Ococo: an online variant and consensus caller

Karel Břinda\textsuperscript{1,2}, Valentina Boeava\textsuperscript{3,4}, and Gregory Kucherov\textsuperscript{5,6}

\textsuperscript{1}Center for Communicable Disease Dynamics, Department of Epidemiology, Harvard TH Chan School of Public Health, Boston MA; \textsuperscript{2}Department of Biomedical Informatics, Harvard Medical School, Boston MA; \textsuperscript{3}Institut Curie – Centre de Recherche, PSL Research University, Mines Paris Tech, INSERM U900, Paris, France; \textsuperscript{4}Institut Cochin, INSERM U1016, CNRS UMR 8104, Université Paris Descartes UMR-S1016, Paris, France; \textsuperscript{5}LIGM/CNRS, Université Paris-Est, Marne-la-Vallée, France; \textsuperscript{6}Skolkovo Institute of Science and Technology (SkolTech), Moscow Region, Russia.

\section*{ABSTRACT}

\textbf{Motivation:} Identifying genomic variants is an essential step for connecting genotype and phenotype. The usual approach consists of statistical inference of variants from alignments of sequencing reads. State-of-the-art variant callers can resolve a wide range of different variant types with high accuracy. However, they require that all reads be available from the beginning of variant calling and be sorted by coordinates. Sorting is computationally expensive, both memory- and speed-wise, and the resulting pipelines suffer from storing and retrieving large alignment files from external memory. Therefore, there is interest in developing methods for resource-efficient variant calling.

\textbf{Results:} We present Ococo, the first program capable of inferring variants in a real-time, as read alignments are fed in. Ococo inputs unsorted alignments from a stream and infers single-nucleotide variants, together with a genomic consensus, using statistics stored in compact several-bit counters. Ococo provides a fast and memory-efficient alternative to the usual variant calling. It is particularly advantageous when reads are sequenced or mapped progressively, or when available computational resources are at a premium.

\textbf{Availability:} \url{http://github.com/karel-brinda/ococo}

\textbf{Contact:} kbrinda@hsph.harvard.edu

\section{1 INTRODUCTION}

Identifying genomic variants is an essential step for connecting genotype and phenotype. The goal of \emph{variant calling} is to identify genomic variants present in the sequenced individual or a population. Most commonly, variant calling proceeds by read mapping and then sliding a small window throughout the genome, collecting statistics for all reads aligned within the window and calculating the likelihood of variants observed in these alignments. We term this approach \emph{offline variant calling} as it requires that all read alignments are available from the beginning. Offline calling is implemented in all major variant callers (see, e.g., \cite{bao2014}).

However, offline variant calling is highly time- and space-demanding. First, all alignments must be available and get sorted by coordinates prior to variant calling; this involves storing and retrieving large alignment files from external memory. Second, variant callers usually apply computationally expensive steps, such as realignments, even for regions where this is not necessary. The resulting performance can be particularly limiting on portable devices, personal computers, or in a cloud environment with restricted resources.

Here, we introduce the concept of \emph{online variant calling}, where variants are inferred in real time, as read alignments are fed in. We implement this approach in a program called Ococo, the first online variant caller. Ococo inputs unsorted alignments from an unsorted SAM/BAM stream \cite{li2009} and infers single-nucleotide variants, together with a genomic consensus, using statistics stored in compact several-bit counters. Ococo provides a fast and memory-efficient alternative to the usual variant calling, which is particularly advantageous when reads are sequenced or mapped progressively, or when available computational resources are at a premium.

\section{2 METHODS}

\textbf{Overview.} Ococo calls variants and consensus directly from an unsorted SAM/BAM file, possibly provided in a stream. To do that, Ococo stores and maintains variant statistics for all genomic positions about previous alignments as well as a consensus sequence, which represents the current internal reference. The consensus can be initialized from a user-provided sequence, typically the same as used for read mapping. Whenever a new alignment is loaded, Ococo updates the statistics and assesses whether they are still concordant with the consensus. If not, the consensus is corrected and the corresponding substitution reported as a novel variant.

\textbf{Compact representation of variant statistics.} In the online approach reads can potentially map to any location. This is a fundamental difference from the offline calling, where reads are sorted and statistical inference uses a small sliding window, collecting information about locally overlapping alignments. Therefore, the main challenge of online calling is to design variant statistics for the whole genome that fit into main memory and at the same time be sufficiently informative for inferring variants. We propose using small, several-bit nucleotide counters and complementing them with fast bit operations.

The Ococo statistics consist of four integer counters per position, one per each nucleotide (Figure 1). Every counter represents the number of nucleotides aligned to that position; however only the most significant bits are stored. Whenever a new alignment is loaded, the corresponding counters are incremented (Figure 1a). If a counter is already saturated and yet is to be incremented, then all counters at the position are first bit-shifted, losing their rightmost (least significant) bit (Figure 1e). This mechanism makes it possible to compute nucleotide frequencies in a limited space and filter out randomly distributed sequencing errors. Ococo supports three counter...
Figure 1. Internal statistics of an online variant caller. Example of update of OCOCO counters for a single position of the genome. The first 4 bits carrying the nucleotide consensus are followed by 4 nucleotide counters, each of them 3 bits long in this case. Vertical axis corresponds to time. The figure shows how the counters and consensus are updated based on the received alignments. a) At the beginning, the consensus base is initialized to the reference base (A). All counters are set to 0. b) The A counter is incremented; the statistics stay concordant with the consensus. c) The G counter is incremented, which triggers a consensus update and reporting a new variant (A→G). d) The G counter is to be incremented, but it is already saturated. Therefore, all counters of the position must be bit-shifted first.

configurations: 16, 32, and 64 bits per position corresponding to 4 bits for the bit-shifted first.

Working modes. OCOCO supports two modes of online calling: the real-time and batch modes. Whereas the real-time mode updates consensus and reports variants immediately after processing each read, the batch mode postpones reporting updates until all reads from the current batch have been processed.

Implementation. OCOCO is implemented in C++ and released under the MIT license. The software package is available from https://github.com/karel-brinda/ococo, BioConda, Trimming et al. [2018], and Zenodo [Brinda et al. 2017].
5 ACKNOWLEDGEMENTS

The work has been supported by ABS4NGS [ANR-11-BINF-0001], Labex Bézout, and The Bill and Melinda Gates Foundation [GCGH OCE, OPP1151010]. VB was supported by the ATIP-Avenir Program, the ARC Foundation [RAC16002KSA-R15093KS], and the “Who Am I?” laboratory of excellence [ANR-11-LABX-0071] funded by the French Government [ANR-11-IDEX-0005-02]. The authors would like to thank Victoria E Jones, Cristina M Herren, and Siân V Owen for critical reading and helpful suggestions.

Conflict of Interest: None declared.

REFERENCES

Bao, R., Huang, L., Andrade, J., Tan, W., Kibbe, W. A., Jiang, H., and Feng, G. (2014). Review of Current Methods, Applications, and Data Management for the Bioinformatics Analysis of Whole Exome Sequencing. Cancer Informatics, 13(s2):CIN.S13779.

Brinda, K., Boeva, V., and Kucherov, G. (2016). Dynamic read mapping and online consensus calling for better variant detection. arXiv:1605.09070.

Brinda, K., Boeva, V., and Kucherov, G. (2016). RNF: a general framework to evaluate NGS read mappers. Bioinformatics (Oxford, England), 32(1):136–9.

Brinda, K., Boeva, V., and Kucherov, G. (2017). karel-brinda/ococo: Ococo 0.1.2.6. Zenodo, 1095696. Available from https://doi.org/10.5281/zenodo.1095696.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., and Durbin, R. (2011). The variant call format and VCFtools. Bioinformatics, 27(15):2156–2158.

Grüning, B., Dale, R., Sjödin, A., Chapman, B. A., Rowe, J., Tomkins-Tinch, C. H., Valieris, R., and Köster, J. (2018). Bioconda: sustainable and comprehensive software distribution for the life sciences. Nature Methods, 15(7):475–476.

Koboldt, D. C., Chen, K., Wylie, B., Larson, D. E., McLellan, M. D., Mardis, E. R., Weinstock, G. M., Wilson, R. K., and Ding, L. (2009). VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics, 25(17):2283–5.

Köster, J. and Rahmann, S. (2012). Snakemake-a scalable bioinformatics workflow engine. Bioinformatics, 28(19):2520–2522.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997, page 3.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25(16):2078–9.

Morris, R. (1978). Counting large numbers of events in small registers. Communications of the ACM, 21(10):840–842.
**Supplementary Figure 1.** Evaluation of OCOCO with Chlamydia Trachomatis (1.046Mbp). a) Online variant calling as a function of time. The blue curve shows the cumulative number of updates of the consensus as a function of the number of processed alignments (or the actual coverage). The red curve shows the edit distance from the simulated sequenced genome. b) Speed comparison. Comparison of time to completion of variant calling using OCOCO and a pipeline based on VARSCAN.

**Supplementary Figure 2.** Evaluation of OCOCO with human chromosome 17 (78,775Mbp). The figure is of the same format as Supplementary Figure 1.