Estimated nucleotide reconstruction quality symbols of basecalling tools for Oxford Nanopore sequencing

Wiktor Kuśmirek

Warsaw University of Technology, Institute of Computer Science, Warsaw, Poland

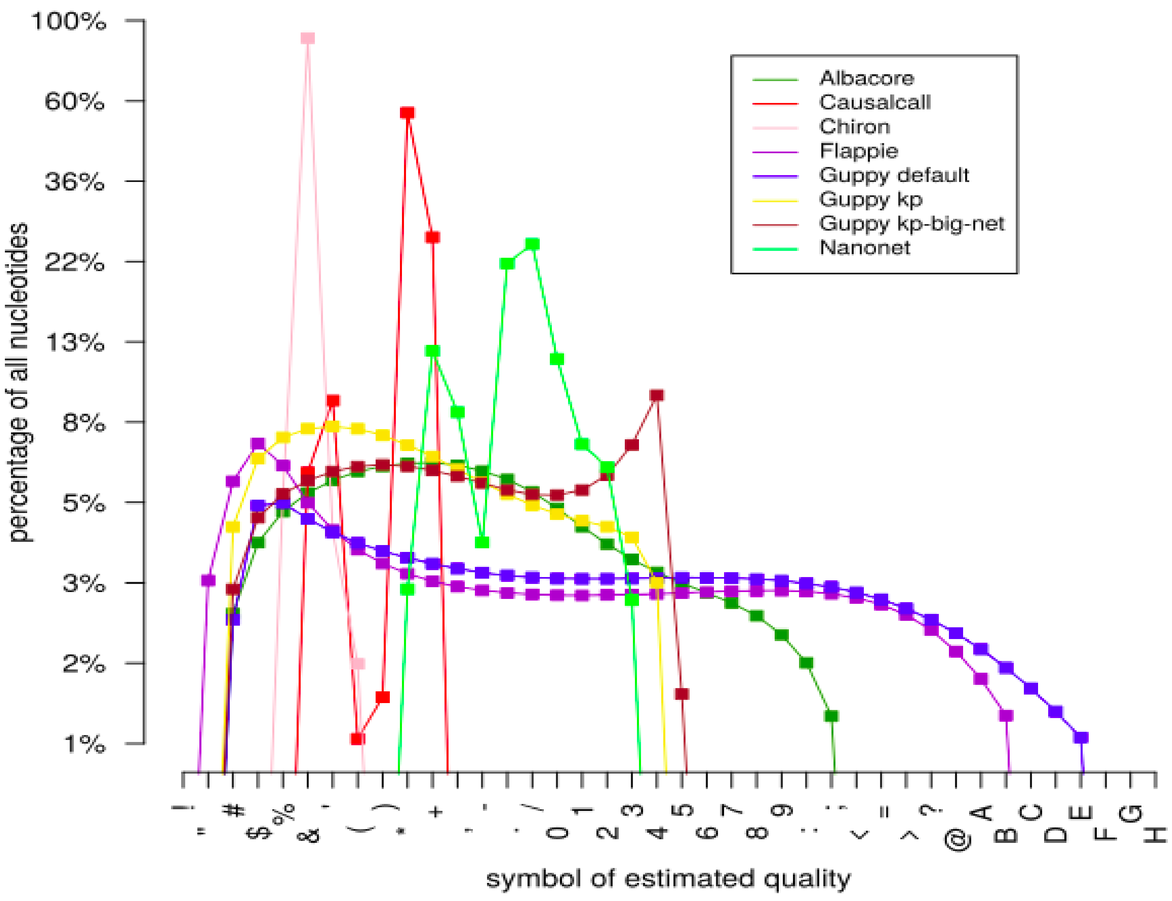
Abstract

Currently, one of the fastest growing DNA sequencing technologies is nanopore sequencing. One of the key stages of processing sequencer data is the basecalling process, which from the input sequence of currents measured on the pores of the sequencer reproduces the DNA sequences called DNA reads. Many of the applications dedicated to basecelling together with the DNA sequence provide the estimated quality of reconstruction of a given nucleotide.

Herein, we examined the estimated quality of nucleotide reconstruction reported by another basecallers. The results showed that the estimated reconstruction quality reported by different basecallers may vary depending on the tool used. In particular, for some tools, along with successive symbols of the estimated reconstruction quality (which theoretically should mean more and more accurate reconstruction), the real quality of the nucleotide increases (the number of matched nucleotides increaces and the number of errors decreases). However, there are tools that report the estimated reconstruction quality in the basecalling results, but these values are in no way interpretable. What is more, the estimated reconstruction quality reported in basecalling process is not used in any investigated tool for processing nanopore DNA reads..

Dataset

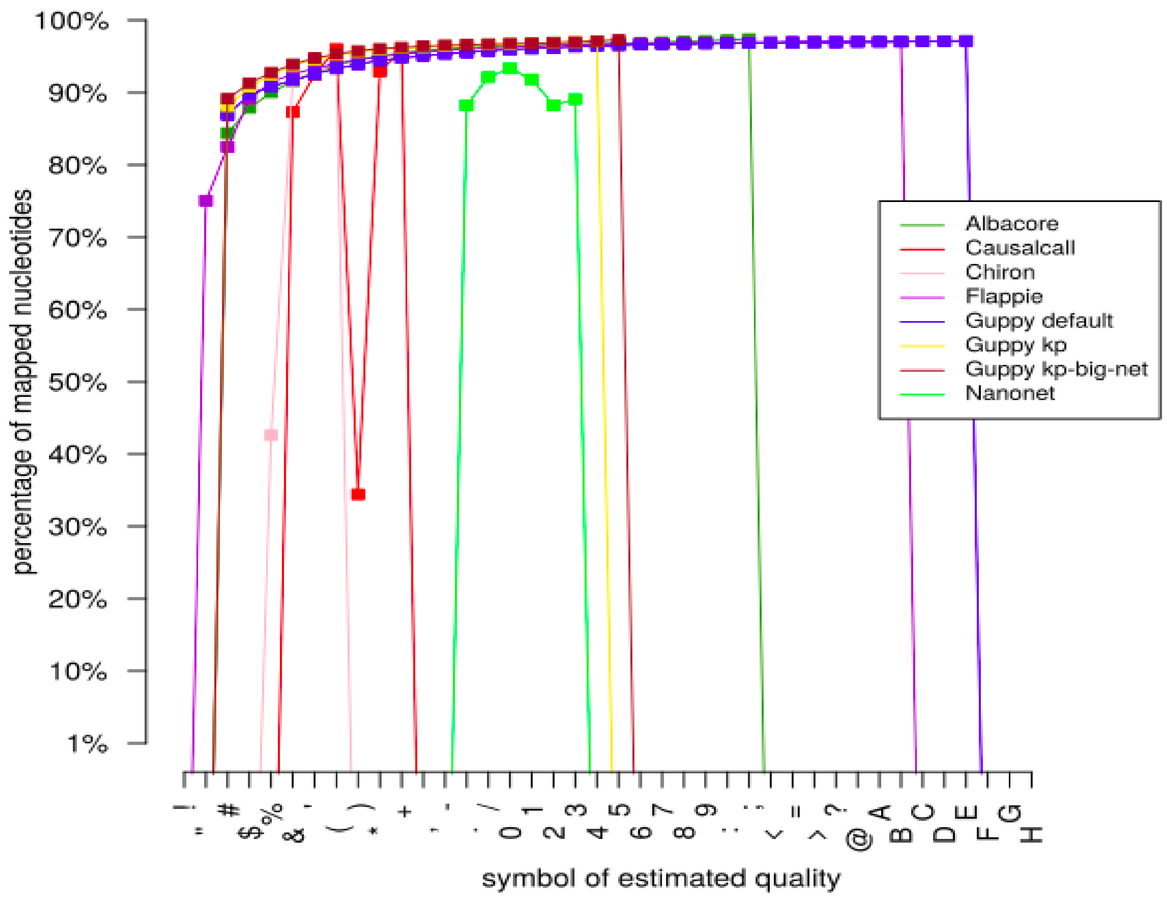
Basecaller	No. of reads	Sum [Mbp]	Mapped [%]	Match [%]
Albacore	4467	116.63	95.77	86.64
Causalcall	4467	115.12	92.21	84.36
Chiron	4467	85.44	81.88	80.43
Flappie	4467	115.04	95.44	89.66
Guppy default	4467	115.48	96.47	89.68
Guppy kp	4467	113.84	96.35	87.60
Guppy kp-big-net	4467	114.99	97.32	89.73
Nanonet	7702	118.18	67.33	84.05



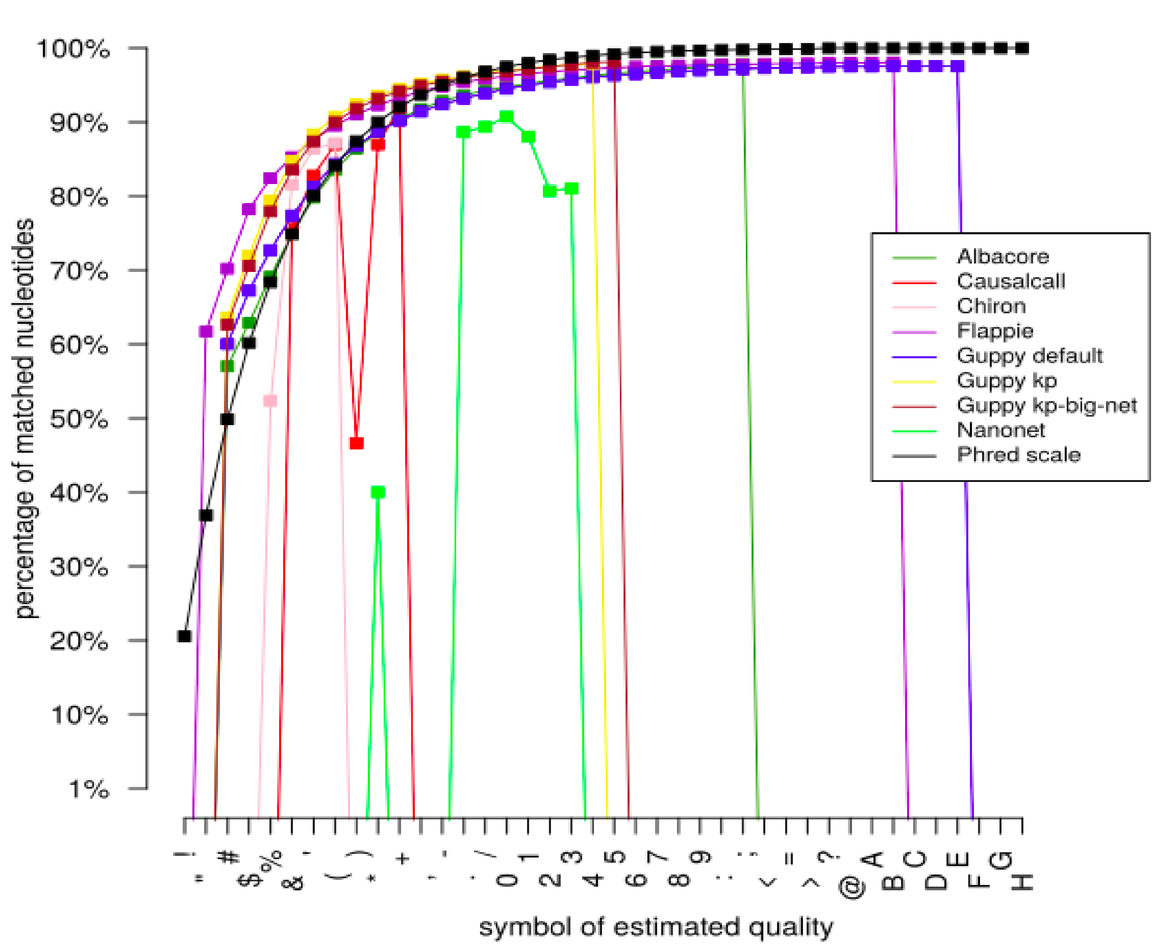
Acknowledgments

The project was funded by POB Research Centre Cybersecurity and Data Science of Warsaw University of Technology within the Excellence Initiative Program - Research University (ID-UB). This work has been also supported by the Polish National Science Center grant Preludium 2019/35/N/ST6/01983.

Results



A



References

Wick, Ryan R., Louise M. Judd, and Kathryn E. Holt. "Performance of neural network basecalling tools for Oxford Nanopore sequencing." Genome biology 20.1 (2019): 129.

David, Matei, et al. "Nanocall: an open source basecaller for Oxford Nanopore sequencing data." Bioinformatics 33.1 (2017): 49-55

Teng, Haotian, et al. "Chiron: translating nanopore raw signal directly into nucleotide sequence using deep learning." GigaScience 7.5 (2018): giy037.

Ḱoren, Sergey, et al. "Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation." Genome research 27.5 (2017): 722-736.

Boža, Vladimír, Broňa Brejová, and Tomáš Vinař. "DeepNano: deep recurrent neural networks for base calling in MinION nanopore reads." PloS one 12.6 (2017): e0178751.

Stoiber, Marcus, and James Brown. "BasecRAWller: streaming nanopore basecalling directly from raw signal." BioRxiv (2017): 133058.

Wrocław University of Environmental and Life Sciences

K. Kotlarz, M. Mielczarek, T. Suchocki, B. Czech, B. Guldbrandtsen, J. Szyda

Wrocław University of Environmental and Life Science, University of Bonn

THETA
Statistical Genetic Group
Institute of Animal Genetics

DIG (DEEP)ER

Deep learning algorithms for the imbalanced classification of correct and incorrect SNP genotypes from WGS pipelines

#1 MATERIALS

Whole-genome DNA sequence of four traditional Danish Red Dairy Cattle bulls:

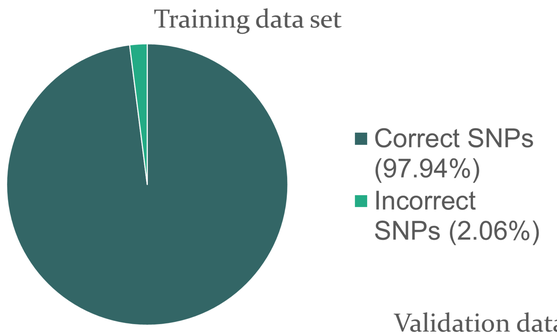
1) The training data set—**three animals**,
2) The validation data set—**the fourth animal**.

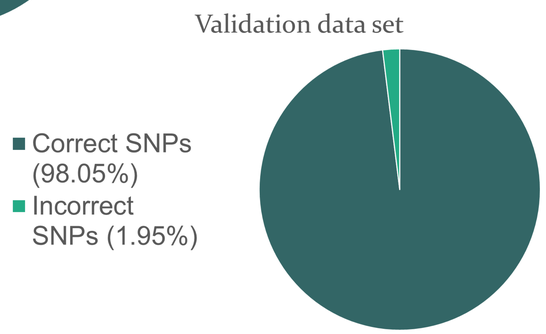
Correct SNPs (concordant WGS—Chip):

1) Training data set: 2 227 995 SNPs,
2) Validation data set: 749 506 SNPs.

Incorrect SNPs (discordant WGS—Chip):

1) Training data set: 46 920 SNPs,
2) Validation data set: 14 940 SNPs.



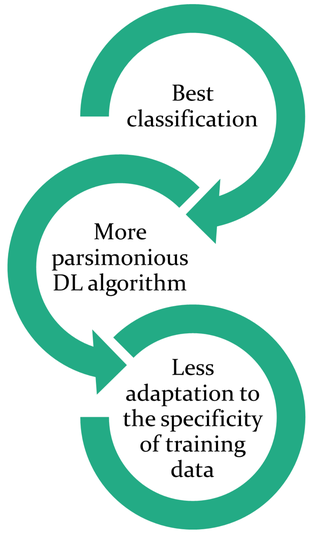


Contact:

Department of Genetics
Wrocław University
of Environmental
and Life Sciences

7 Kożuchowska Street
51-631 Wrocław
Poland
krzysztof.kotlarz@upwr.edu.pl

#4 CONCLUSIONS



#2 METHODS

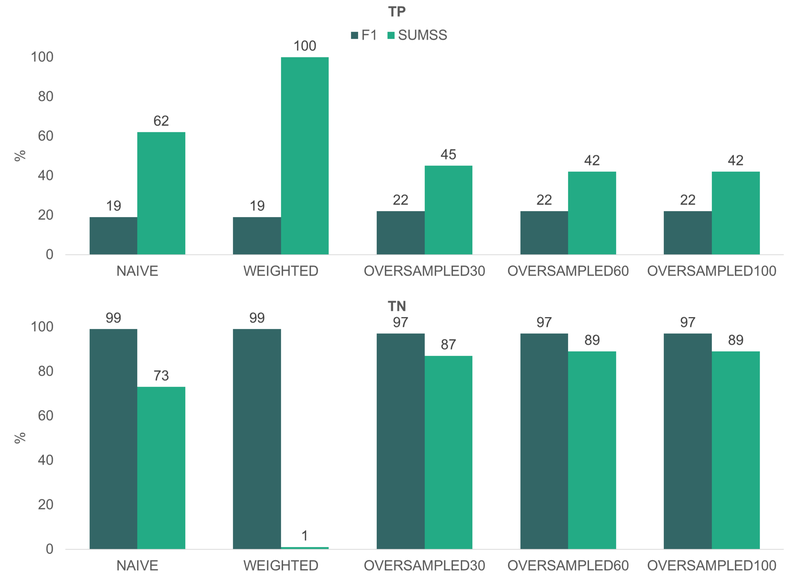
Deep Learning algorithms

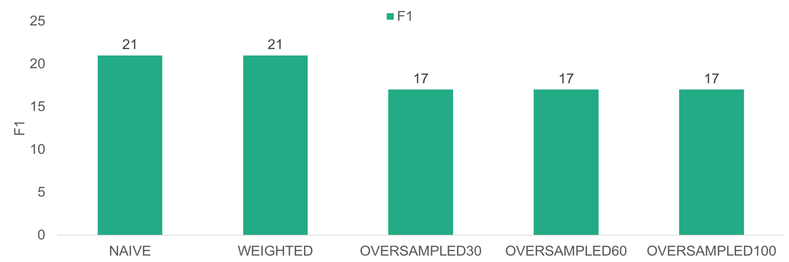
1) Naïve algorithm
2) Weighted algorithm
3) Oversampled algorithm; oversampled of the incorrect SNP:
   - 30%
   - 60%
   - 90%

Cutoff points

1) The estimated cutoff points for each model by:
   - $F1 = \frac{2TP}{2TP+FN+FP}$
   - $SUMSS = \frac{TN}{TN+FP} + \frac{TP}{TP+FN}$

#3 RESULTS





Classification of **validation data** by the algorithms, based on the cutoff thresholds for the **F1** or **SUMSS** metrics.

1) **True positive (TP)**—an incorrect SNP classified as incorrect,
2) **False negative (FN)**—an incorrect SNP classified as correct,
3) **True negative (TN)**—a correct SNP classified as correct,
4) **False positive (FP)**—a correct SNP classified as incorrect,
5) **F1**—values of the F1 metric.

DNA sequence features underlying large-scale duplications and deletions in humans

Mateusz Kołomański[1], Joanna Szyda[1,2], Magdalena Frąszczak[1], Magda Mielczarek[1,2]

[1] Biostatistics group, Department of Genetics, Wroclaw University of Environmental and Life Sciences
[2] National Research Institute of Animal Production

Objective

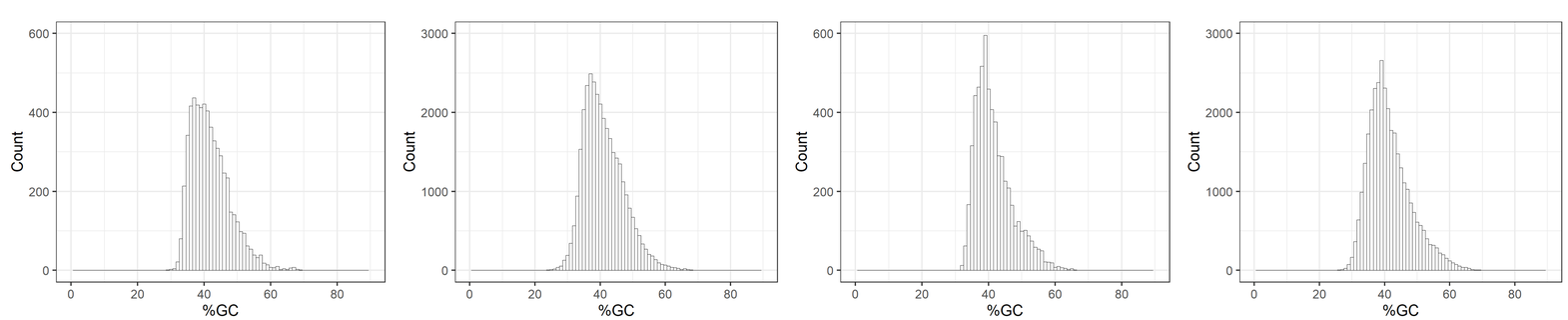
Characterizing regions of human genome that are susceptible to formation of Copy Number Variants.

Conclusions

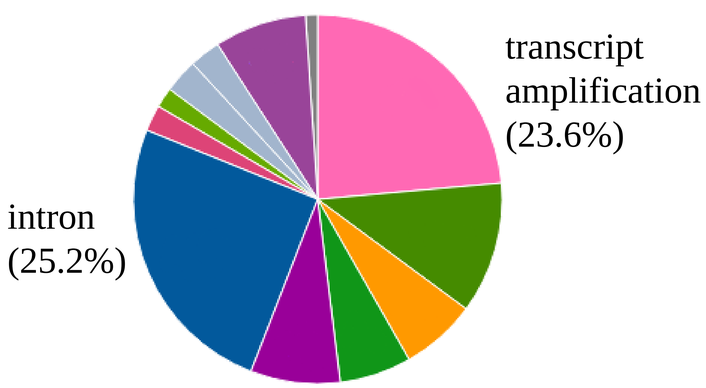
- **Deletions and sequences upstream of Copy Number Variants have low sequence complexity.**
- **Large proportion of CNVs overlap with introns.**

Results

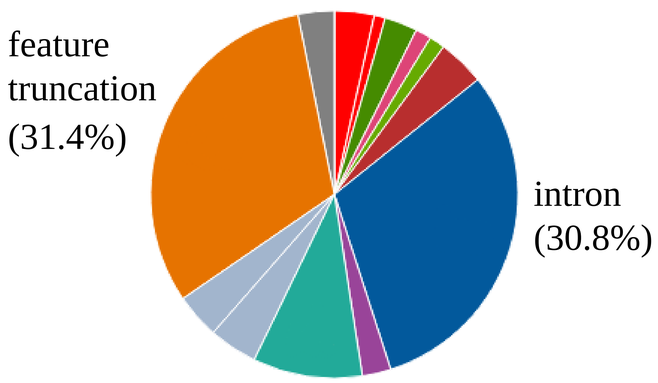
GC-pairs content



Functional annotation



duplications



deletions

Material & Methods

- Database of *1000 Genomes Project*
- 5 867 duplicated and 33 181 deleted regions
- 100 bp-long sequences flanking CNVs
- Random regions
- Seqences extracted from reference genome (GRCh38)

- Analysis regarded:
- Unknown nucleotide contents (14 CNVs)
- Guanine-Cytosine pairs content
- Sequence complexity → sDust software
- CNV-related and randomised regions comparison → Wilcoxon test
- Functional annotation → VEP