Abstract

The introduction of next-generation sequencing (NGS) technologies enables scientists to analyze millions of DNA sequences in a single run. The hereby produced gigabytes of raw data need to be further analyzed in order to gain biological meaningful results. Although NGS has lowered the cost for whole genome sequencing dramatically, its application for high-throughput screening studies still remains expensive. Exome sequencing provides a more cost effective approach where only the protein coding regions of a genome is utilized to find mutations which cause and maintain human diseases.

Spurred by NGS technologies, new efficient and well designed bioinformatics tools emerged which are addressing different tasks in the downstream analysis of NGS data. Since combining these tools into an analysis pipeline greatly facilitates the interpretation of NGS results, an exome sequencing pipeline was developed in this thesis which connects all necessary analysis steps into a unified application. The pipeline supports input data generated by the NGS platforms Illumina and ABI SOLiD™, handles correct execution of all integrated tools, and automatically distributes computational expensive tasks on a high-performance computing (HPC) cluster. It performs quality statistics on raw and processed reads, allows users to trim and filter sequence reads, and aligns the processed reads to a reference genome. Post alignment analysis includes the calculation of alignment statistics, region filtering, and the detection of variants resulting in a list of potential disease driving candidates. The developed pipeline was applied in a joint project with clinical research partners to detect potential causes for Mendelian disorders.

The integration of well established tools and newly developed promising algorithms into a unified solution eases the analysis of NGS data and may provide a valuable method for detecting and investigating therapeutical targets of diseases such as cancer and hereditary disorders.

Keywords: next-generation sequencing, exome analysis, pipeline development, high-performance computing, distributed analyses

Kurzfassung

Durch die Entwicklung von ‘Next Generation Sequencing’ (NGS) wurde die Analyse von Millionen von DNA Sequenzen in einem einzigen Sequenzierdurchlauf ermöglicht. Die auf diesem Wege gewonnen Rohdaten erfordern weitere Analysen um biologisch aussagekräftige Resultate zu liefern. Trotz drastisch gesunkener Kosten für die Sequenzierung vollständiger Genome, bleiben die absoluten Kosten für vergleichende Parameterstudien hoch. Exom-Sequenzierung bietet eine kosteneffizientere Methode, welche nur Eiweiß kodierende Regionen des Genoms zur Detektion von krankheitsauslösenden und -relevanten Mutationen im Menschen heranzieht.

Verschiedenste bioinformatische Werkzeuge wurden entwickelt, um die unterschiedlichsten Aufgaben der Analyse von NGS Daten zu bewerkstelligen. Die gegenwärtige Dissertation beschäftigt sich mit der Kombination einiger dieser Werkzeuge zu einer Analysekette, welche alle notwendigen Analysen in eine einheitliche Applikation vereint. Die hierbei erstellte Software unterstützt Exom-Sequenzdaten der NGS Plattformen Illumina und ABI SOLiD^{TM}, stellt die korrekte Ausführung aller Werkzeuge sicher und verteilt rechnerisch aufwendige Aufgaben auf Hochleistungsrechner. Sie berechnet Qualitätsmerkmale der Sequenzdaten, ermöglicht Trimmen und Filtern von Sequenzen und detektiert die Position der aufgearbeiteten Daten im Referenzgenom. Folgeanalysen beinhalten die Berechnung von Alignment Statistiken, das Filtern anhand der Position im Genom und die Detektierung von Mutationen, welche in eine Liste von potentiellen Krankheitsauslösern resultieren. Die entwickelte Software wurde bereits in einer Kooperation mit einem klinischen Forschungspartner zur Identifikation von potentiellen Ursachen von Erbkrankheiten angewandt.

Die Integration von etablierten sowie innerhalb der gegenwärtigen Arbeit neu entwickelten Algorithmen in eine einheitliche Software erleichtert die Analyse von NGS Daten und kann eine wertvolle Methode zur Detektierung und Erforschung von therapeutischen Targets für Erbkrankheiten und Krebs darstellen.

Stichwörter: Next Generation Sequencing, Exomanalyse, Pipeline Entwicklung, Hochleistungsrechnen, verteilte Analysen