A high-throughput analysis pipeline for large next generation DNA sequencing studies

Zayed Albertyn [1], Jörg Hakenberg [1], Hongjin Bian, Huifeng Niu, James Cai [1]

[1] Disease and Translational Informatics, pREDi-DTI, Hoffmann-La Roche, Inc., Nutley, New Jersey 07110, USA
2 Oncology Discovery, pRED, Hoffmann-La Roche, Inc., Nutley, New Jersey 07110, USA

Next-generation sequencing technologies offer extraordinary high-throughput capacity and broad applications in biological research, drug development and molecular diagnostics. The use of targeted exome capture provides an efficient strategy sequence the complete coding regions (the exome) of the genome at much deeper read-depths without the high costs of the whole genome sequencing approaches. With the increasingly wide acceptance of high-throughput sequencing technologies in the bio-pharmaceutical industry, it becomes critical to have fast, robust, and high quality *in silico* methods for secondary and tertiary analyses of large amounts of exome sequencing data in an industrial environment.

We introduce a computational pipeline for the secondary analysis of short read exome resequencing data, including alignment, recalibration, variant calling, the *in silico* prediction of functional impacts of mutations, as well as exhaustive quality controls. The pipeline is highly parallelized to take advantage of the large number of CPU cores available in a high-performance compute cluster.

We validate called variants in regions with known variants and regions of varying coverage and base call qualities. Taking Sanger sequencing as a benchmark, the comparison shows 98% specificity and 88% to 100% sensitivity of our target capture, high-throughput sequencing, and *in silico* analysis pipeline. These results establish our pipeline as a valid means to generate hypotheses driving research into origins and therapeutics of cancer.

978-1-4673-2560-8/12/$31.00 ©2012 IEEE