Expanding the utility of third generation sequencing with mobile k-mer counting

Kaden King*, Marco Oliva*, Christina Boucher, Franco Miiccho, and Mattia Prosperi

*: indicates co-first authors

Mobile (third-generation) sequencing technologies, including Oxford Nanopore's MinION and SmidgION, have the benefit of outputting long sequence reads--up to hundred thousands of bases--in a portable manner. These sequencing devices fit in the palm of a hand and only require a USB outlet. Unfortunately, the development of data analysis tools for these technologies is in a nascent stage, impeding on the portability of these devices. Many bioinformatics applications---such as de Bruijn graph based genome assembly---require k-mer counting as a first step. In addition, k-mer counting can inform what species or subspecies are likely contained within a sample. In this work, we present and compare two possible approaches to bring k-mer counting to mobile platforms. One approach explored is porting the existing k-mer counting application DSK directly to mobile devices using the Android NDK. The other is using Nanopore Portable Analytics Library (NanoPAL), implemented in ISO C++ v.14 and compiled for Android devices. In order to accomplish the above mentioned portability, NanoPAL uses cache-oblivious data structures and out-of-core processing methods.

We tested the performance of both DSK and NanoPAL on a range of real DNA reads ranging in size from 500 MB to 5 GB on a Samsung Galaxy S9+ with 6 GB of memory with 1.5 MB of L2 cache and 64 GB of storage, We show that although NanoPAL and DSK both have the ability to the count the k-mers of reasonably sized input on mobile devices and present an output suitable for de Brujin graph assembly, NanoPAL requires more efficient. For example, we evaluated both applications on a 4GB sized metagenomic sample (ERR2529201) and showed DSK required 945 seconds and NanoPAL only required 339 seconds to complete the counting.