SeqBIM

Nanopore MinION long read sequencer: an overview of its error landscape

Clara DELAHAYE
PhD supervisor: Jacques NICOLAS

Nov. 23th 2020

IRISA - Team Genscale

Introduction

How does Nanopore sequencing works ?

Constraints linked to the basecaller

Our working basis

Nanopore: how does it work ?

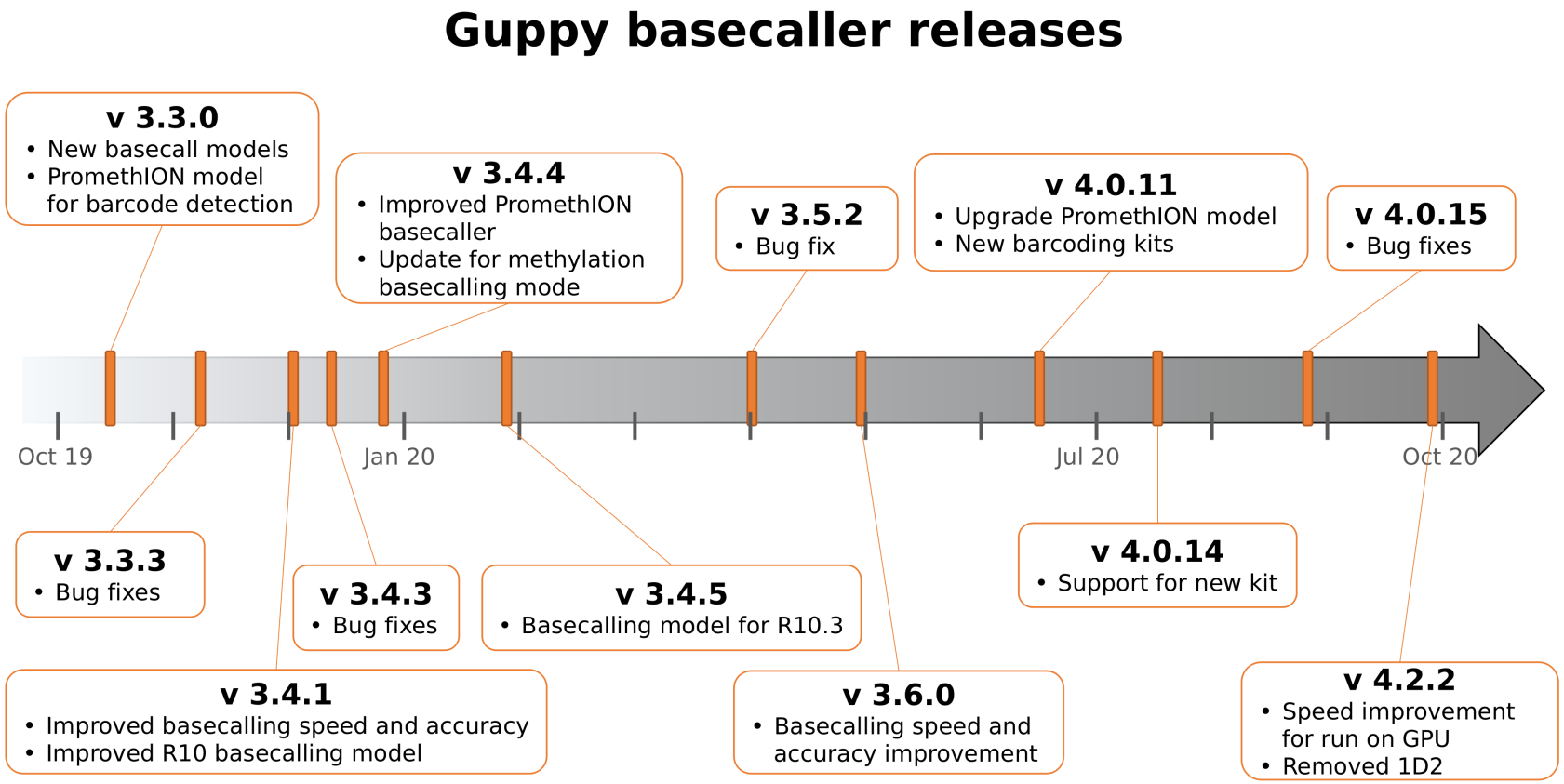
DNA passing through nanopore alters voltage.

The basecaller exploits these variations to retrieve DNA sequence.

Limitation: bunch of same bases $\rightarrow$ no variation.

Basecaller: a core element.

Basecaller: a core element. That is updated almost monthly



(+ Many other basecallers prior to Guppy [1] and to come.)

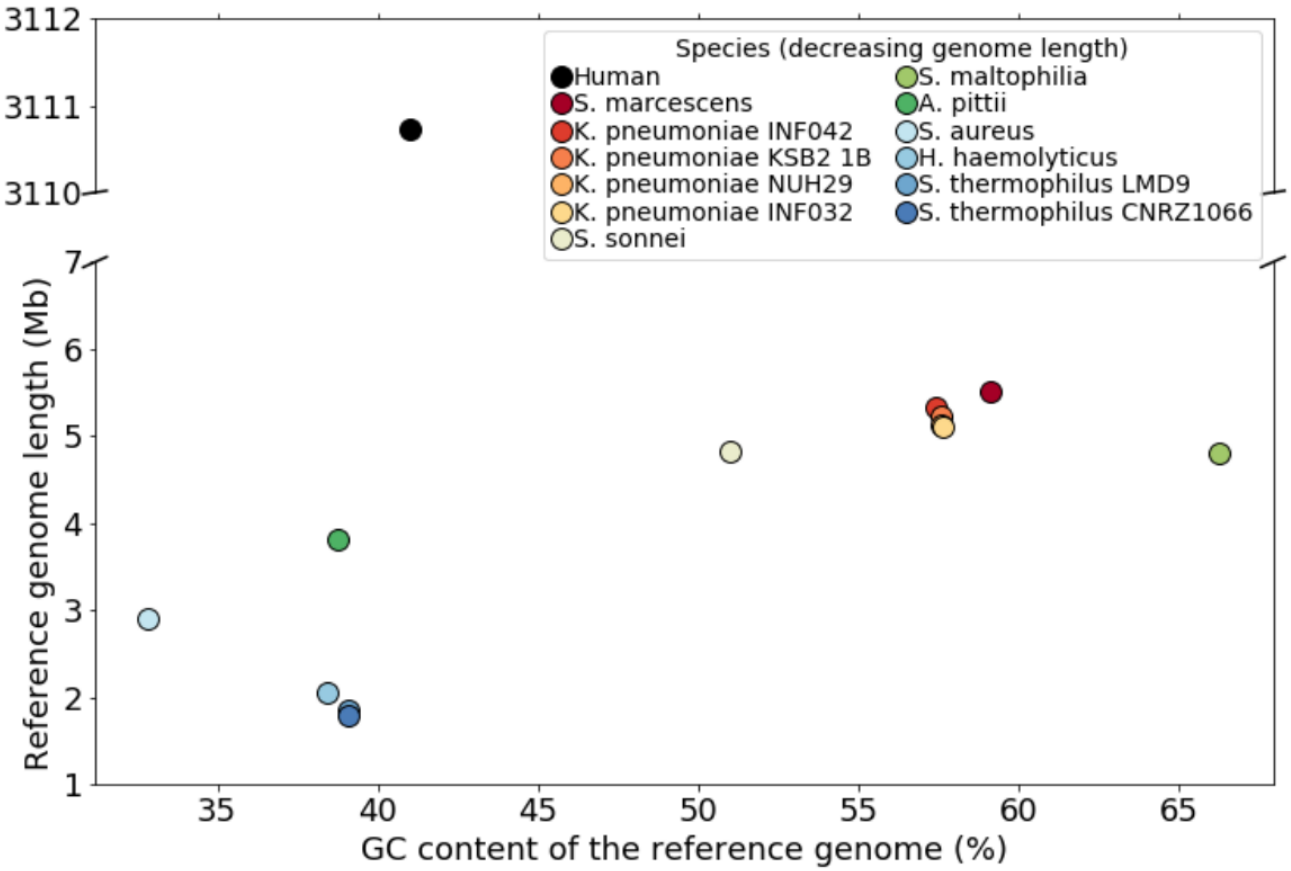
What we have been working on

- Raw reads from various sources, sequenced with the MinION
  - Our sequencing experiments [1]
  - Wick *et al.* 2019 [1]
  - Shafin *et al.* 2019 [2]

[1] Thanks E. Roux, STLO INRA/Agrocampus and INRIA

What we have been working on

- Raw reads from various sources, sequenced with the MinION
- 12 bacterial species with various GC content + 2 human datasets



What we have been working on

- Raw reads from various sources, sequenced with the MinION
- 12 bacterial species with various GC content + 2 human datasets
- **Guppy basecaller** (latest v4.2.2)
  - HAC mode: High ACcuracy
  - FAST mode: faster but higher error rate

What we have been working on

- Raw reads from various sources, sequenced with the MinION
- 12 bacterial species with various GC content + 2 human datasets
- Guppy basecaller (latest v4.2.2)

Goal: outline landscape of MinION long read sequencing errors

+ 2 preliminary steps:

- estimating error rates of reads
- evaluate HAC and FAST basecalling modes

Estimate error rate from raw reads

How to estimate error rates of reads?

Phred quality score

Estimate Nanopore error rates from quality

Estimate error rate

How to get error rate of reads ?

- Align reads on reference genome
- Count differences treated as sequencing errors

Estimate error rate

How to get error rate of reads ?

- Align reads on reference genome
- Count differences treated as sequencing errors

Problems

- Reference genome of poor quality (or even absent)
- Different strain / variants
- Requires to compute alignments

Estimate error rate

How to get error rate of reads ?

- Align reads on reference genome
- Count differences treated as sequencing errors

Problems

- Reference genome of poor quality (or even absent)
- Different strain / variants
- Requires to compute alignments

→ **Estimate error rates directly from raw reads ?**

Filtering reads on quality score

Phred quality score: confidence score for each sequenced base
Ranging from 0 to 93 (the higher the better)

Base	T	G	A	T	A	G	T	T	A	T	G
Score	32	40	41	35	29	23	26	32	36	32	14
ASCII	A	I	J	D	>	8	;	A	E	A	/

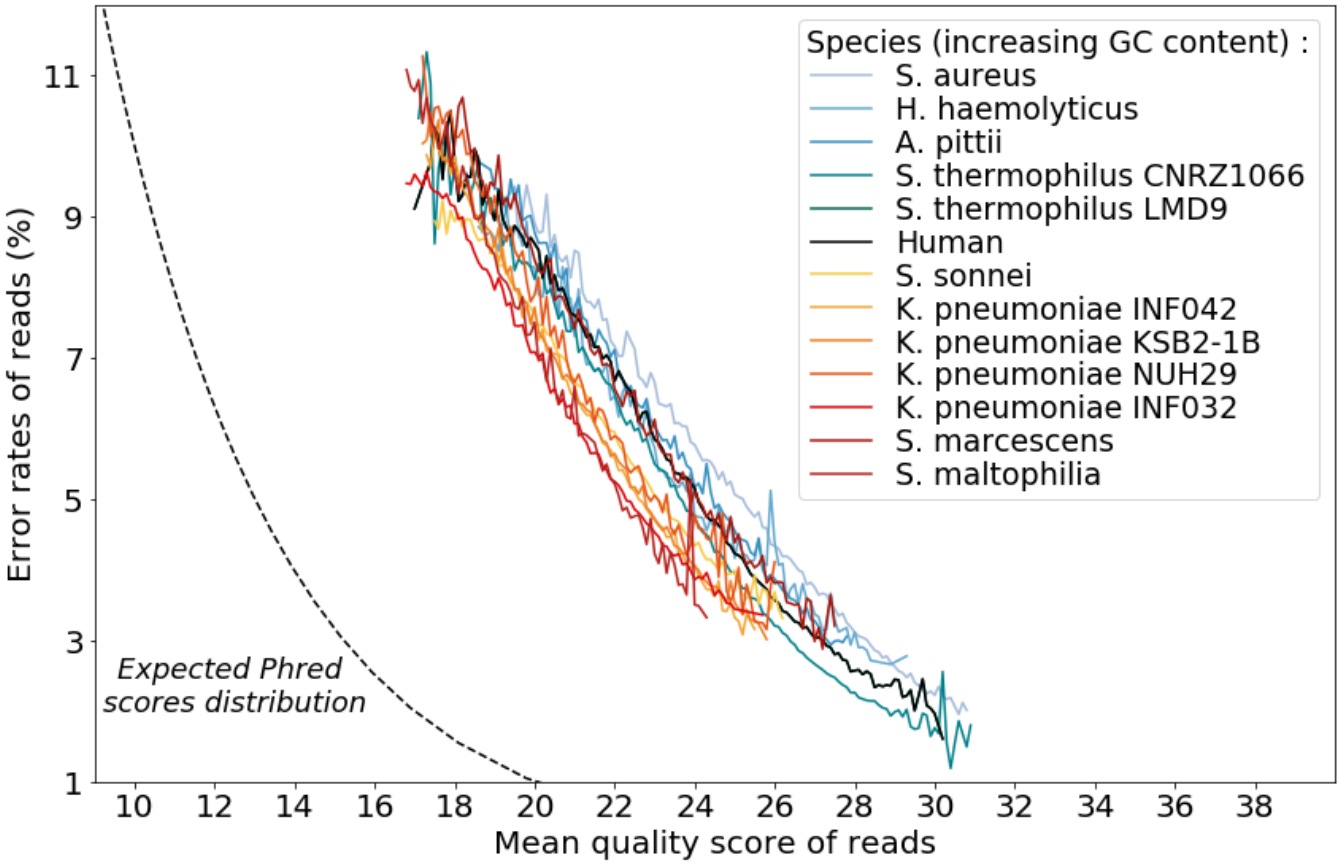
In FASTQ files scores are encoded in ASCII characters

Score indicates probability **P** of a wrong base:

$$\boxed{P = 10^{\frac{-Q}{10}}}$$

Phred score of 10 $\leftrightarrow$ 10% error rate ; score of 20 $\leftrightarrow$ 1% error rate

Estimate error rate from quality score



Nanopore quality score (Q) does not follow Phred scores

Yet enables to **estimate error rate (E)** (locally and at read level):

$E = 0.015Q^2 - 1.15Q + 24$

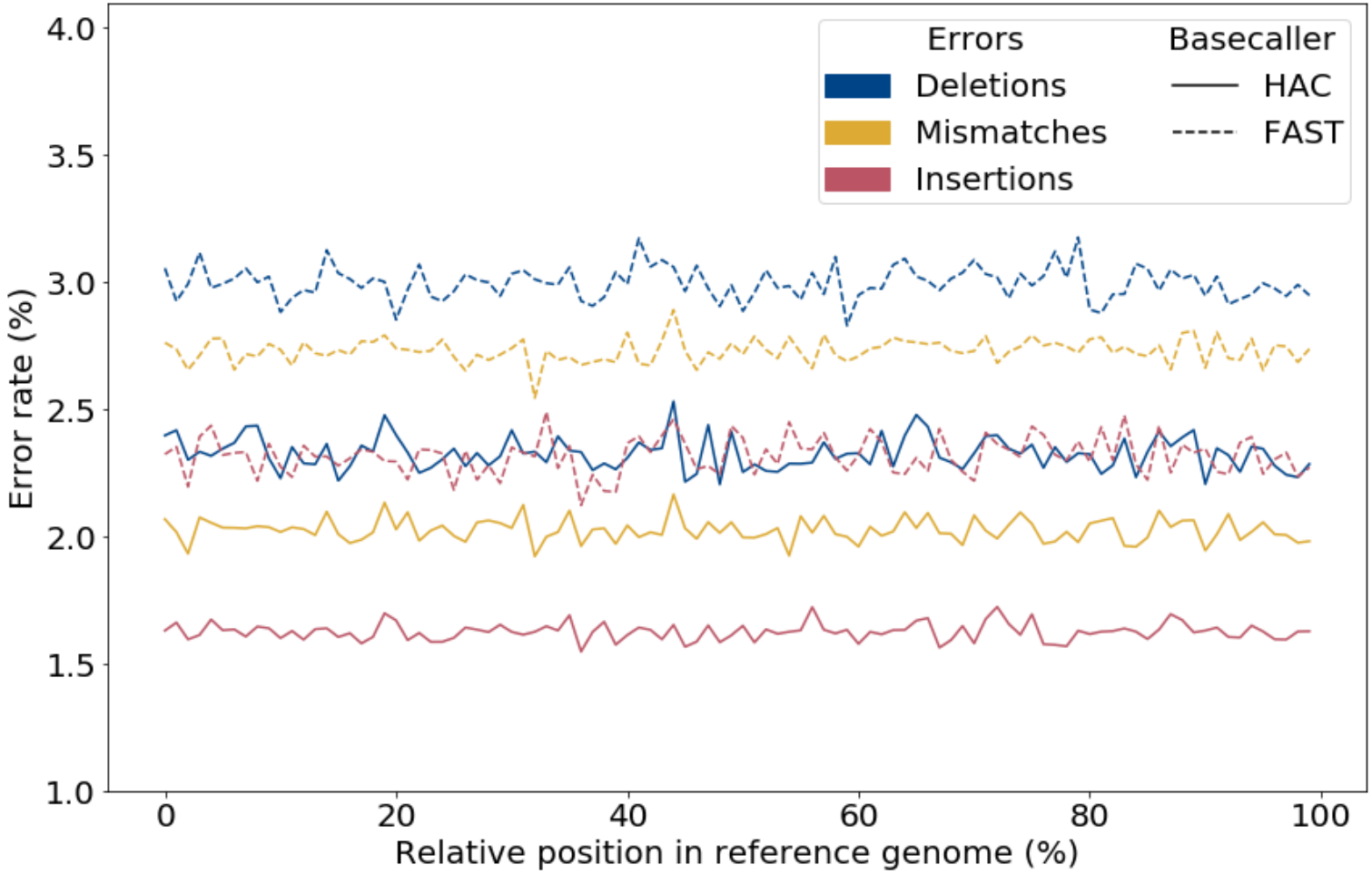
Comparison between two main modes of the basecaller: HAC *vs* FAST

Error rate

Calling homopolymers

Running time

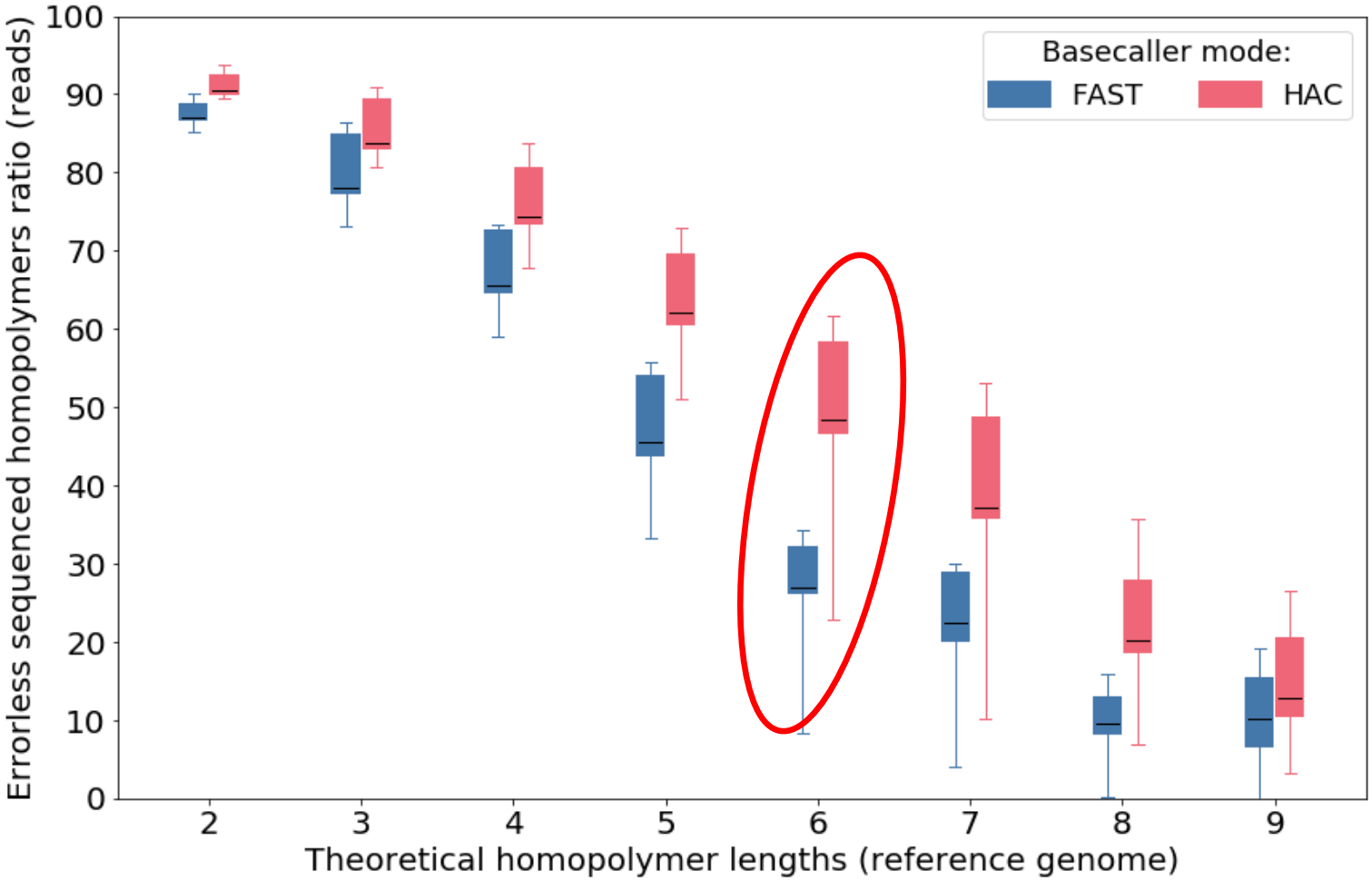
HAC *vs* FAST - Error rates



Similar error profile : deletions > mismatches > insertions

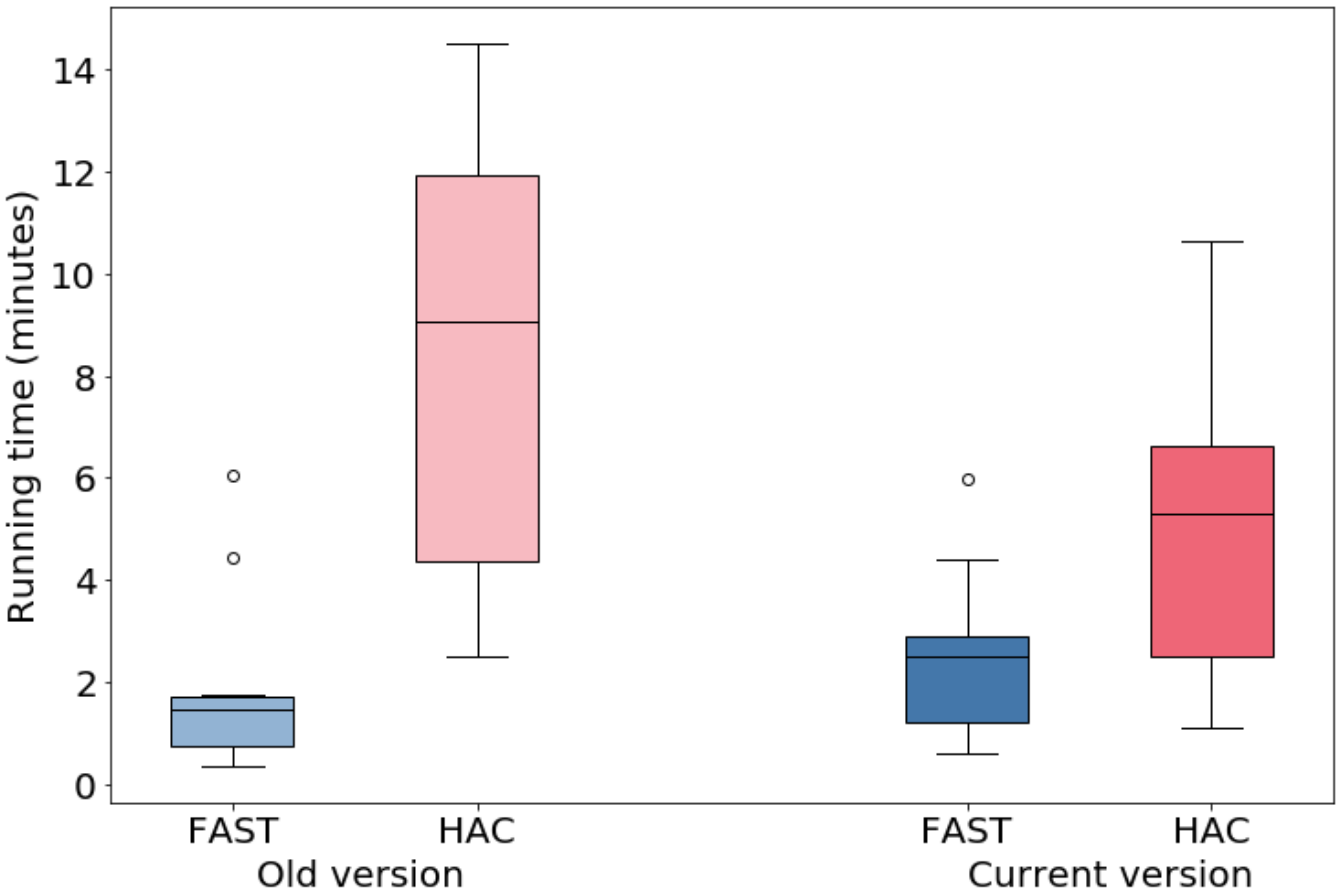
HAC mode reduces error rate by 2%

HAC *vs* FAST - Calling homopolymers



HAC mode basecalls homopolymers up to twice better than FAST

HAC *vs* FAST - Running times (now and before)

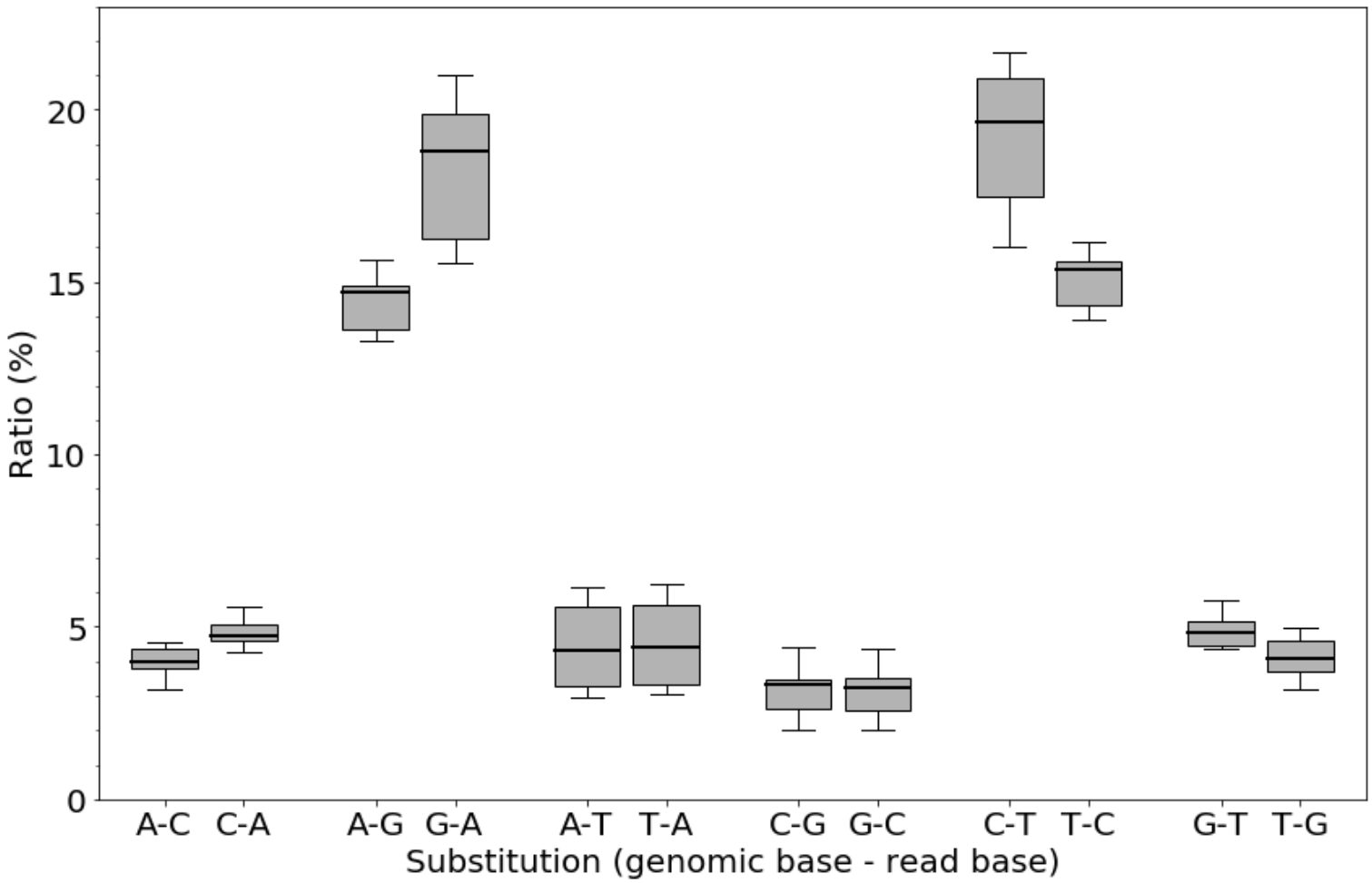


FAST mode is **only about 2 times faster** now

FAST mode is no longer of practical interest

Bias in substitution errors

Substitution errors bias



Transitions are three times more frequent than transversions

Non-symmetrical bias: decrease of reads' GC content

(-0.2% compared to reference genome)

GC bias

Depth of coverage

Error rate

Reads quality

GC bias - Depth of coverage

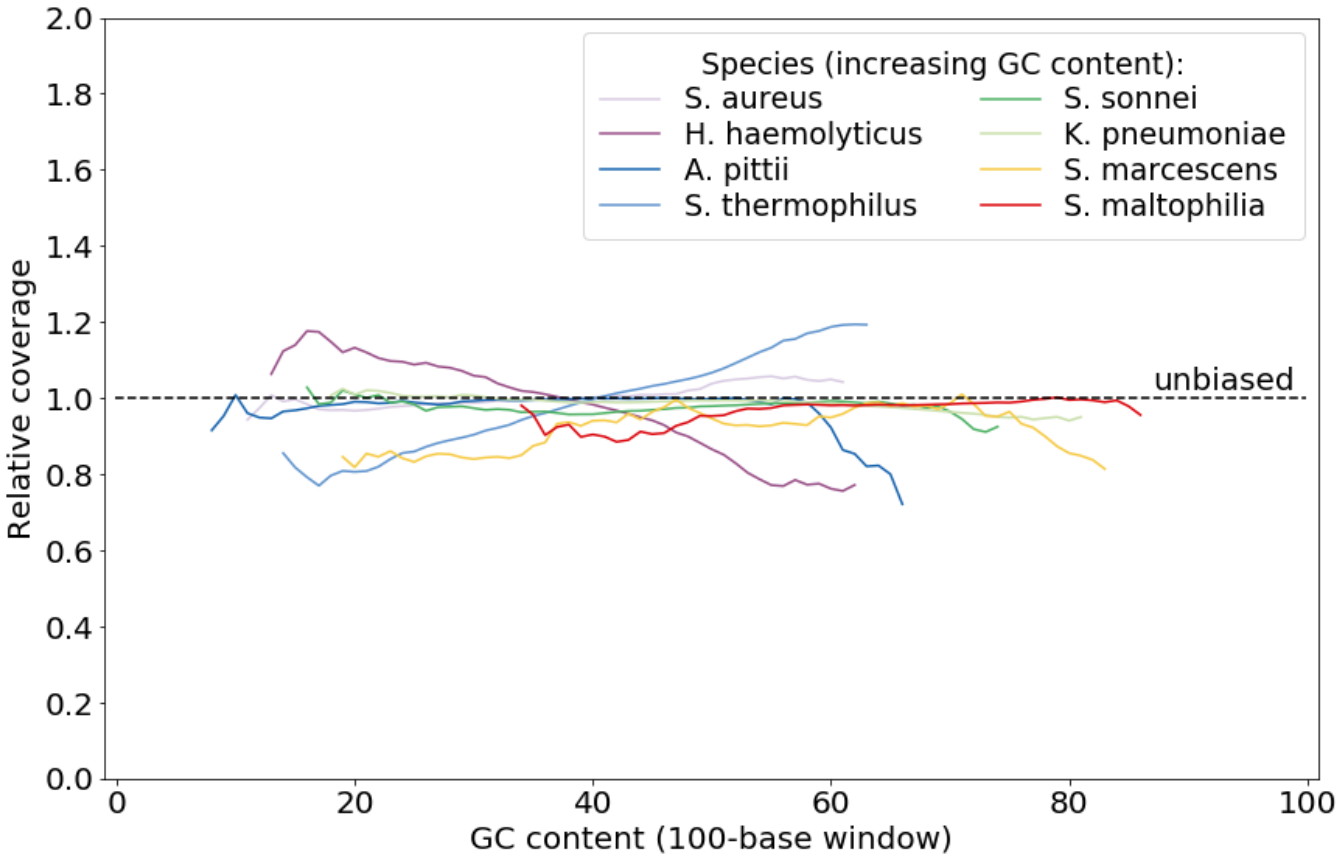
NGS data are subject to GC-bias due to PCR amplification

What about Nanopore data (no PCR needed) ?

GC bias - Depth of coverage

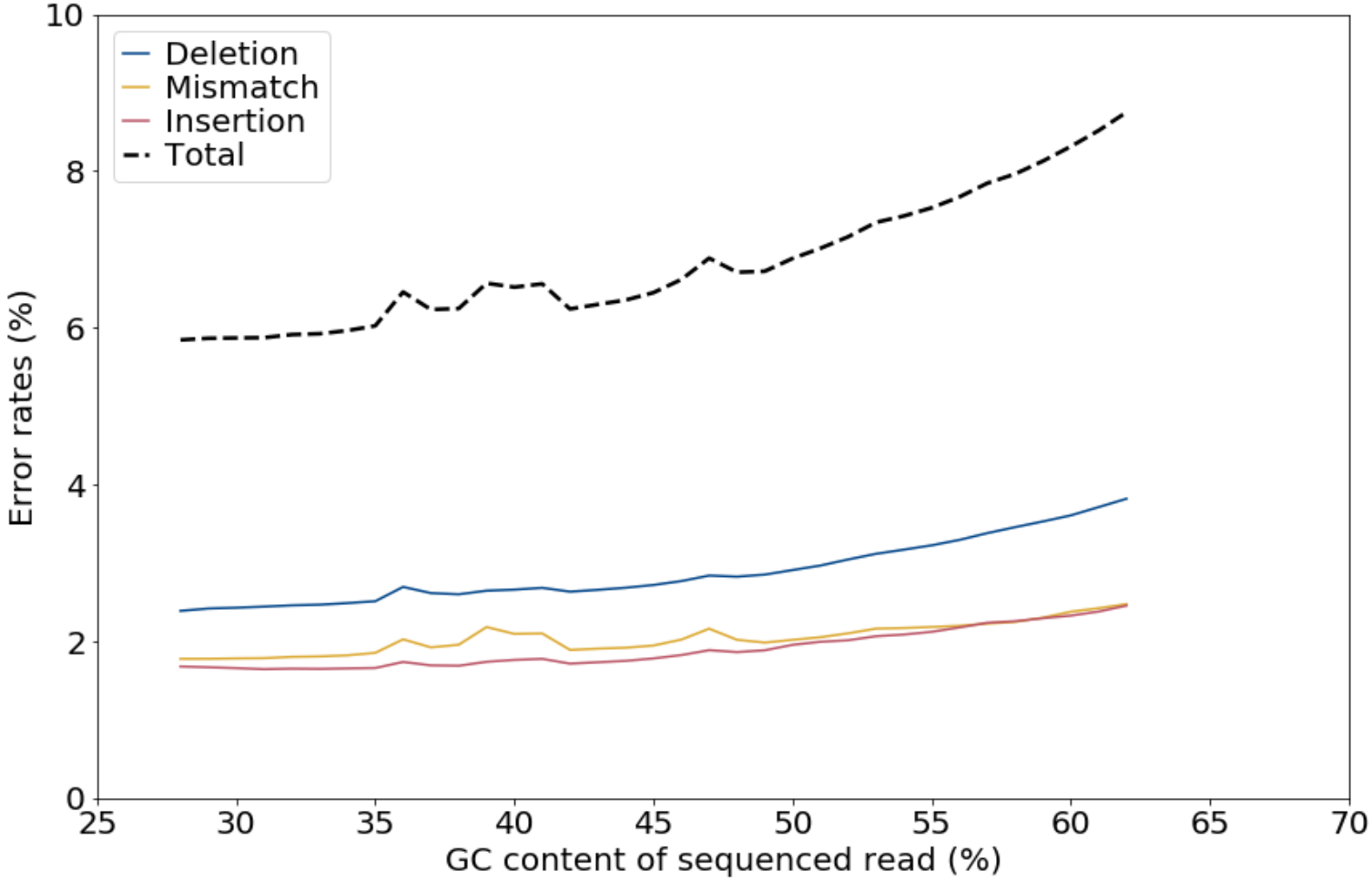
NGS data are subject to GC-bias due to PCR amplification

What about Nanopore data (no PCR needed) ?



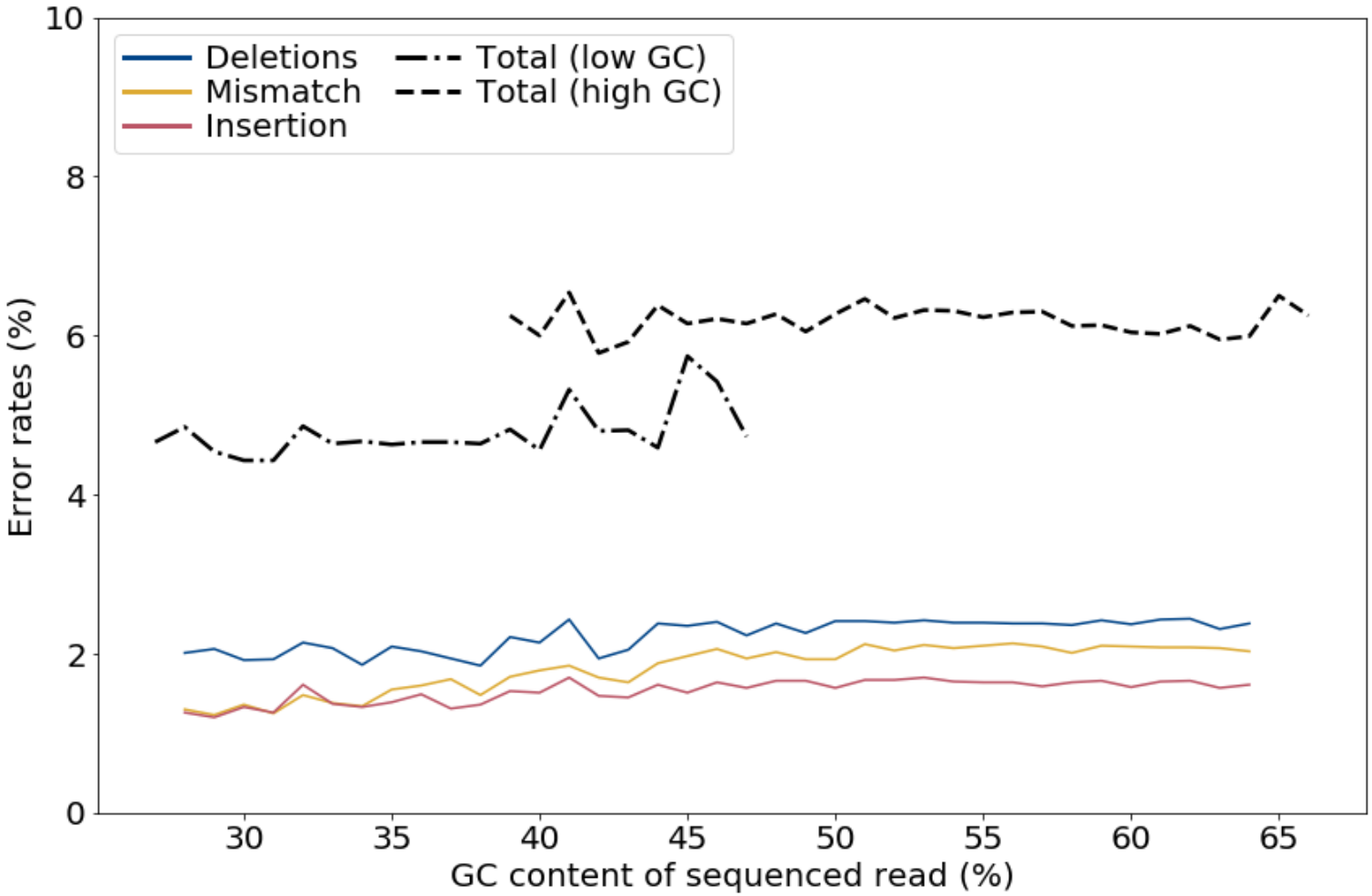
Roughly no bias, but drop for high GC + 2 species more impacted

GC bias - Error rate (human)



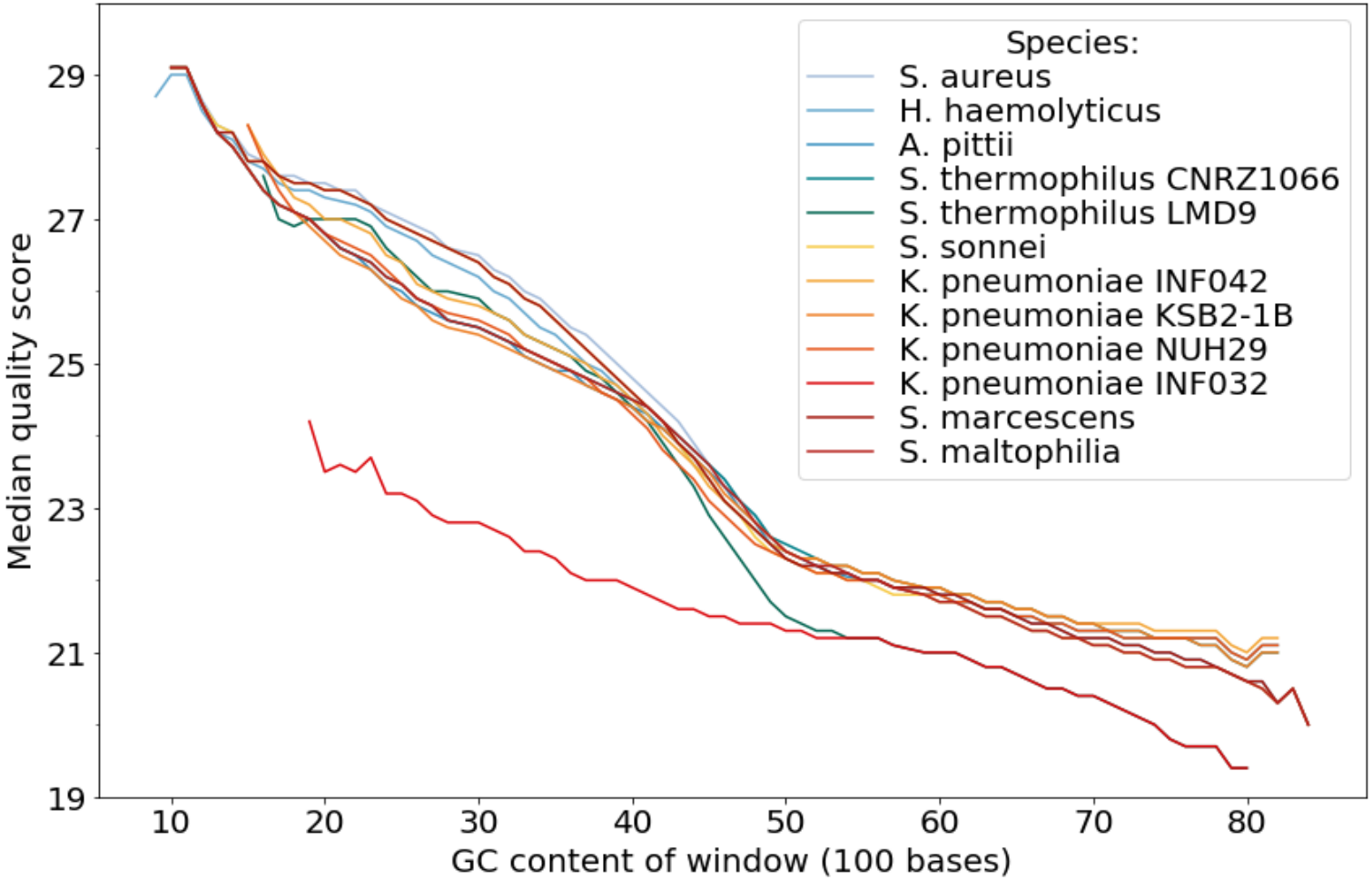
Low GC reads have about 1.5% lower error rates

GC bias - Error rate (bacteria)



Low GC species have about 1.5% lower error rates

GC bias - Reads quality



Quality decreases with GC content + break around 50%

Sequencing low complexity regions

Homopolymers $(X)^n$

Heteropolymers $(XY)^n$

Trinucleotides $(XYZ)^n$

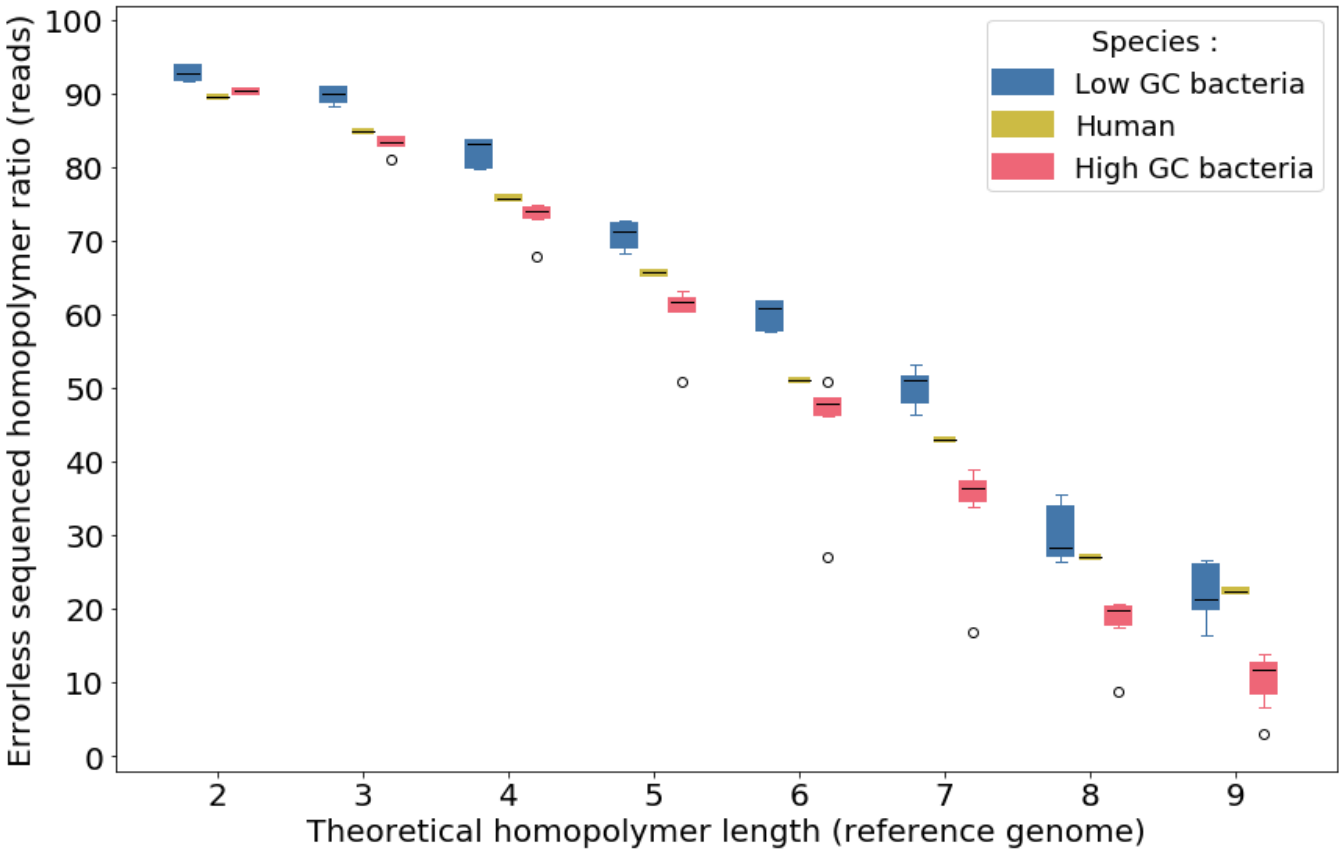
Share of sequencing errors due to homopolymers

	Mismatches	Insertions	Deletions	Global
Bacteria	52.61	18.95	55.66	**44.59**
Human	56.57	25.07	60.51	**49.43**

Table 1: Ratio (%) of sequencing errors induced by homopolymers over all sequencing errors

Near 50% of sequencing errors are linked to homopolymers
(mostly deletions and mismatches)

Homopolymer sequencing accuracy

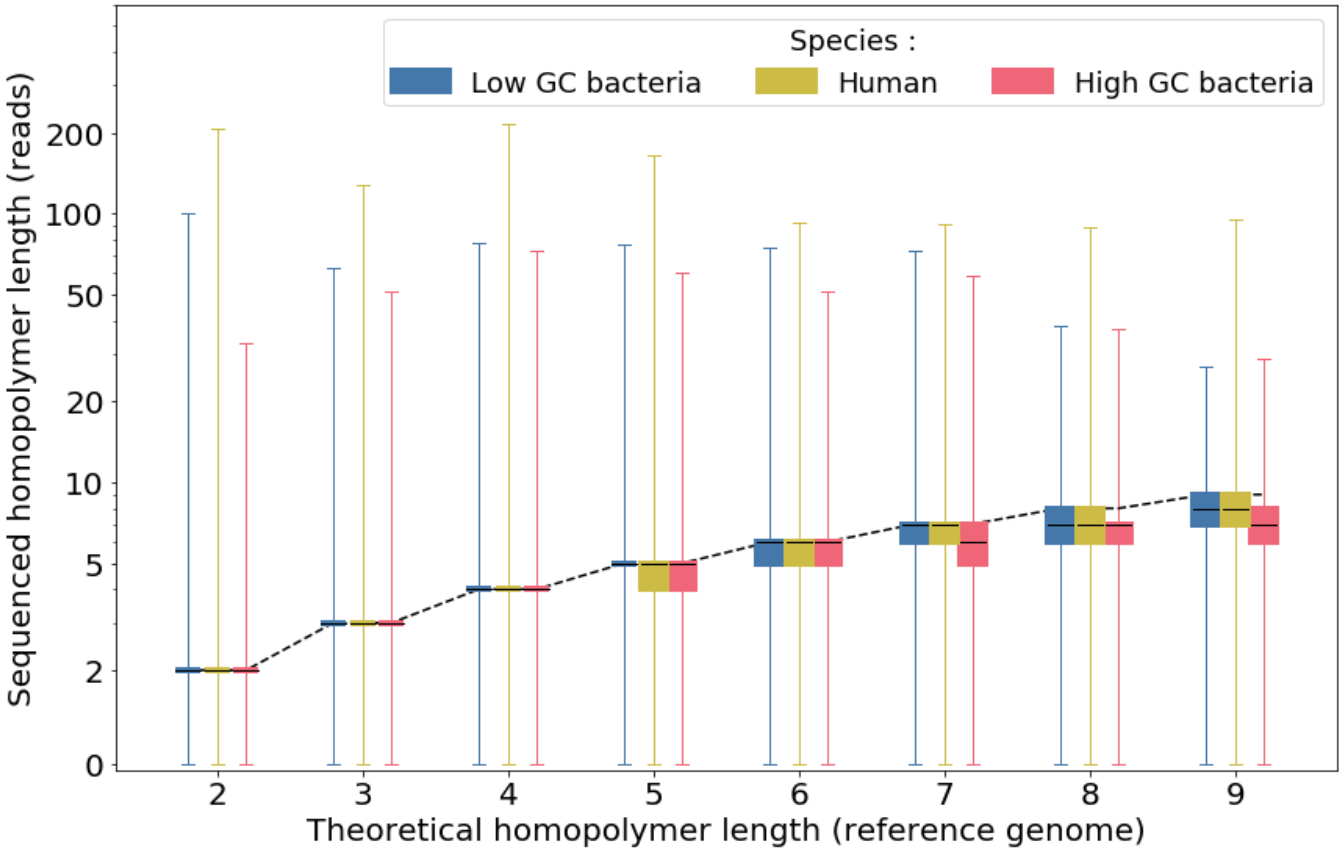


Rather good accuracy (>70%) for **short homopolymers** (< 4-5 bases)

Then **drop for higher lengths**: only 25% for length 8

Better results for **low GC**

Estimating homopolymer lengths

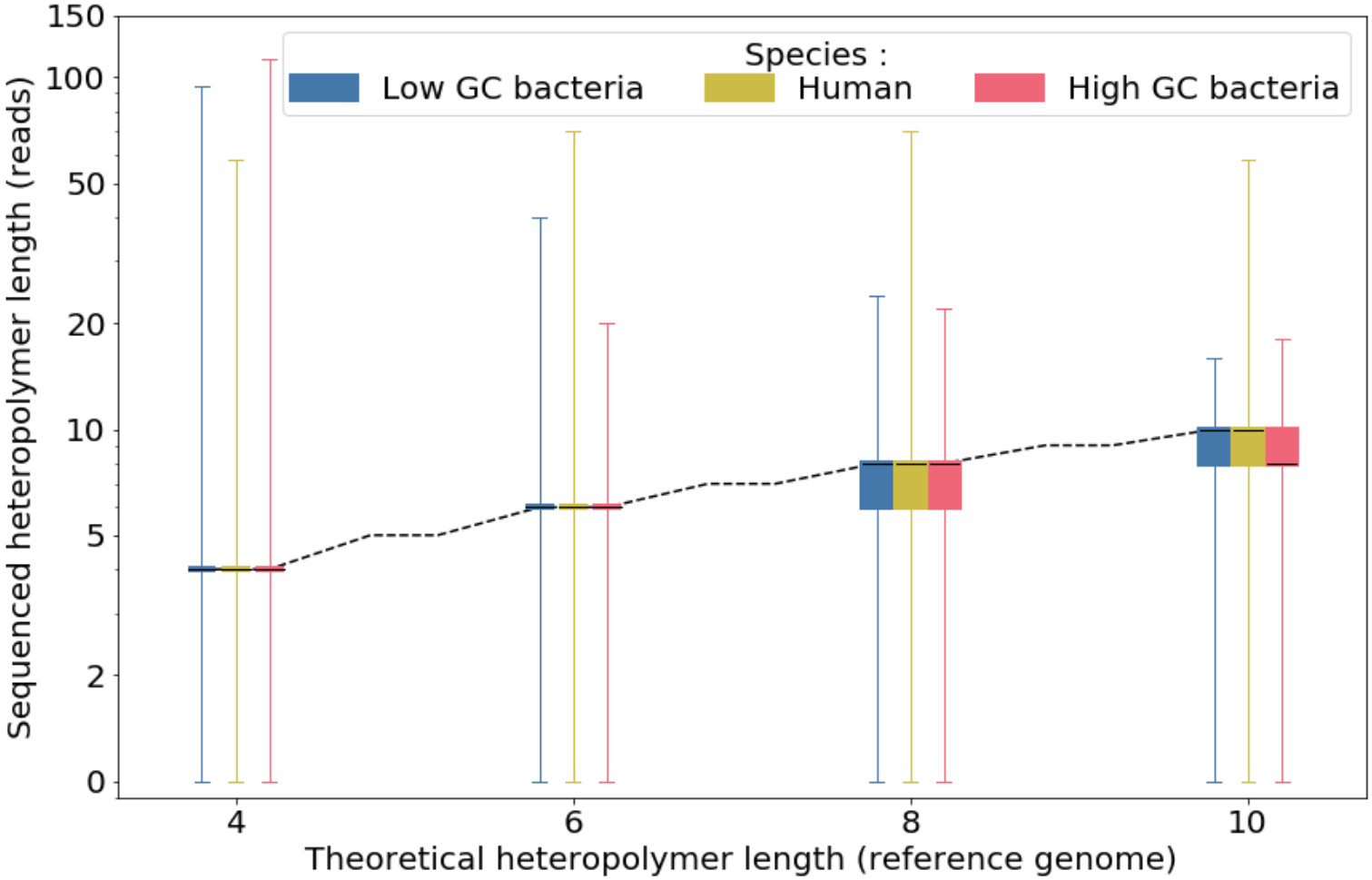


Rather good accuracy for **short homopolymers** (< 5 bases)

Underestimated length for longer homopolymers

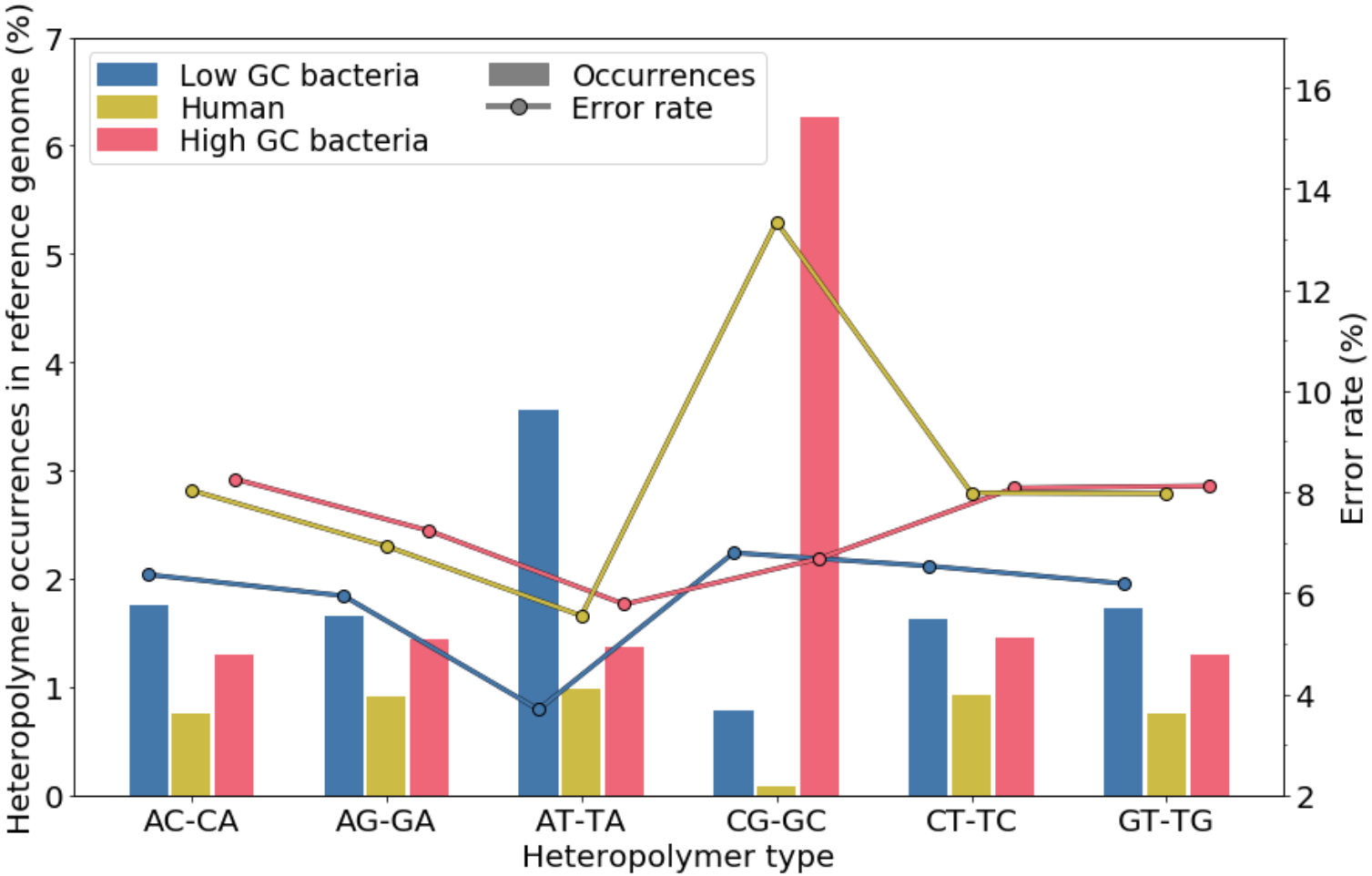
High variability: overestimation of one order of magnitude

Estimating heteropolymer lengths



Same trends as for homopolymers

Heteropolymer sequencing accuracy



CpG islands are marginal, except for high GC species

Error rate proportional to species GC rate (inverted for CpG islands)

Minimal error rate for AT-TA heteropolymers

Trinucleotides

Biological relevance: trinucleotide repeat expansion (mutation implied in several genetic diseases)

Notation:

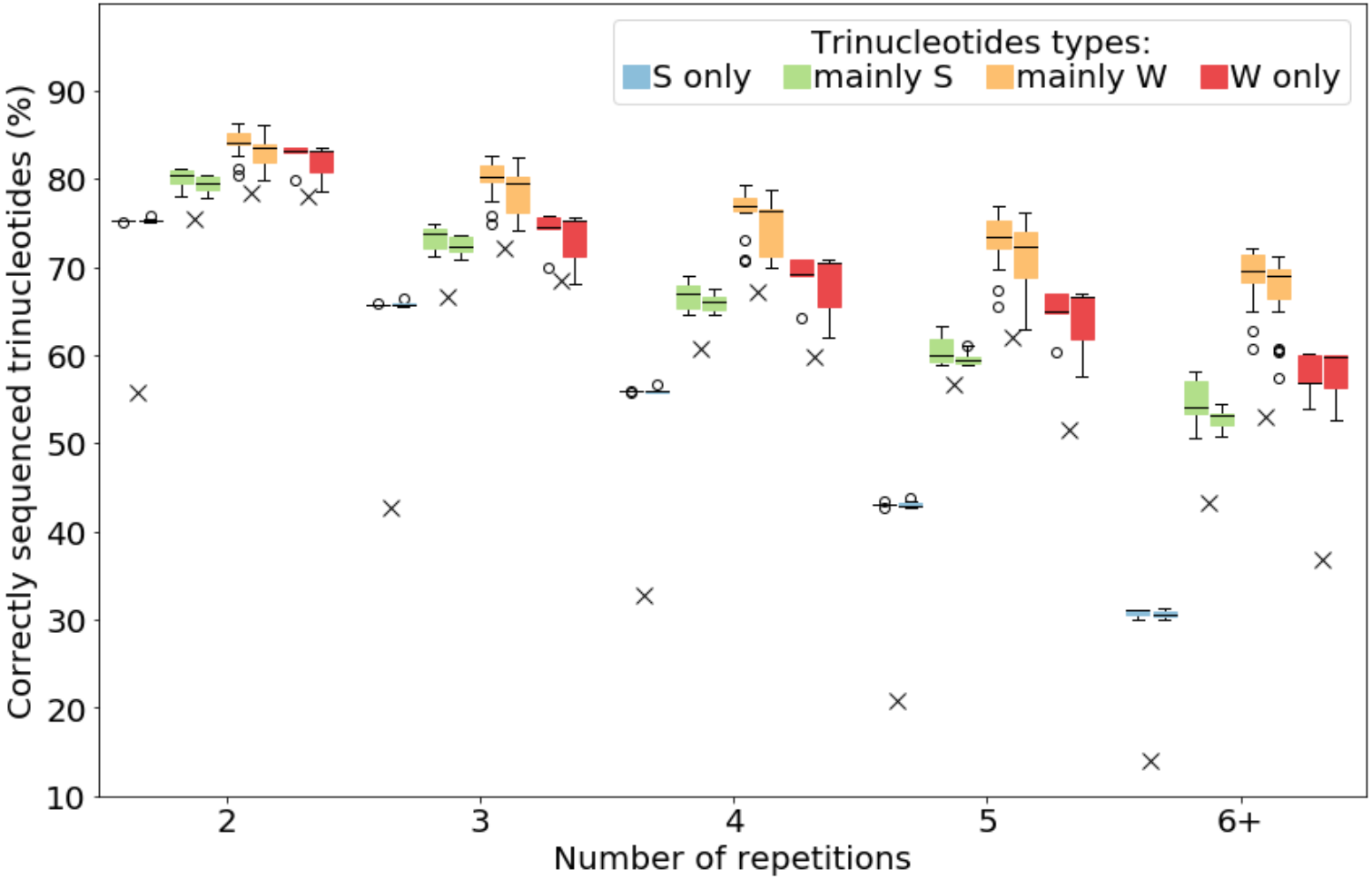
- S (Strong) for C and G bases
- W (Weak) for A and T bases

For example: (homopolymers removed as analysed in previous part)

- WWW = {AAT, ATA, TAA, ATT, TAT, TTA}
- WSW = {ACA, ACT, AGA, AGT, TCA, TCT, TGA, TGT}

→ **S only** (SSS), **mainly S** (SSW, SWS, WSS),
W only (WWW) and **mainly W** (WWS, WSW, SWW)

Trinucleotides sequencing accuracy



S-only are the worst (GC content) but W-only are not the best !

Low GC trinucleotides are better sequenced

Decrease of accuracy with increase of repetition number

Conclusions and prospects

Conclusions and prospects

- estimate error rate from quality score
- use HAC basecalling mode rather than FAST one
- substitution bias
- GC bias: no depth bias but error bias
- low complexity regions

Few papers addressing error profile of Nanopore sequencing
$\hookrightarrow$ Provide basis for future work on basecallers, read correction or assembly algorithms.

References i

- Wick, R. R., Judd, L. M., and Holt, K. E. (2019) Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biology,* **20**(1), 129.
- Shafin, K., Pesout, T., Lorig-Roach, R., Haukness, M., Olsen, H. E., Bosworth, C., Armstrong, J., Tigyi, K., Maurer, N., Koren, S., Sedlazeck, F. J., Marschall, T., Mayes, S., Costa, V., Zook, J. M., Liu, K. J., Kilburn, D., Sorensen, M., Munson, K. M., Vollger, M. R., Eichler, E. E., Salama, S., Haussler, D., Green, R. E., Akeson, M., Phillippy, A., Miga, K. H., Carnevali, P., Jain, M., and Paten, B. (2019) Efficient de novo assembly of eleven human genomes using PromethION sequencing and a novel nanopore toolkit. *bioRxiv,*.