**First Name:** Peter Kimani



**Surname:** Muchina

**Email address:** kimanimuchina@gmail.com

**Institution:** Pwani University/ ICIPE

Please copy and paste a recent picture of yourself in the box:

**Biography**

Peter Muchina has a BSc. In Biochemistry and Molecular Biology. He has worked for the International Center of Insect Physiology and Ecology(ICIPE). Currently, Peter is a MSc scholar under the Eastern African Network for Bioinformatics Training fellowship. His Masters research is focused on Parallelizing long-read de novo genome assemblers on mixed architectures (CPU/GPU) with OpenACC and CUDA. His research interest is in Advanced algorithms, Genomics and Machine Learning  for Bioinformatics.

**Parallelizing long-read de novo genome assemblers on mixed**

**architectures (CPU/GPU) with OpenACC and CUDA**

1,2 Peter Muchina , 3 Jean-Baka Domelevo Entfellner

1Department of Biochemistry and Biotechnology, Pwani University, P.O Box 195-80108, Kilifi Kenya.  
2International Center of Insect Physiology and Ecology, P.O. Box 300772-00100 Nairobi, Kenya.  
3International Livestock Research Institute (BecA-ILRI Hub), PO Box 30709 Nairobi 00100 Kenya.

**Abstract**

The computational challenge of de novo genome assembly (when no reference genome is available) is a hard one to tackle, when millions to billions of short reads must be assembled into contigs and further into scaffolds. Current de novo aligners (one can cite SGA, Spades, Velvet, ABySS or SOAPdenovo) try their best at finding overlaps between reads or building De Bruijn graphs on k-mers, but doing so on relatively large genomes (say, over tens or hundreds of megabases) requires both huge amounts of available RAM and CPU cycles. The very large number (in the millions or billions) of reads now routinely generated by Next-Generation Sequencing jobs (e.g. on Illumina MiSeq, HiSeq and NovaSeq) has rendered virtually intractable the OLC (overlap layout consensus) methods, which were clearly superseded by De Bruijn graph-based methods. Since the advent of GPU-based computing (where vast arrays of individually slower graphical processors are available, but with high bandwidth to main memory and high parallelism), some avenues have been explored very recently for the acceleration of De Bruijn-based (Arioc, 2018) or OLC-based (GRASShopPER, 2018) assembly methods on mixed architectures (CPU + GPU).

The advent of long-read sequencing technologies (Oxford Nanopore and Pacific Biosciences) brought some disruption in the landscape, and it is now possible to generate reads whose length extends over the tens or hundreds of kilobases (the current record being around 2.4MB for one read). This gives a new youth to OLC methods, and there is now the hope that such OLC methods (which were in full use at the time of the construction of the first reference human genome, based on Sanger sequencing reads) would give better accuracy than graph-based methods on long reads. The problem of their computational tractability, though, still needs to be addressed properly, and could benefit greatly from the most recent mixed architectures and software stacks enabling developers to produce accelerated code.

This project, explores the development of new algorithms for accelerated long-read de novo assembly on GPU architectures. The software will be written in Python with the support of CUDA libraries. We will run test benchmarks for the software on the Jetson Nano developer kit released in March 2019 by NVIDIA, and possibly on a larger GPU-based server.