Improved Assembly Accuracy
by Integrating Base-calling, Error Correction
and Assembly

NARZISI G, MENGES F, MISHRA B
NYU Bioinformatics Lab, New York University, New York

Motivation
With the recent advent of a multitude of next-generation sequencing technologies (typically characterized by massively high throughput but relatively shorter read length), de novo DNA sequence assembly has become again one of the most prominent problems in Genomics and Computational Biology. Although algorithmic improvements play an important role in sequence assembly, the complexity of the problem is strongly reduced if higher quality (low error rate) sequences can be generated. For this reason, base-calling and error correction tools, together with novel assembly strategies, have begun to play a significant role in generating more accurate sequence assemblies.

Methods
We present a new de novo assembly pipeline that integrates, in a Bayesian manner, two of our recent tools: TotalReCaller (for base-calling and error correction) and SUTTA (for sequence assembly). TotalReCaller was designed to improve base-calling quality by interpreting the analog signals from sequencing machines while simultaneously aligning the sequence reads to a source reference (draft or finished) genome, whenever available, to reduce the error rate. SUTTA is an accurate sequence assembler, based on a flexible branch-and-bound framework that forcefully and quickly eliminates incorrect solutions (i.e., implausible layouts). To achieve this goal SUTTA relies on technology agnostic score functions that enable combining data from multiple sources and distinct technologies.

Results
This novel pipeline is demonstrated to improve the assembly quality significantly, when compared to the standard SUTTA pipeline (without error correction). Extensive comparison results with respect to some of the best state-of-the-art assembly algorithms (e.g., SOAPDenovo and ABySS) for short read next-generation technologies are presented. The results point to a competitive performance improvement, achievable by combining an insightful base-caller with a foresighted assembler.

Contact email
narzisi@nyu.edu