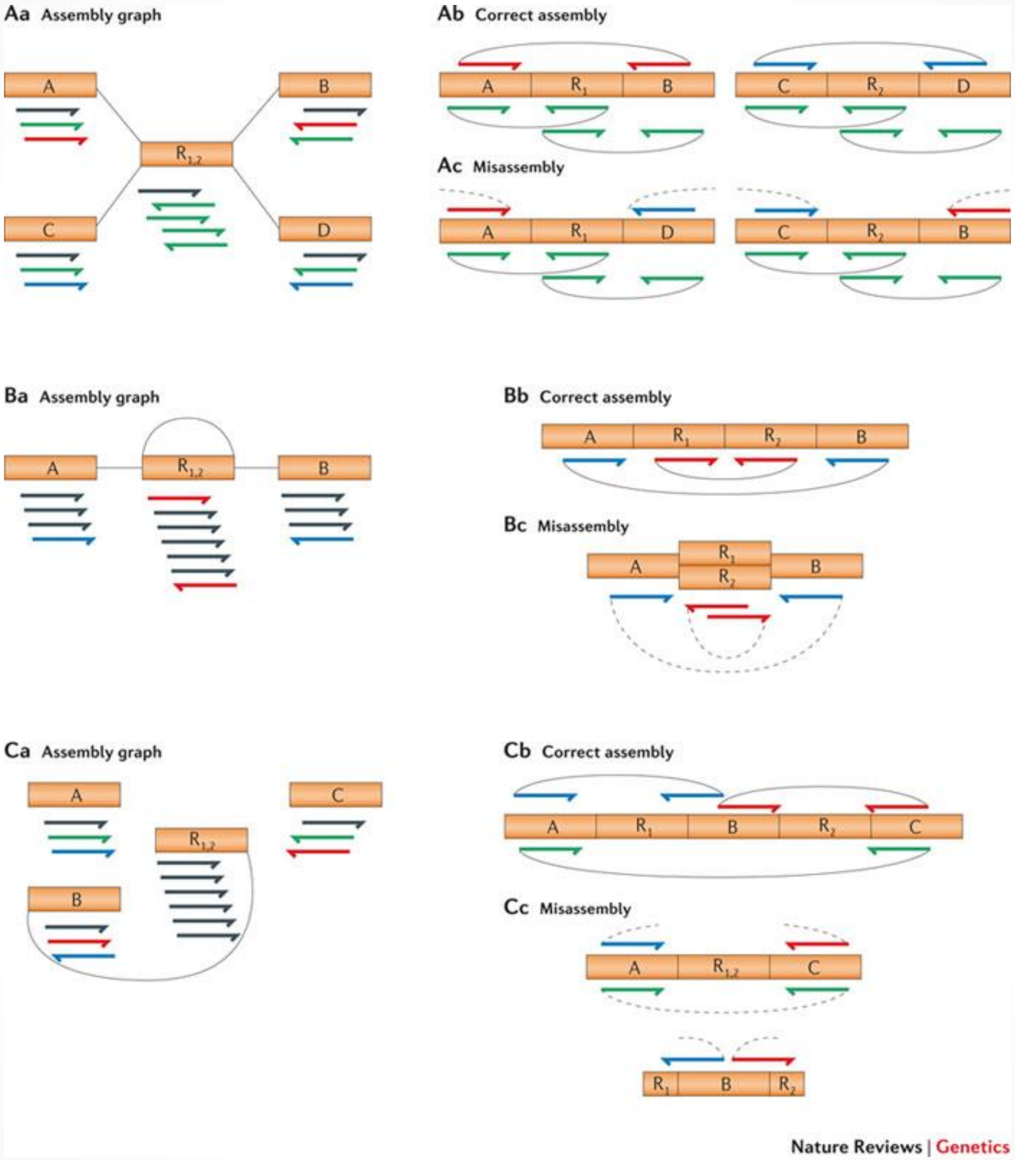
Assembled sequence

GCTATCAGGCTAGGTTACAGTGCATGCATACACGTAGCTATACG

A R_1 B

A R_2 B

A R_1 R_2 B


Aa Assembly graph

Ab Correct assembly

Ac Misassembly

Ba Assembly graph

Bb Correct assembly

Bc Misassembly

Ca Assembly graph

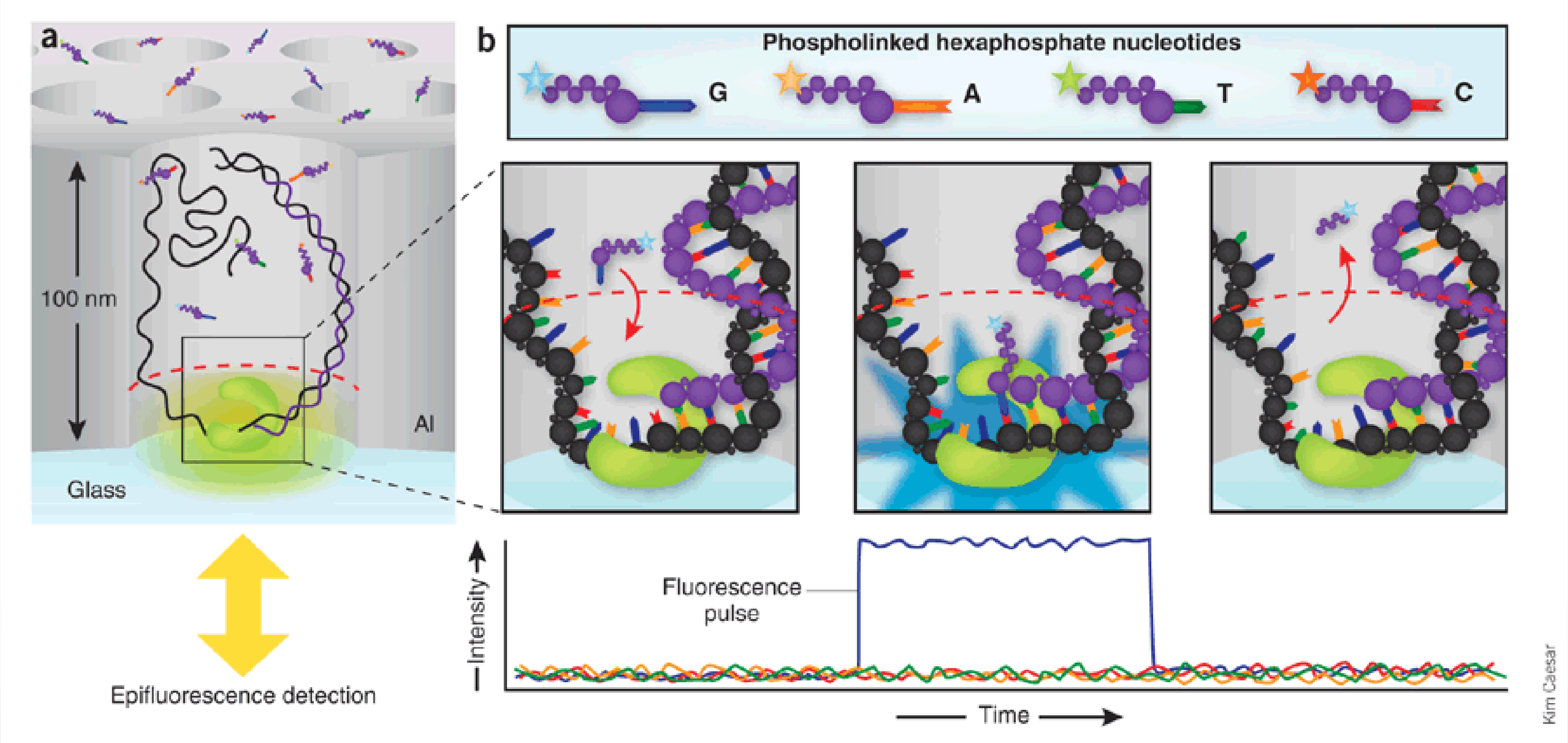
Cb Correct assembly

Cc Misassembly

Nature Reviews | Genetics

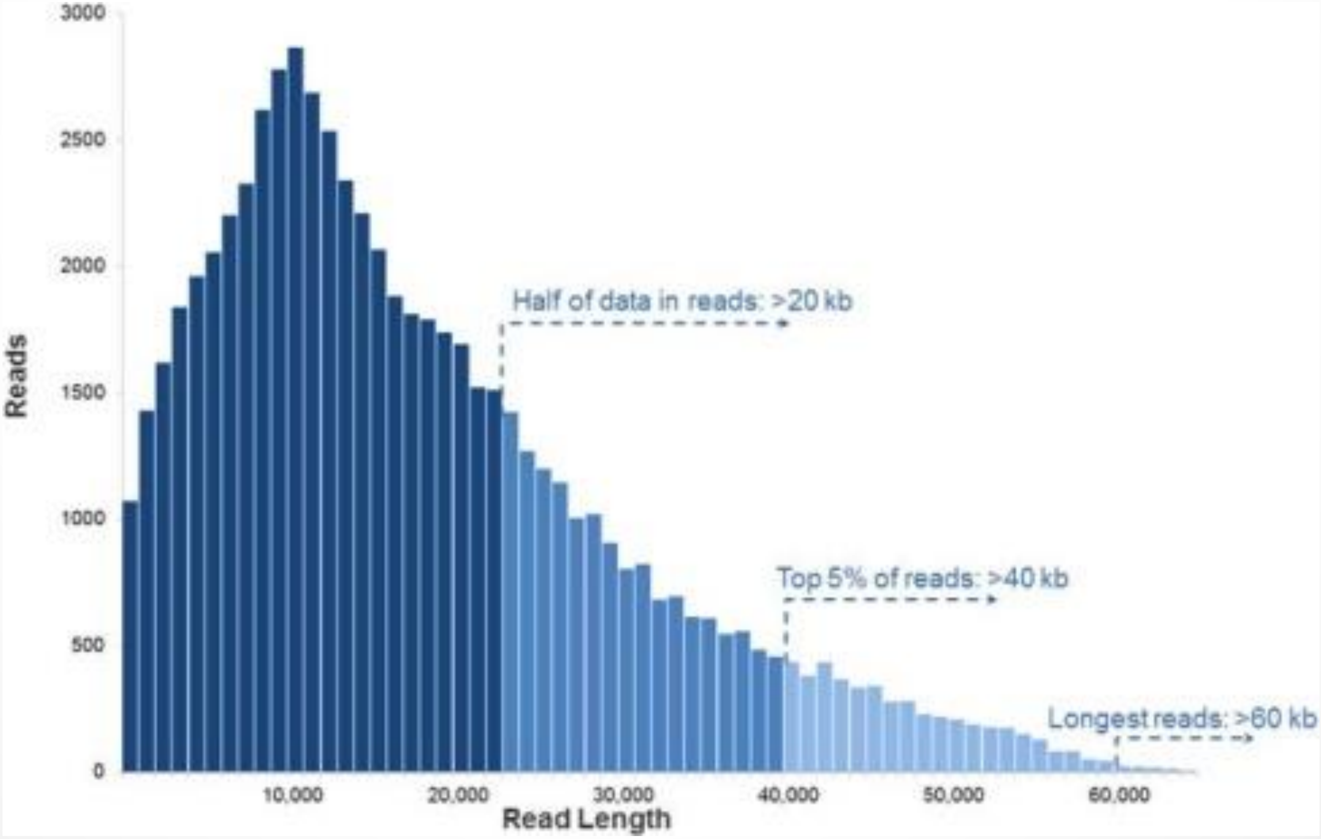
SMRT PacBio sequencing

The idea


a
100 nm
Al
Glass
Epifluorescence detection
b
Phospholinked hexaphosphate nucleotides
G
A
T
C
Intensity
Fluorescence pulse
Time
Kim Caesar

PacBio read advantages

- offers much longer read lengths and faster runs than SGS (reaches 60,000 bp)

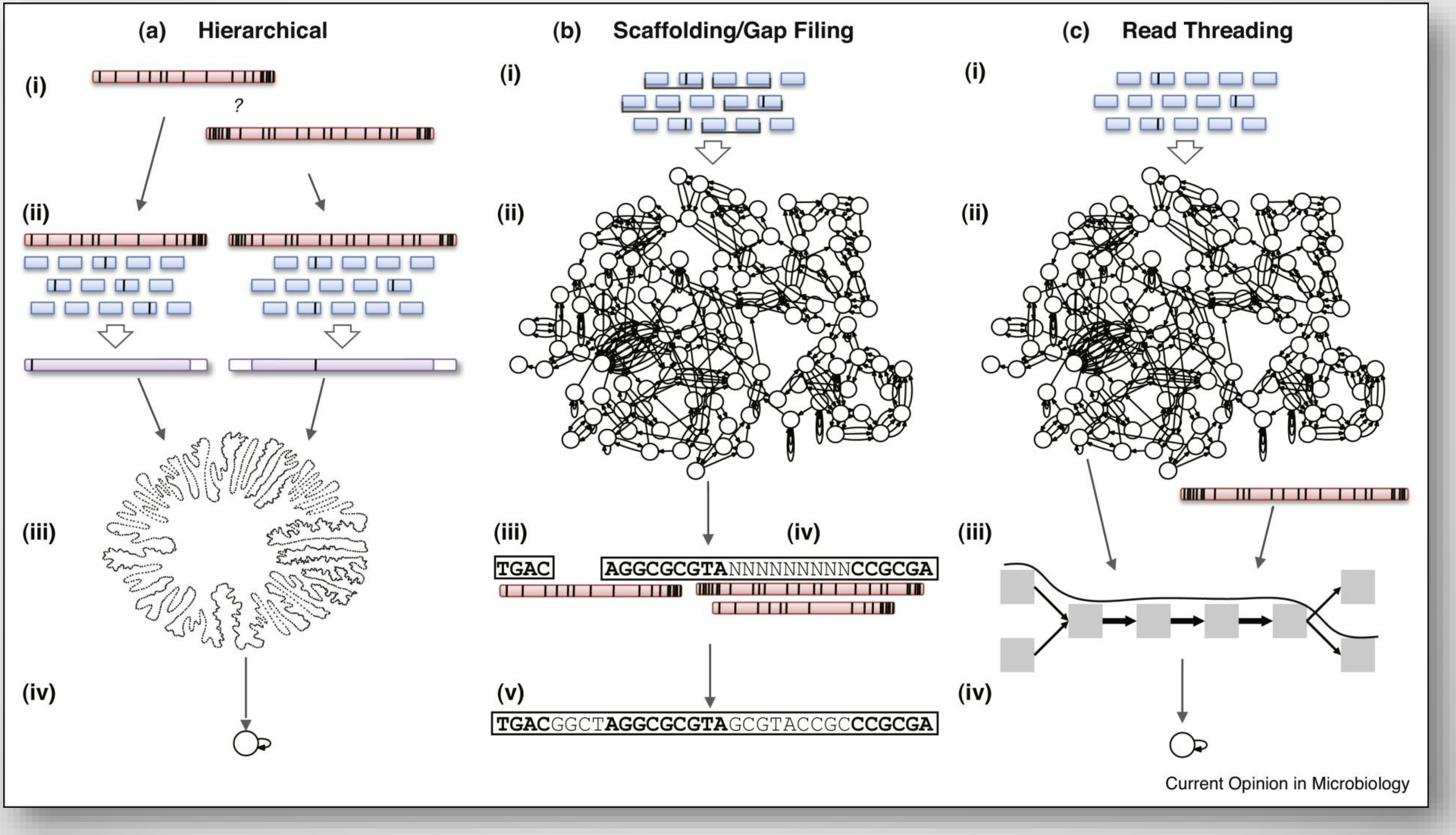


repetitive
structural
genes
mutations
variation
hybrid
detect
sequencing
isoforms
novel
assemblies
gaps
gene
novo
annotated
close
regions
methylation
resolve

Where is the catch ?!

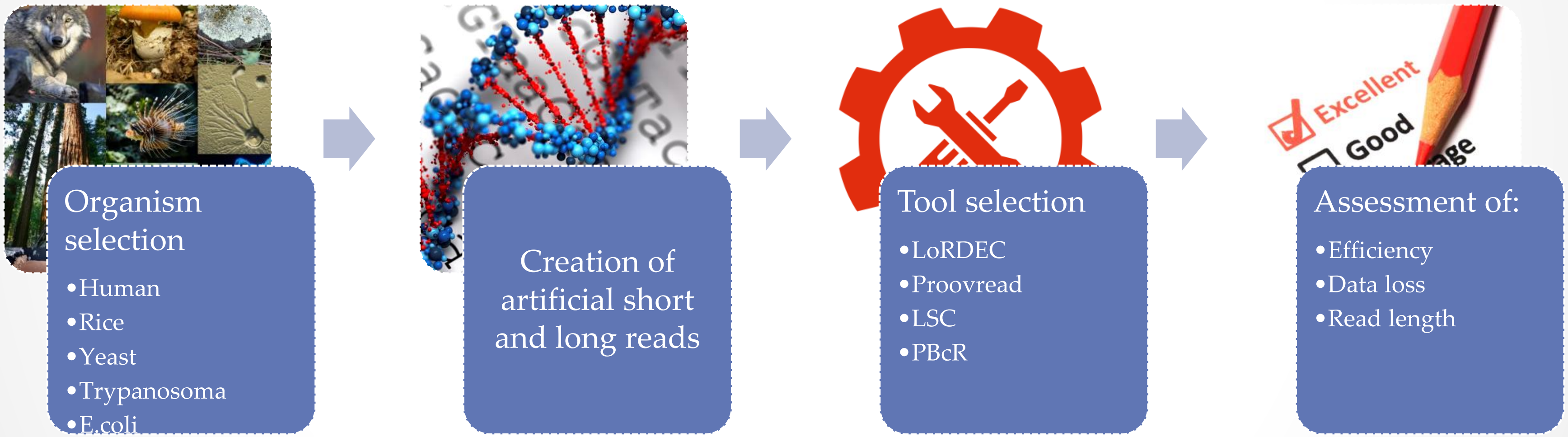
higher error rate, and higher cost per base

PacBio assembly usage example



Which tool to choose?

Assessment steps



Organism selection

- Human
- Rice
- Yeast
- Trypanosoma
- E.coli

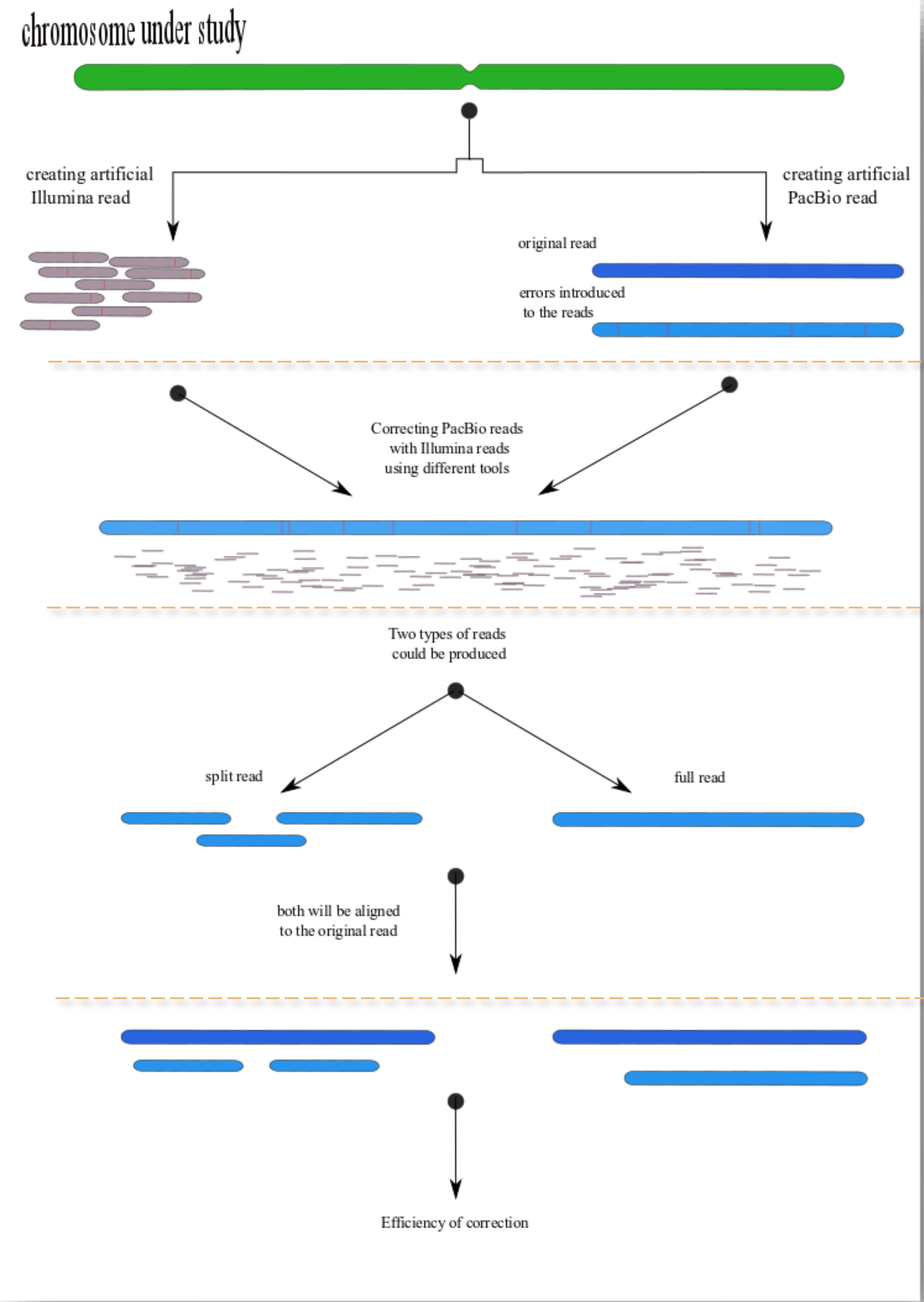
Creation of artificial short and long reads

Tool selection

- LoRDEC
- Proovread
- LSC
- PBcR

Assessment of:

- Efficiency
- Data loss
- Read length



chromosome under study

creating artificial Illumina read

creating artificial PacBio read

original read

errors introduced to the reads

1

Correcting PacBio reads with Illumina reads using different tools

2

Two types of reads could be produced

split read

full read

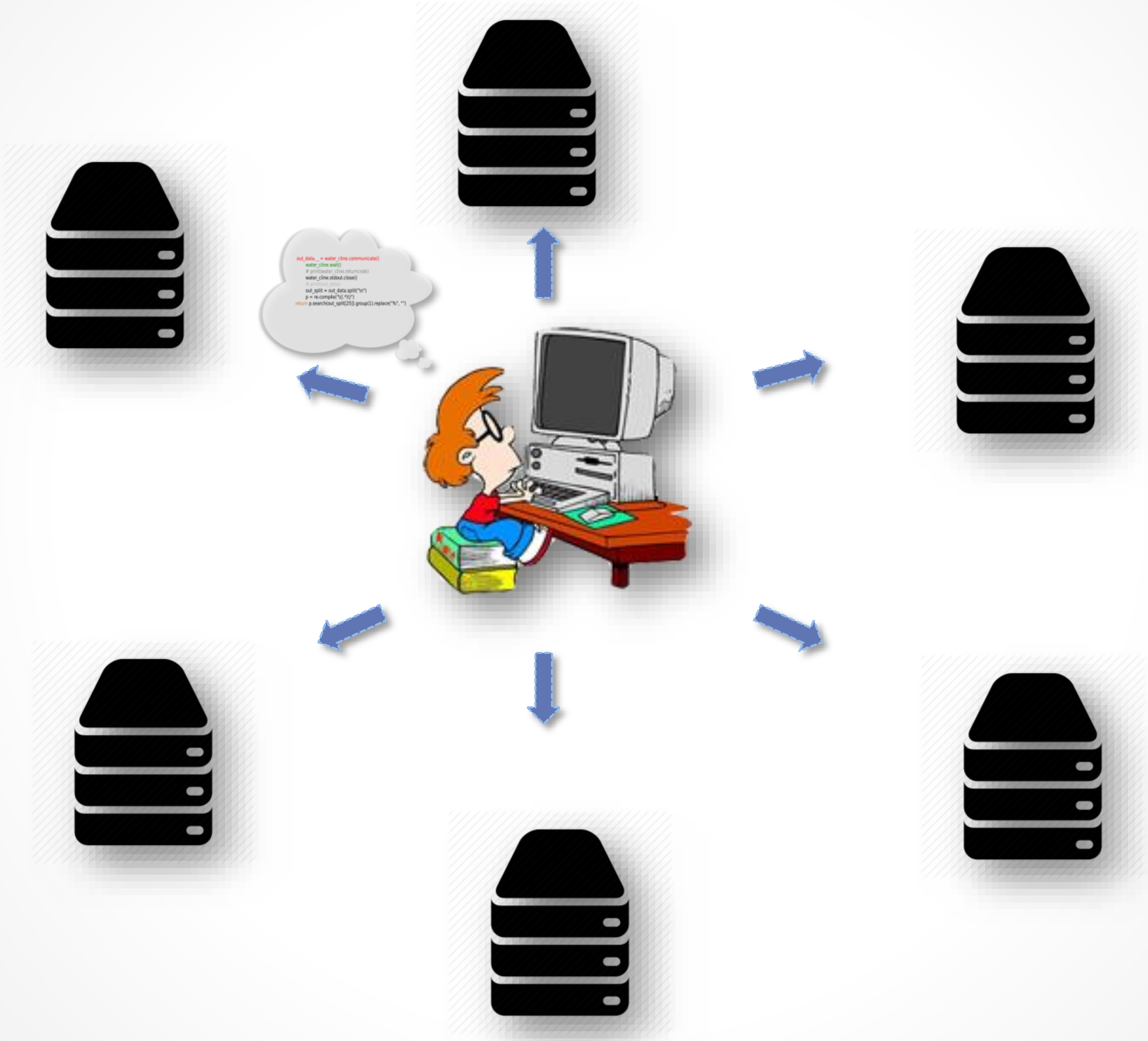
both will be aligned to the original read

3

4

Efficiency of correction

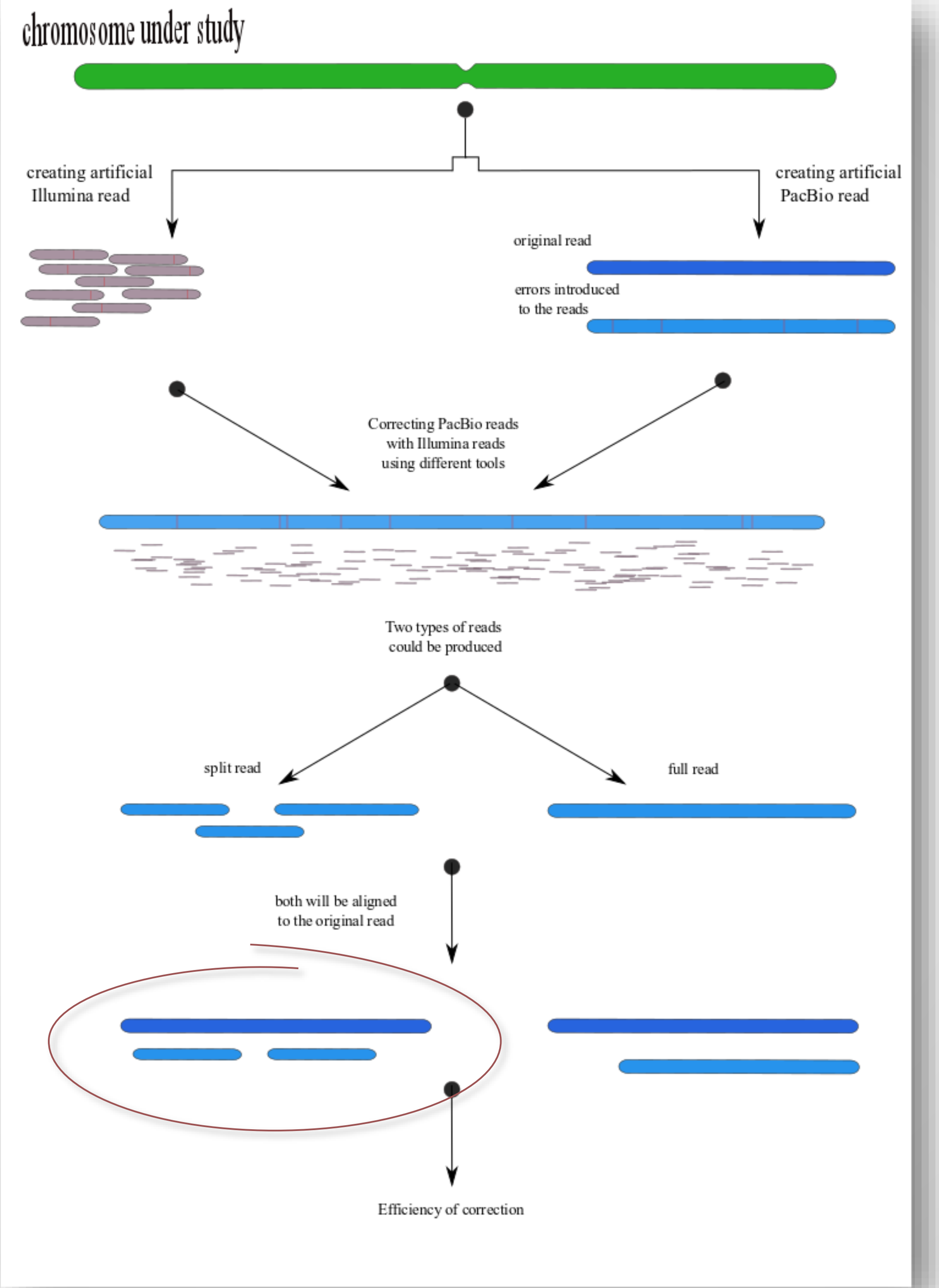
Eagle



```
out_data, _ = water_cline.communicate()
        water_cline.wait()
        # print(water_cline.returncode)
        water_cline.stdout.close()
        # print(out_data)
        out_split = out_data.split("\n")
        p = re.compile("\/(.*)\/")
return p.search(out_split[25]).group(1).replace("%", "")
```

The result

For the split reads



chromosome under study

creating artificial Illumina read

creating artificial PacBio read

original read

errors introduced to the reads

Correcting PacBio reads with Illumina reads using different tools

Two types of reads could be produced

split read

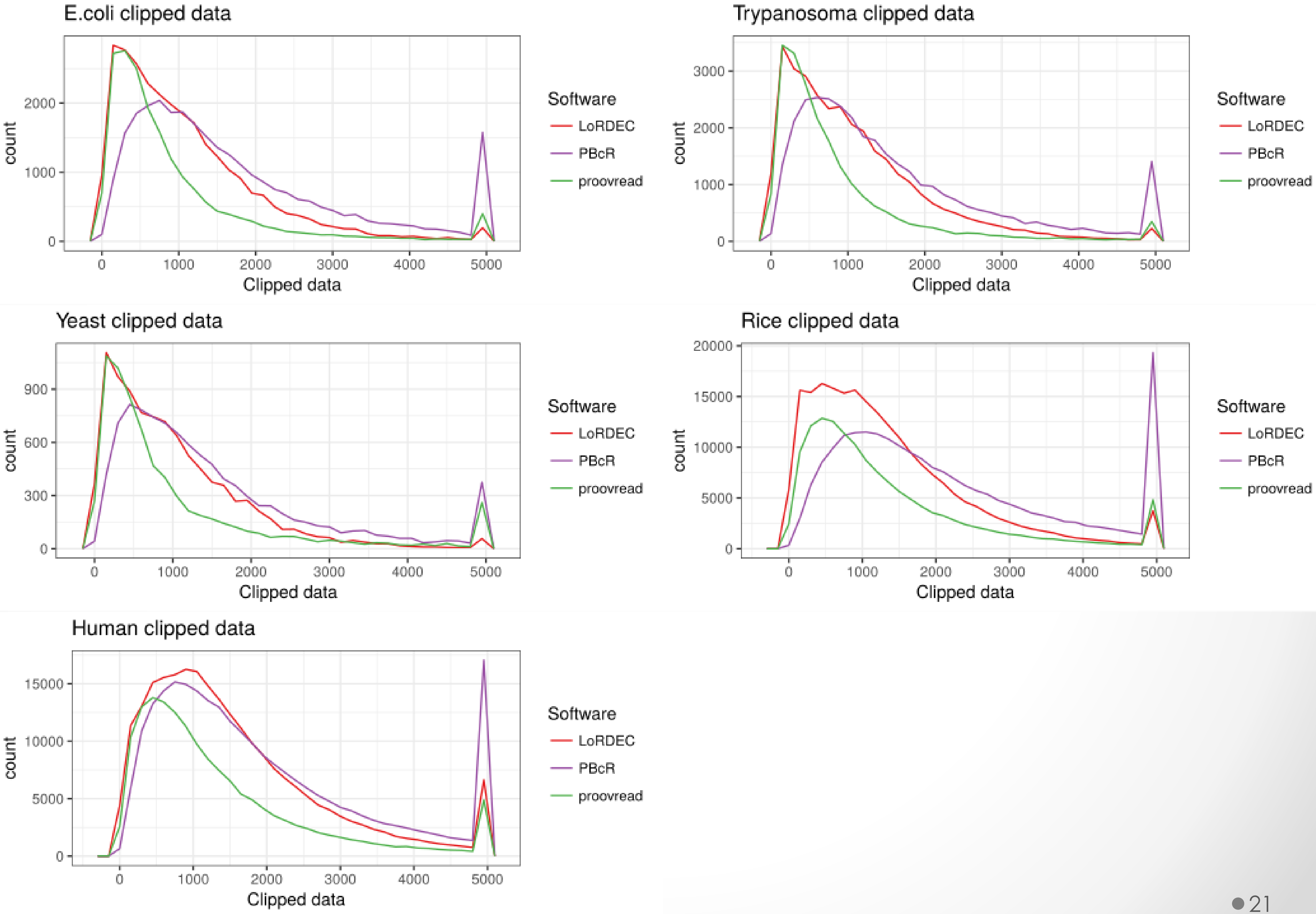
full read

both will be aligned to the original read

Efficiency of correction

Data loss

Clipped sequence distribution



E.coli clipped data

Trypanosoma clipped data

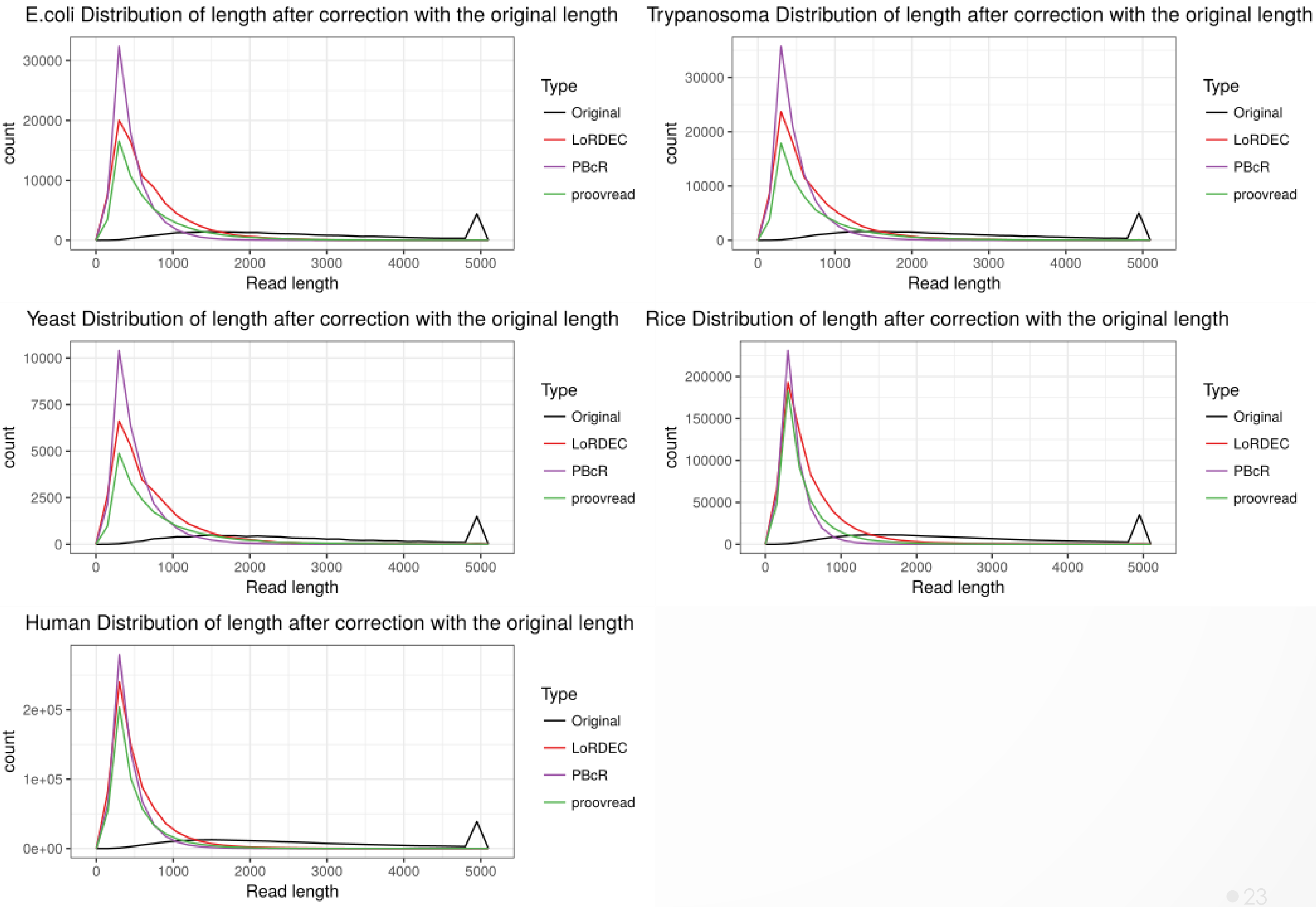
Yeast clipped data

Rice clipped data

Human clipped data

Read length count

Correction length with original



E.coli Distribution of length after correction with the original length

Trypanosoma Distribution of length after correction with the original length

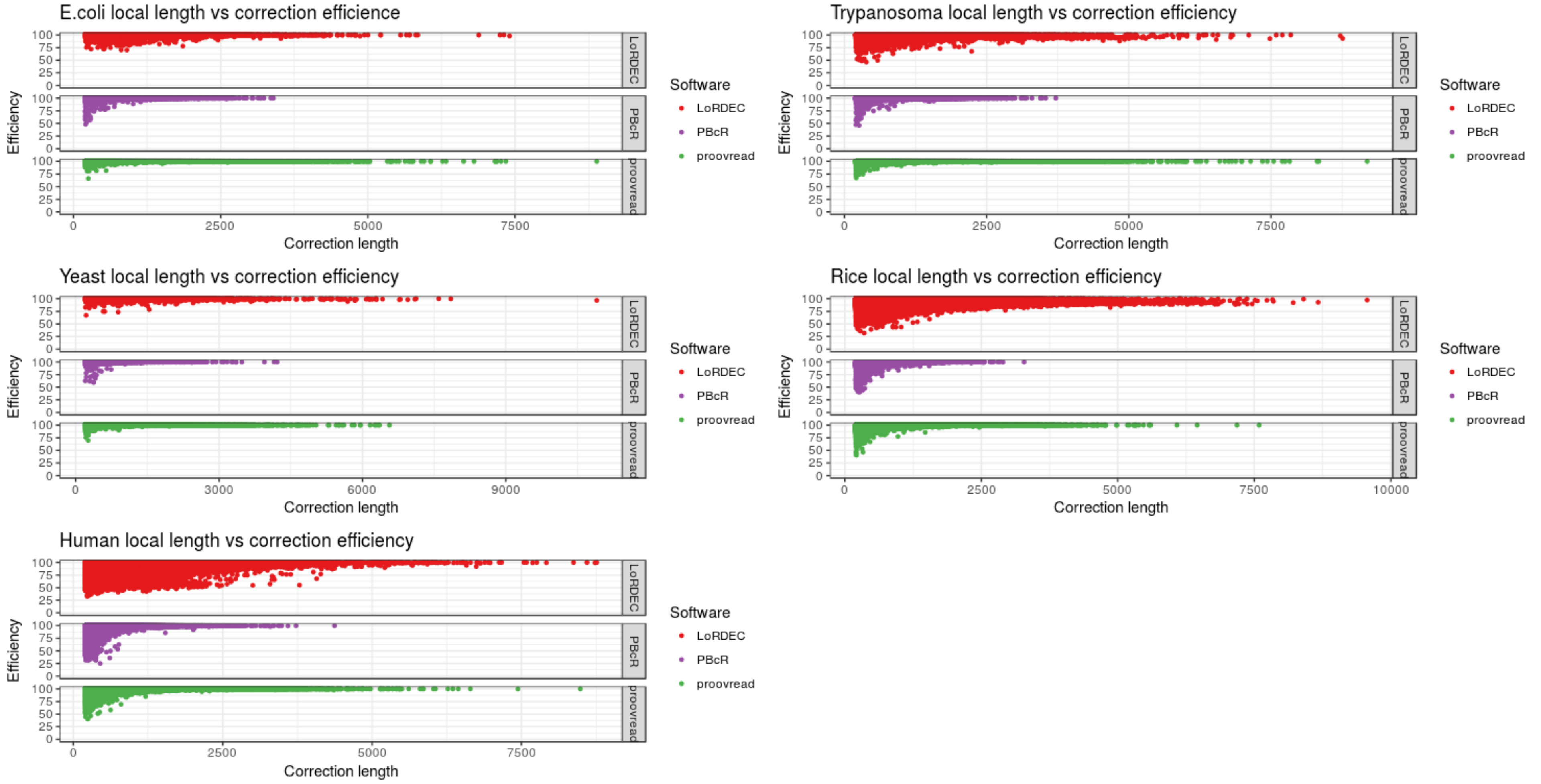
Yeast Distribution of length after correction with the original length

Rice Distribution of length after correction with the original length

Human Distribution of length after correction with the original length

Correction efficiency

Correction vs Length local


E.coli local length vs correction efficience

Trypanosoma local length vs correction efficiency

Yeast local length vs correction efficiency

Rice local length vs correction efficiency

Human local length vs correction efficiency

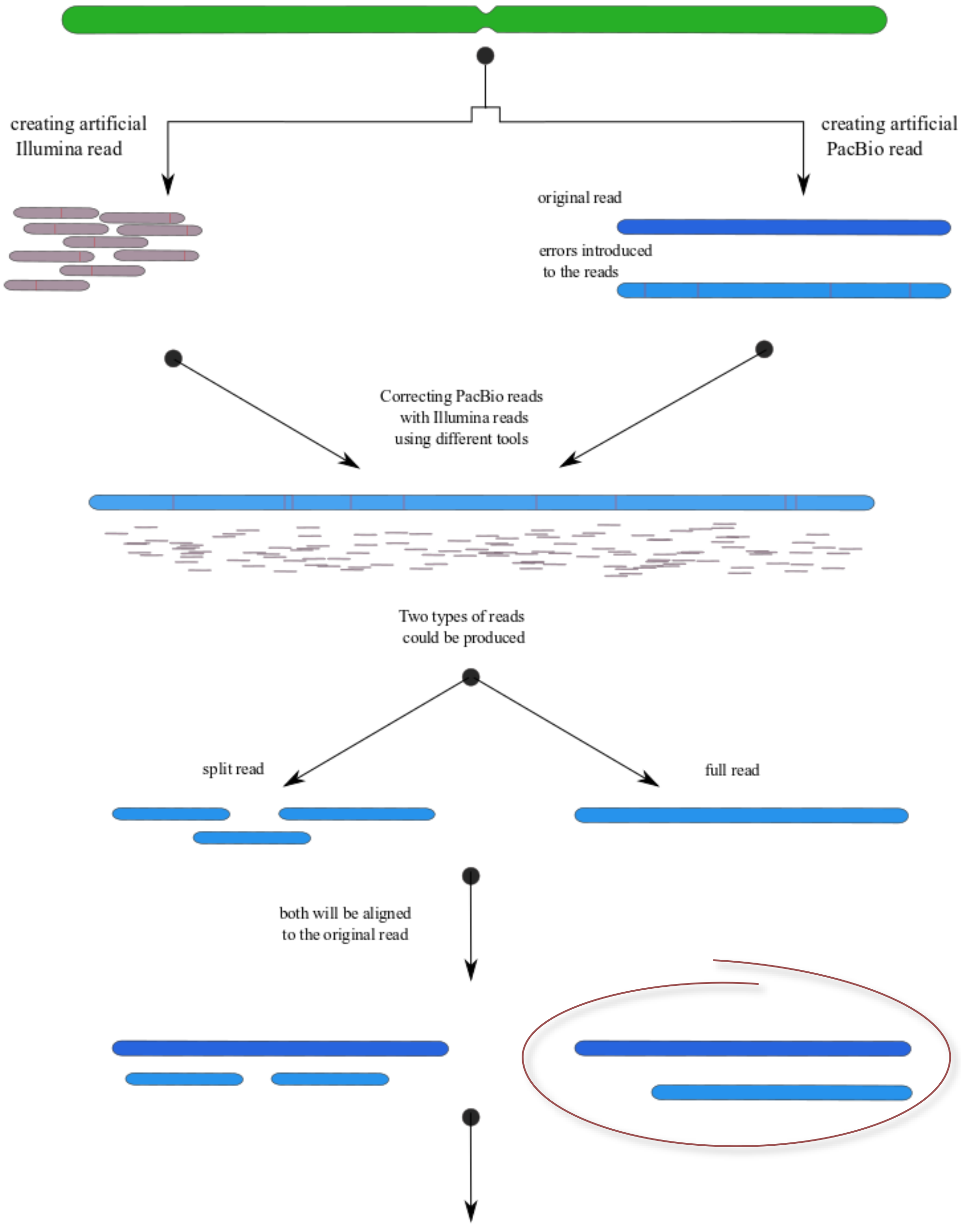
Split reads result conclusion

1st Proovread

The result

For the full reads

chromosome under study



creating artificial Illumina read

creating artificial PacBio read

original read

errors introduced to the reads

Correcting PacBio reads with Illumina reads using different tools

Two types of reads could be produced

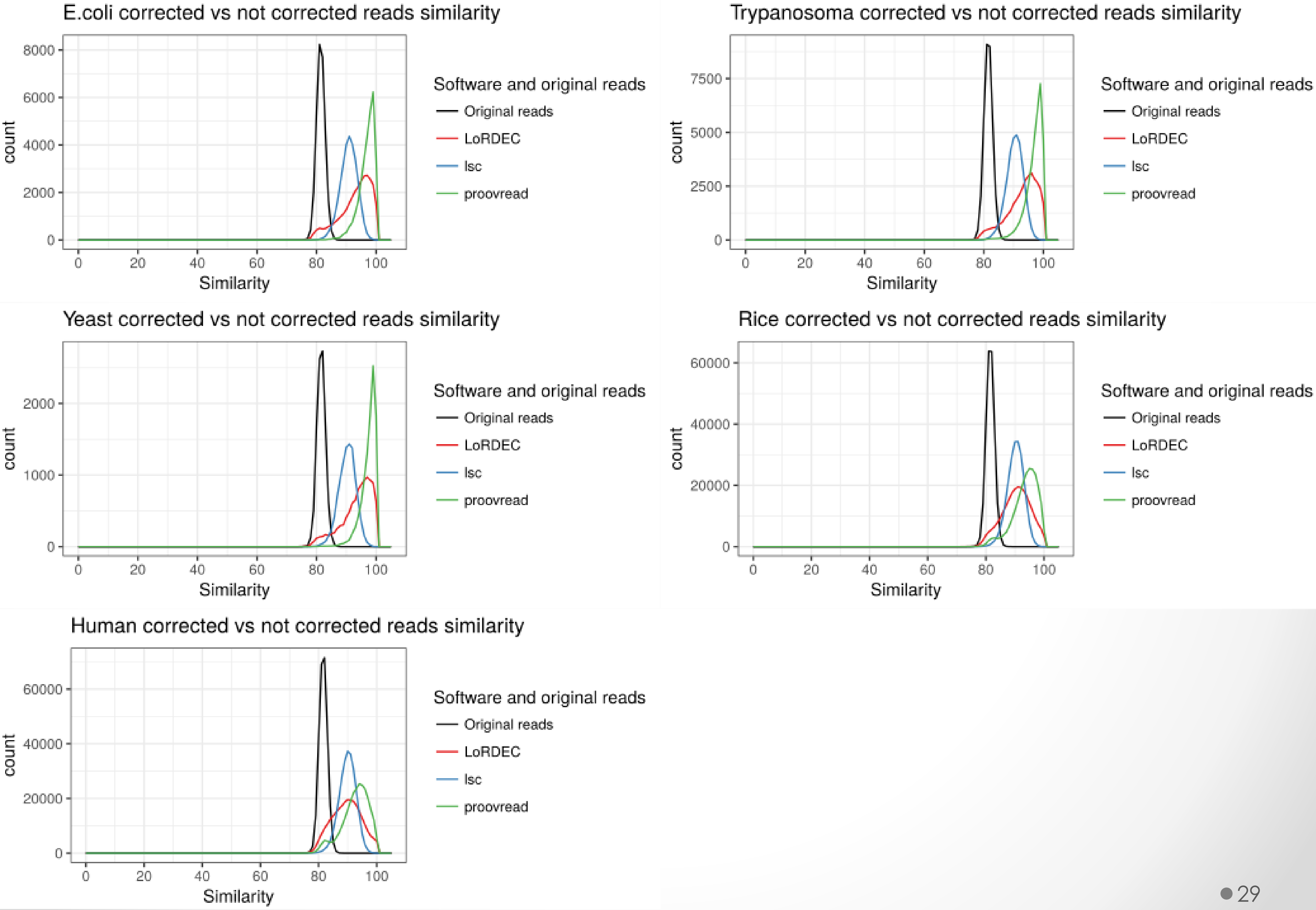
split read

full read

both will be aligned to the original read

Efficiency of correction

Similarity before and after correction



Full reads result conclusion

1st Proovread

Performance

	User	System	CPU	Total
Proovread	49302.87s (13.69h)	6811.14s (1.89h)	676%	2:18:11.62
LoRDEC	2943.50s (0.81h)	532.17s (0.14h)	790%	7:19.84
LSC	58944.97s (16.37h)	195.09s (0.05h)	934%	1:45:27.54
PBcR	20538.56s (5.70h)	1140.17s (0.31h)	383%	1:34:11.75

Chimera

- Read correction

- Genome assembly

ACKNOWLEDGEMENTS

Department of Computational Biology (AMU):

Prof. dr hab. Wojciech Karłowski
Dr Marek Żywicki

Department of Protein Biosynthesis (IBCh):

Prof. dr hab. Tomasz Twardowski
(The head of the project)
Dr Agata Tyczewska
Dr Joanna Gracz

UAM
ADAM MICKIEWICZ
UNIVERSITY
IN POZNAŃ

INSTITUTE OF BIOORGANIC CHEMISTRY POLISH ACADEMY OF SCIENCES

National
Science
Centre
Poland

PCSS

www.combio.pl