Comparison of Nanopore DNA Sequencing Basecallers on Whole Human Data

Erik Jagyugya, Peter Sarkozy
Department of Measurement and Information Systems
Budapest University of Technology and Economics
Budapest, Hungary
psarkozy@mit.bme.hu

***Abstract*—Since the release of Oxford Nanopore Technologies' MinION single molecule real-time (SMRT) DNA sequencing platform, a multitude of approaches have been evaluated to identify the exact nucleotide sequence passing through the individual pores based on the raw, picoampere level current recorded from the device. Multiple Hidden Markov Model (HMM) and artificial neural network (ANN) based basecalling approaches have been released.**

We examined the most promising academic approaches, and compared them to the reference solution provided by the platform vendor, using the NA12878 whole genome shotgun sequencing dataset. Multiple types of systematic errors offer challenges to each individual solution, thus we propose a framework to unify the strengths of each basecaller, and to aggregate their output in order to increase their accuracy over any single solution.

***Index Terms*—DNA sequencing, Basecalling, Nanopore**

I. INTRODUCTION

The MinION released by Oxford Nanopore Technologies (ONT) is a single-molecule real-time (SMRT) DNA sequencing technology that offers unprecedented read lengths in a novel, compact form factor while greatly simplifying library preparation procedures compared to current next-generation sequencing (NGS) platforms.

The method of identifying the individual nucleotides comprising a DNA sequence utilizes a nano-scale pore embedded in an artificial membrane across a semiconductor surface. The blockage of the picoampere level current passing through the pore while a DNA molecule passes through it is measured and sampled at 4kHz. The traversal of each molecule is slowed down by a ratcheting enzyme, to ensure enough time to sample the current respective of the DNA sequence. The current level is determined by 5-6 consecutive nucleotides (k-mers), and is a characteristic of the types of bases. Despite offering extremely long reads (up to 100kb), the accuracy of the platform is only approximately 90%, compared to the 99-99.9% accuracy of NGS platforms [1].

The original method to transform the current measurement into a sequence of nucleotides (basecalling) splits the current into consecutive segments (events) where the current levels were relatively static. The series of events are then passed into a hidden markov model (HMM) for decoding into a sequence of k-mers [2]. Finally, empirically calibrated quality scores are assigned to each base, and the results stored in the original fast5 HDF [3] container file.

Recently, neural networks models have been developed to perform nanopore DNA basecalling.

In our study we reviewed some of the available basecallers. We investigated characteristics such as read identity, insertion and deletion rate, the size and frequency distribution of insertions and deletions among the software tools and we summarized the overall accuracy. We unified this information and propose consensus calling by pairwise merging.

II. MATERIALS AND METHODS

A. Raw and reference data

We used the publicly available whole genome shotgun sequencing dataset of human sample NA12878 [4]. This is a well characterized genome, and is used as a gold standard in measuring the quality of human DNA sequencing.

For our study, we compared the source of the chromosome 1 data. We chose a subset of chromosome 1 which represent the whole genome well in our case, because the entire dataset was deemed too large. The subset is approximately 300 GB and contains 100000 individual fast5 files with a mean read length of 10kb. Using the rel3 version with R9.4 chemistry. We used the GRCh37 human reference genome from the Reference Genome Consortium [5] instead of NA12878 reference because the deviation between them is two orders of magnitude lower than the expected read error rate.

B. Basecalling DNA sequences

Raw reads were processed by four basecallers, Albacore [11], Chiron [10], Metrichor and Scrappie. The Metrichor data was available directly from the raw fast5 files, as it is a cloud-based basecaller that cannot be run locally.

Albacore, Chiron and Scrappie all utilize GPU support for the majority of their computations, however our version of Chiron had issues with our GPU setup so it was run in CPU only mode. The same neural network architecture is run on the CPU as the GPU, but the computations are orders of magnitude slower. We attempted to include a fifth basecaller, basecRAWler [6], but it had to be removed due errors and poor support for our hardware.

Metrichor

Metrichor is the ONT cloud-based platform basecaller [8], it works by segmenting the raw current signal into events which represent the traversion of a single nucleotide through the pore,

and utilizes a HMM on the event sequence to indentify the nucleotide order. Instead of each event representing each the base individually, the current level is influenced by adjacent bases which correlate to the length of the narrowest region of the nanopore. The model supports 5 and 6 adjacent bases in the decoding step [7]. The individual events are decoded into stay and skip probabilities by the HMM, according to their duration, mean current and noise. These state transitions are decoded into the final basecalled sequence.

The HMM model was superseded by a more accurate recurrent neural network (RNN) model in early 2016 [7]. We used the data from the cloud-based version as integrated into the EPI2ME service which relied on HMM. The raw fast5 files already contained the required Metrichor basecall results, and were used subsequently without modification.

Albacore

Albacore is one of the official command-line basecallers of ONT, and is considered the "gold-standard" in accuracy. As the software is unfortunately not open source, the neural network structure used in its model is not public. We used version 2.1.3, the most recent at the time of writing. The neural network does not require segmenting the reads into events, and instead operates on the raw current levels, allowing for great improvements in basecalling accuracy.

Albacore enforces stricter limits on minimum sequence quality, and will refuse to call low quality reads, contributing to it's higher overall performance. In our study, this has no effect as we only used reads called by all 4 basecallers.

Chiron

Chiron is a novel third-party neural network based basecaller [10]. It couples a recurrent neural network (RNN), a convolutional neural net (CNN) and a connectionist temporal classification (CTC) decoder. This structure also enables it to model the raw signal data directly, without use of any segmentation step. It is the first publicly available artificial neural network which can translate raw current signals directly to nucleotide base sequences. The basecaller was trained on only a small subset of data from the lambda phage genome and from the E. coli genome, and yet it has a surprising ability to generalize to larger genomes such plant and mammal genomes. Chiron is as accurate as the ONT designed and trained Albacore in some cases, and outperforms all other existing basecallers on bacterial genomes.

Chiron consists of two sets of layers: a set of convolutional and a set of recurrent ANN see Fig. 1. The convolutional layer defines local patterns from raw signal, whereas the recurrent layer combines these pattern into base probabilities. The CTC decoder makes data segmentation unnecessary through the use of blank labels inserted into the sequences. This allows it to model input and output sequences of varying lengths. It models a two dimensional graph similar to a pairwise alignment matrix, and the neural network predicts the probabilities of the transitions along the alignment matrix.

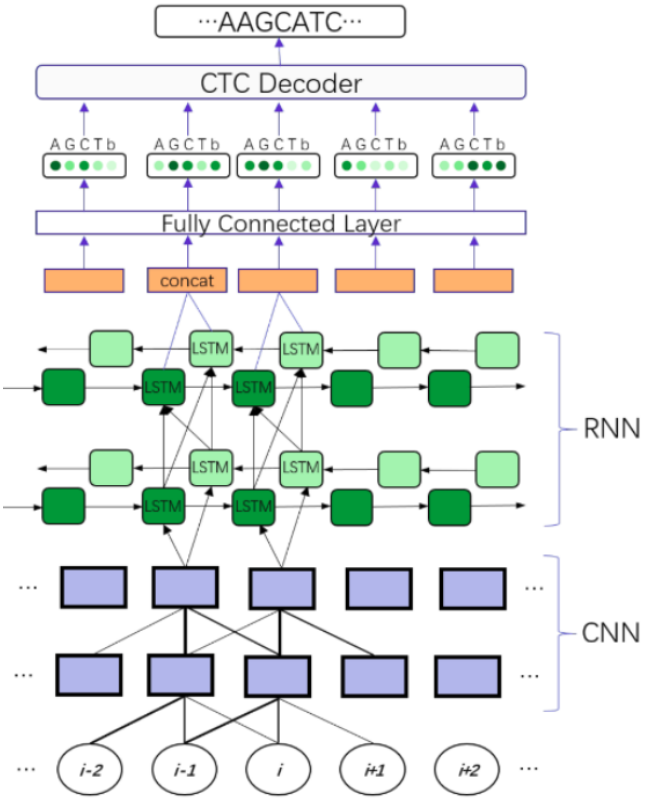

Fig. 1. The neural network architecture used by Chiron [10]

Scrappie

Scrappie is another basecaller developed by ONT, this is the first basecaller that was developed specifically to address homopolymer deletions. One of the major hurdles in SMRT sequencing is accurately determining length of homopolymers [9]. Scrappie was run parallel with Metrichor and NanoNet (a first generation experimental NN basecaller) on human chromosome 20. It was found that Scrappie indeed called homopolymer areas more accurate than the other two basecaller. Homopolymeric stretches of up to 16 bases were called accurately which referred to the transducer-based homopolymer calling[8][12]. We utilized version 1.3.0 with the *raw model* to perform basecalling.

C. Processing DNA sequences

We prepared our dataset by first running all of the basecallers and extracting the relevant data for the entire set. Not all basecallers produce output sequences for every input, so we selected a subset of reads that had calls from each basecaller. All of the software tools use different naming schemes when outputting results, so these had to be unified to contain only the read ID string. The resulting sequences were aligned to the reference genome using the Burrows-Wheeler Aligner which is tolerant with errors given longer query sequences [13][14], and all alignments with mapping qualities below 30 were

discarded, because low alignment quality is indicative of the read originating from a different chromosome, especially at the read lengths produced by the ONT platform. In the alignment processing BWA uses soft clipping. Some reads had secondary alignments, but only the highest-scoring alignment was used. We investigated the overall match, mismatch and insertion rates of each dataset.

III. Results

A. Systematic errors

While all tested basecallers perform reasonably well on our dataset, homopolymer stretches (repeating identical nucleotides) still present a challenge. HMM based basecallers struggle with calling homopolymers longer than 5 bases, as the event segmentation process hides the very long times that the current stays static because there are identical nucleotides in the pore, as shown on Fig. 4. Albacore called the overall lowest number of deletions and the highest identity rate too, as NN based tools tend to cope better with homopolymer stretches. Insertions were overall less common, and mostly randomly distributed along the sequence without any sequence-specific bias Fig. 3. A insertions and deletions were twice as common as T indels, with G and C indels being equally common. A/G and C/T nucleotide substitutions are approximately 3x more frequent than other substitutions, as they share similar current signatures due to their similar purine/pyrimidine molecular structures, respectively as shown in Tab. III-A and Tab. III-C.

B. Unified Basecaller

The aggregation of the output of multiple basecallers to obtain a consensus read is nontrivial in the case of ONT reads. While multiple sequence alignment approaches are feasible for individual genes and transcripts, even pairwise alignment quickly becomes computationally expensive with ONT read lengths. We implemented a method to perform the pairwise merging of all basecaller output sequences. We find the optimal pairwise alignments with the Smith-Waterman algorithm, using a match score of 1, a mismatch, gap open and gap extension penalty of -1. Each read is successively merged, taking to account the PHRED-scaled quality scores. On matches, the maximum of the quality is passed on. On mismatches, the higher quality base overrides the lower quality one, and on gaps, a configurable threshold is specified to select the sensitivity in which insertions and deletions are selected. Unfortunately, Scrappie did not provide base quality scores, so we used a flat PHRED score of 10 to approximate it.

C. Read identity

Read identity was derived from total base matches. Albacore and Scrappie have the best performance as shown on Fig. 2. These basecallers developed by ONT. However Metrichor performed the worst which is also an ONT product. As Metrichor relies on a HMM its accuracy is far away from basecallers based on NN. Chiron and Consensus medians are near each other but their distribution is markedly different. Chiron produces more accurate read against Consensus. The overall identity rates are shown on Tab. III-C.

Albacore				
Matches	**A**	**C**	**G**	**T**
A	22.98%	0.35%	1.11%	0.34%
C	0.30%	21.16%	2.35%	0.40%
G	0.26%	0.26%	20.91%	0.37%
T	0.34%	0.47%	0.31%	23.95%
Chiron				
Matches	**A**	**C**	**G**	**T**
A	20.87%	0.41%	1.34%	0.36%
C	0.33%	20.06%	0.23%	0.59%
G	1.71%	0.35%	19.86%	0.39%
T	0.27%	0.39%	0.25%	22.40%
Metrichor				
Matches	**A**	**C**	**G**	**T**
A	20.92%	0.45%	1.44%	0.37%
C	0.27%	19.48%	0.22%	0.46%
G	1.34%	0.34%	19.30%	0.35%
T	0.37%	0.55%	0.32%	22.54%
Scrappie				
Matches	**A**	**C**	**G**	**T**
A	22.32%	3.73%	1.32%	0.32%
C	0.31%	20.62%	0.23%	0.37%
G	0.11%	0.27%	20.34%	0.27%
T	0.37%	0.52%	0.33%	23.58%

TABLE I
The match and mismatch rate of all 4 basecallers. Columns specify reference bases, while rows show the read bases.

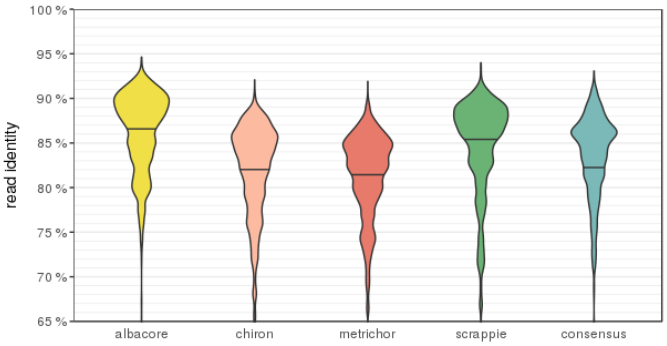

Fig. 2. Read identity distribution and medians which was weighted by read lengths it is marked by black horizontal line.

IV. Conclusion

The characteristics of each basecaller show that currently Albacore performs the best on our dataset. We have confirmed that most of the errors in the sequences crop up at identical positions in the sequences, meaning that they are artifacts of the underlying measurement process, and unifying the output of multiple basecallers can only offer limited improvements in accuracy. The pairwise merging of the output of each basecaller into a consensus sequence benefits from the specification of each basecaller's error characteristics, as they provide

Consensus				
Matches	**A**	**C**	**G**	**T**
A	22.18%	4.42%	1.02%	0.40%
C	0.46%	20.88%	0.37%	0.46%
G	0.97%	0.36%	20.53%	0.35%
T	0.50%	0.54%	0.47%	23.51%

TABLE II
The match and mismatch rate of consensus sequences. Columns specify reference bases, while rows show the read bases.

Basecaller	**Insertion rate**	**Deletion rate**	**Identity rate**
Albacore	3.57%	5.54%	89.01%
Chiron	1.92%	10.16%	83.2%
Metrichor	1.52%	11.20%	82.25%
Scrappie	2.63%	7.31%	86.86%
Consensus	5.21%	5.33%	88.06%

TABLE III
The overall accuracy of the 4 tested basecallers, and the results of the consensus sequences

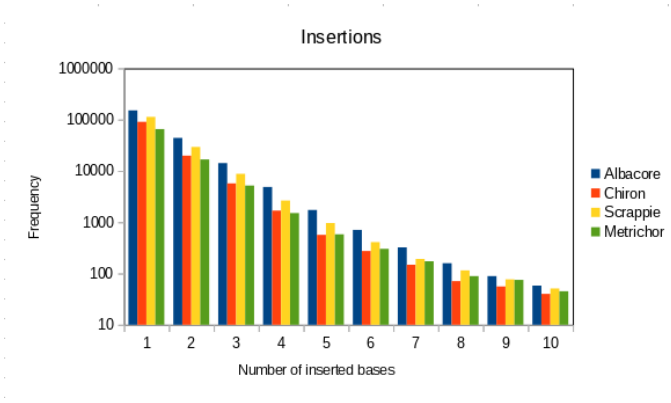

Fig. 3. The size and frequency distribution of insertions among the 4 basecallers

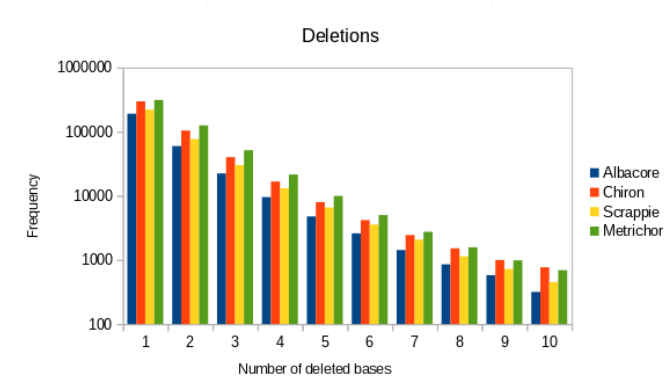

Fig. 4. The size and frequency distribution of deletions among the 4 basecallers

additional information when considering the individual base quality scores. However, the quality scores provided by each basecaller show marked differences, and it could prove useful to perform base quality score recalibration based on empirical results. While our unified basecalls did not show an overall lead above all individual basecallers, it did provide more robust results overall, with a more reliable per-base quality score.

ACKNOWLEDGMENT

The authors acknowledge that they are participants in the Oxford Nanopore Technologies's MinION Access Program. The Titan Xp used for this research was donated by the NVIDIA Corporation. This research was supported by the OTKA-K-112915 Grant, and the Multipurpose Health Monitoring Platform bilateral Croatian-Hungarian grant. The authors state that they have no other conflicts of interest.

REFERENCES

[1] Quail MA, et al, A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 2012; 13:341; DOI:10.1186/1471-2164-13-341
[2] Matei D., et al, Nanocall: an open source basecaller for Oxford Nanopore sequencing data. Bioinformatics 2017 DOIL10.1093/bioinformatics/btw569
[3] The HDF Group. Hierarchical Data Format, version 5, 1997 http://www.hdfgroup.org/HDF5/.
[4] Oxford Nanopore Human Reference Datasets ,https://github.com/nanopore-wgs-consortium/NA12878, 11-15-2017
[5] Human Genome Overview - Genome Reference Consortium, https://www.ncbi.nlm.nih.gov/grc/human, 03-12-2017
[6] Stoiber M., Brown J.: BasecRAWller: Streaming Nanopore Basecalling Directly from Raw Signal BioarXiv 2017 DOI:10.1101/133058
[7] C. V. de Lannoy, D. de Ridder, J. Risse, A Sequencer Coming Of Age: De Novo Genome Assembly Using MinION Reads, https://www.biorxiv.org/content/early/2017/05/26/142711, 2017-12-07
[8] Oxford Nanopore Technologies, https://nanoporetech.com/, 2018
[9] P. Antal P. Sarkozy, Á. Jobbágy. Calling Homopolymer Stretches from Raw Nanopore Reads by Analyzing k-mer Dwell Times. 2016. DOI: 10.1007/978-981-10-5122-7-61.
[10] T. Haotian, C. M. Duc, H. M. B., D. Tania, W. Sheng and C. Lachlan, Chiron: Translating nanopore raw signal directly into nucleotide sequence using deep learning, bioRxiv
[11] Oxford Nanopore Technologies, https://nanoporetech.com/about-us/news/new-basecaller-now-performs-raw-basecalling-improved-sequencing-accuracy, 11-15-2017
[12] S. Wermter, Knowledge Extraction from Transducer Neural Networks, Journal of Applied Intelligence, 12, 27, 44 (2000)
[13] Li H. and Durbin R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics, 26, 589-595. [PMID: 20080505]
[14] Li H. Burrows-Wheeler transform repository, https://github.com/lh3/bwa, 11-15-2017