Improved Error Correction of NGS Data

May 16, 2016

The work done for this doctorate thesis focuses on error correction of Next Generation Sequencing (NGS) data in the context of High Performance Computing (HPC).

Due to the reduction in sequencing cost, the increasing output of the sequencers and the advancements in the biological and medical sciences, the amount of NGS data has increased tremendously. Humans alone are not able to keep pace with this explosion of information, therefore computers must assist them to ease the handle of the deluge of information generated by the sequencing machines. Since NGS is no longer just a research topic (used in clinical routine to detect cancer mutations, for instance), requirements in performance and accuracy are more stringent. For sequencing to be useful outside research, the analysis software must work accurately and fast. This is where HPC comes into play. NGS processing tools should leverage the full potential of multi-core and even distributed computing, as those platforms are extensively available. Moreover, as the performance of the individual core has hit a barrier, current computing tendencies focus on adding more cores and explicitly split the computation to take advantage of them.

This thesis starts with a deep analysis of all these problems in a general and comprehensive way (to reach out to a very wide audience), in the form of an exhaustive and objective review of the NGS error correction field. We dedicate a chapter to this topic to introduce the reader gradually and gently into the world of sequencing. It presents real problems and applications of NGS that demonstrate the impact this technology has on science. The review results in the following conclusions: the need of understanding of the specificities of NGS data samples (given the high variety of technologies and features) and the need of flexible, efficient and accurate tools for error correction as a preliminary step of any NGS postprocessing.

As a result of the explosion of NGS data, we introduce MuffinInfo. It is a piece of software capable of extracting information from the raw data produced by the sequencer to help the user understand the data. MuffinInfo uses HTML5, therefore it runs in almost any software and hardware environment. It supports custom statistics to mould itself to specific requirements. MuffinInfo can reload the results of a run which are stored in JSON format for easier integration with third party applications. Finally, our application uses threads to perform the calculations, to load the data from the disk and to handle the UI.

In continuation to our research and as a result of the single core performance limitation, we leverage the power of multi-core computers to develop a new error correction tool. The error correction of the NGS data is normally the first step of any analysis targeting NGS. As we conclude from the review performed within the frame of this thesis, many projects in different real-life applications have opted for this step before further analysis. In this sense, we propose MuffinEC, a multi-technology (Illumina, Roche 454, Ion Torrent and PacBio -experimental), any-type-of-error handling (mismatches, deletions insertions and unknown values) corrector. It surpasses other similar software by providing higher accuracy (demonstrated by three type of tests) and using less computational resources. It follows a multi-steps approach that starts by grouping all the reads using a k-mers based metric. Next, it employs the powerful Smith-Waterman algorithm to refine the groups and generate Multiple Sequence Alignments (MSAs). These MSAs are corrected by taking each column and looking for the correct base, determined by a user-adjustable percentage.

This manuscript is structured in chapters based on material that has been previously published in prestigious journals indexed by the Journal of Citation Reports (on outstanding positions) and relevant congresses.