

NGS as tool for discovery disease related differences in circulating nucleic acids (can)

Julia Beck and Ekkehard Schütz¹

Abstract

Circulating nucleic acids (CNA) isolated from serum or plasma are playing an increasing role as biomarkers for various diseases, such as cancers, chronic inflammatory diseases and fetal aneuploidy. The trace amounts of these unique biomarkers present in serum, requires pre-amplification or high serum volumes and can then be investigated using next generation sequencing, which provides high numbers of DNA sequences. The aim of this study was to find signature CNA sequences from patients with breast cancer (BrCa) compared to healthy controls. CNA extracted from the serum of breast cancer patients and healthy controls was sequenced using on a Roche/454 high-throughput sequencing platform. The extracted trace amount of DNA per 200 μ L specimen was preamplified using a commercially available whole genome amplification (WGA) kit. The obtained product amplicons of 300 – 500 bp were molecular barcoded and sequenced. Repetitive elements present in CNA were detected and classified, and each repetitive element was normalized based on total sequence count or repeat count. Multivariate regression models were calculated using an information-theoretical approach and multimodel inference. Data from 26 BrCa patients with stages II to IV tumors and from 67 apparently healthy female controls were used as the training data set. Using a bootstrap method to avoid sampling bias, a five-parameter model was developed. When this model was applied to a validation data set consisting of patients with tumor stage I ($n = 10$) compared with healthy and nonmalignant disease controls ($n = 87$) a sensitivity of 90% at a diagnostic specificity level of 95% obtained.

In addition, a digital single molecule counting approach was used to assess cancer-related differential representation of non-repetitive genomic regions. In order to increase the information content of 454 (Titanium) sequencing in terms of countable sequences we developed and applied a SAGE-like pre -procedure to 56 BrCa and 35 control samples. Briefly, from each WGA fragments a 26 (24-28) bp tag was generated and the short tags were ligated to form concatemers that were then sequenced on Roche/454 platform. Instead of one countable element per sequence we obtained 12 countable tags per 454 read. After the origin of the CNA tags was investigated by local alignment analyses genomic regions that show differential representation between patients and controls were selected by applying a cluster analysis. Genomic clusters with an overrepresentation of CNA hits from patients were defined, of which about 25 regions with a total of about 110k non-repetitive bases was used. These clusters of unambiguous genomic regions yielded a sensitivity of up to 98% with no false positive result in simple classification, whereas 96% had two or more representation of the 25 regions.

¹Chronix Biomedical, Goetheallee 8, 37073 Göttingen, Germany

These studies underline the usefulness of next generation-high throughput sequencing of CNAs for the development of biomarkers for malignant diseases. Further improvement concerning the cost-efficiency of sequencing may allow a broader use of CNA sequencing for malignant and non-malignant disease discovery and therapeutic management as well.