Karolinska Institutet

Department of Medical Epidemiology and Biostatistics

Statistical Methods for the Detection, Analyses and Integration of Biomarkers in the Human Genome and Transcriptome

AKADEMISK AVHANDLING
som för avläggande av medicine doktorsexamen vid Karolinska Institutet offentligen försvaras i Atrium, Nobels väg 12B, Solna, Karolinska Institutet.

Fredagen den 26 September, 2014, kl 09.00

av

Chen Suo

Huvudhandledare:
Professor Yudi Pawitan
Karolinska Institutet
Department of Medical Epidemiology and Biostatistics

Bihandledare:
Doctor Stefano Calza
University of Brescia
Department of Molecular and Translational Medicine

Doctor Agus Salim
La Trobe University
Department of Mathematics and Statistics

Fakultetsopponent:
Professor Wolfgang Huber
European Molecular Biology Laboratory
Genome Biology Unit

Betygsnämnd:
Professor Mats Gustafsson
Uppsala University
Department of Medical Sciences
Cancer Pharmacology and Computational Medicine

Docent Keith Humphreys
Karolinska Institutet
Department of Medical Epidemiology and Biostatistics

Docent Erik Kristiansson
Chalmers University of technology
Department of Mathematical Science

Stockholm 2014

ABSTRACT

Most human diseases have been shown to have a genetic basis that is linked to regulation of gene expression at the transcriptional or post-transcriptional level. In the central dogma of biology, deoxyribonucleic acid (DNA) is transcribed to messenger ribonucleic acid (mRNA), and then translated into proteins; dysfunction in any of these processes may contribute to the development of disease. Sources of such potential irregularities include, but not limited to, the following: point mutations in DNA sequences, copy number alterations (CNAs) and abnormal mRNA and microRNAs (miRNAs) expression. MiRNAs are a type of non-coding RNA that inhibit the transcription and/or translation of specific target mRNAs. Current technologies allow the identification of biomarkers and study of the complex interplay between DNA, mRNA, miRNA and phenotypic variation. This thesis aims to tackle the statistical challenges that have arisen with the application of these technologies to investigate various genomic and transcriptomic alterations.

In study I, modified least-variant set normalization for miRNA microarray, a new algorithm and software were developed for microRNA array data normalization. The algorithm selects miRNAs with the least array-to-array variation as the reference set for normalization. The selection process was refined by accounting for the considerable differences in variances between probes. Data are provided to show that this algorithm results in better operating characteristics than other methods.

In study II, joint estimation of isoform expression and isoform-specific read distribution using multi-sample RNA-Seq data, a joint model and software were developed to estimate isoform-specific read distribution and gene isoform expression, using RNA-sequencing data from multiple samples. Observation of similarities in the shape of the read distributions solves the problem that the non-uniform read intensity pattern is not identifiable from the data provided by one sample.

In study III, integrated molecular portrait of non-small cell lung cancers, molecular markers at the DNA, mRNA and miRNA level that can distinguish between different histopathological subtypes of non-small cell lung cancer were identified. Additionally, using integrated genomic data including CNAs and mRNA and miRNA expression data, three potential driver genes were identified in non-small cell lung cancer, namely *MRPS22*, *NDRG1* and *RNF7*. Furthermore, a potential driver miRNA, *hsa-miR-944*, was identified.

In study IV, integration of somatic mutation, expression and functional data reveals potential driver genes predictive of breast cancer survival. An analytic pipeline to process large-scale whole-genome and transcriptome sequencing data was created, and an integrative approach based on network enrichment analyses to combine information across different types of omics data was proposed to identify putative cancer driver genes. Analysis of 60 patients with breast cancer provided evidence that patients carrying more mutated potential driver genes had poorer survival.