Unraveling expression and DNA methylation landscapes in cancer

Lydia Hopp

Summary

Cancer is a complex, heterogeneous disease and associated with a pluralism of distinct molecular events occurring on multiple layers of cell activity. It is a disease of genomic regulation driven by genetic and epigenetic mechanisms. Consideration of these regulatory levels is inevitable for understanding cancer genesis and progression. Improved high-throughput techniques developed in the last decades enable a highly resolved view on these mechanisms but at the same time the technologies produce an incredible amount of molecular data. Hence it needs advances in computational methods to master the data.

In this thesis we demonstrate how to cope with high-dimensional data to characterize molecular aspects of cancer. The main aim of this thesis is to develop and to apply bioinformatics methods to unravel molecular mechanisms, with special focus on gene expression and epigenetics, underlying cancer. Therefore, we selected two cancer entities, B-cell lymphoma and glioblastoma, for a more detailed, exemplary study.

Bioinformatics methods dealing with molecular cancer data have to tackle tasks like data integration, dimension reduction, data compression and proper visualization. One effective method that fulfills the mentioned tasks is self organizing map (SOM) machine learning, a technique to 'organize' complex, multivariate data. We present an analytic framework based on SOMs that aims at characterizing single-omics landscapes, here either regarding genome wide expression or methylation, to describe the heterogeneity of cancer on the molecular level. Molecular data of each sample is presented in terms of 'individual' maps, which enable their evaluation by visual inspection. The portrayal method also realizes comprehensive downstream analysis tasks such as marker selection and clustering of co-regulated features into modules, stratification of cases into subtypes, knowledge discovery, function mining and pathway analysis. Further, we describe how to detect and to correct outlier samples.

In a novel combining approach all these analytic tasks of the single-omics SOM are embedded in a workflow to integratively analyze gene expression and DNA methylation data of unmatched patient cohorts. We showed that this approach provides detailed insights into the transcriptome and methylome landscapes of cancer. Furthermore, we developed a new inter-omics method based on SOM machine learning for the combined analysis of gene expression and DNA methylation data obtained from the same patient cohort. The method allows the visual inspection of the data landscapes of each sample on a personalized and class-related level, where the relative contribution of each of both data entities can be tuned either to focus on expression or methylation landscapes or on a combination of both.

Using the single-omics SOM approach, we studied molecular subtypes of B-cell lymphoma based on gene expression data. The method disentangles tumor heterogeneity and provides suited markers for the cancer subtypes. We proposed a refined subtyping of B-cell lymphoma into four subtypes, rather than a previously assumed three-group classification. In a second application of the single-omics SOM we studied a gene expression data set concerning glioblastoma for which we confirmed an established four-subtype classification. Our results suggested a similar gene activation pattern as observed in the lymphoma study characterized by an antagonistic switching between transcriptional modes related to immune response and cell division.

Our integrative study on a larger lymphoma cohort comprising additional subtypes confirmed previous results about the role of stemness genes during development and maturation of B-cells. Their dysfunctions in lymphoma are governed by widespread epigenetic effects altering the promoter methylation of the involved genes, their activity status as moderated by histone modifications, and also by chromatin remodeling. We identified subtype-specific signatures that associate with epigenetic effects such as remodeling from transcriptionally inactive into active chromatin states, differential promoter methylation, and the enrichment of targets of transcription factors such as *EZH2* and *SUZ12*.

While studying the transcription of epigenetic modifiers in lymphoma and healthy controls, we found that the expression levels of nearly all modifiers are strongly disturbed in lymphoma and concluded that the epigenetic machinery is highly deregulated. Our results suggested that Burkitt's lymphoma and diffuse large B-cell lymphoma differ by an imbalance of repressive and poised promoters, which is associated with an imbalance of the activity of histone- and DNA-modifying enzymes.

Our inter-omics method was applied to a high-grade glioblastomas. Their expression and methylation landscapes were segmented into modes of co-expressed and co-methylated genes, which reflect underlying regulatory modes of cell activity. We found antagonistic methylation and gene expression changes between the *IDH1* mutated and *IDH1* wild type subtypes, which affect predominantly poised and repressed chromatin states. Therefore we assume that these effects deregulate developmental processes either by their blockage or by aberrant activation.

Our methods presented in this thesis enable a holistic view on high-dimensional molecular data collected in large-scale cancer studies. The examples chosen illustrate the mutual dependence of regulatory effects on genetic, epigenetic and transcriptomic levels. Our finding revealed that epigenetic deregulation in cancer must go beyond simple schemes using only a few modes of regulation. By applying the tools and methods described above to large-scale cancer cohorts we could confirm and supplement previous findings about underlying cancer biology.