Statistische Methoden in Genomik und Proteomik (IN5089)

Titel	Statistical Methods in Genomics and Proteomics	
Typ	Vorlesung mit Übung	
Credits	6 ECTS	
Lehrform/SWS	3V + 1Ü	
Sprache	Deutsch	
Modulniveau	Master	
Arbeitsaufwand	Präsenzstunden	60 Stunden
	Eigenstudium	120 Stunden
	Gesamtaufwand	180 Stunden
Angestrebte Lernergebnisse	Die Studierenden kennen die bei der Analyse von Hochdurchsatz-Daten aus Genomik und Proteomik auftretenden statistischen Herausforderungen wie z.B. das multiple Testen und das Dimensionsproblem (mehr Variablen als Beobachtungen: das sogenannte n<<p Problem) sowie die Kriterien zur geeigneten Auswahl statistischer Verfahren zur Kontrolle der Fehlerrate beim Testen oder zur Erstellung von Prädiktionsmodellen. Die Studierenden sind in der Lage, für die oben genannten Probleme bzw. zugehörige experimentelle Daten (wie z.B. Genexpressionsdaten und SNP-Daten) geeignete Modelle und Methoden für die statistische Analyse auszuwählen, die zugehörigen Daten entsprechend eigenständig zu analysieren und die Ergebnisse kritisch zu bewerten und zu interpretieren, ohne der Versuchung des „fishing for significance“ zu verfallen.	
Intended Learning Outcomes	Students are familiar with modern statistical challenges occurring during the analysis of genomics and proteomics high-throughput data, e.g., multiple testing, and the dimension problem (more variables than observations: the so-called n<<p problem) as well as criteria for appropriate selection of statistical methods to control error in testing processes or for creation of prediction models. Students are able to select appropriate models and methods for the statistical analysis of the problems mentioned above and for corresponding experimental data (e.g. gene expression data or SNP data), to analyze independently corresponding data, and to critically evaluate and to interpret these results without to succumb for “fishing for significance.”	

Inhalt	Technologische Innovationen in der Genomanalyse (Genomik) und Identifikation von Proteinen (Proteomik) ermöglichen die Untersuchung biologischer und biomedizinischer Fragestellungen mit Hilfe von so genannten Hochdurchsatz-Daten, die aus genomischen und proteomischen Experimenten resultieren. Die adäquate Analyse solcher Daten erfordert neue Methodiken in Biostatistik und Bioinformatik. Das Modul gibt eine Einführung und einen Überblick zu Problemen und Konzepten der statistischen Modellierung und statistischen Inferenz von hochdimensionalen Daten, die sich aus substantiellen Fragestellungen in molekularer Biologie und Biomedizin ergeben. Nach einer Einführung in den biologischen Hintergrund, liegt der Schwerpunkt auf der statistischen Modellierung und Analyse von Daten, die aus den betrachteten Experimenten resultieren. Die Themen im Einzelnen sind: • Biologischer Hintergrund • SNP-Daten, Einführung in die genetische Epidemiologie • Univariate Analysen • Multiples Testen • Statistisches Lernen • Penalisierte Regression, Random Forest • Schätzung des Prädiktionsfehlers • Validierung, gute statistische Praxis
Contents	Technological innovation in genome analysis (genomics) and identification of proteins (proteomics) allows the investigation of new biological and medical research questions using so-called high-throughput data resulting from genomic and proteomic experiments. Adequate analysis of such data requires new methods in Biostatistics and Bioinformatics. This module introduces problems and concepts of statistical modeling and statistical inference with high-dimensional data, which arise form research problems in molecular biology and biomedicine. After an introduction into the biological background, the main focus is on the statistical modeling and analysis of data, which stem from the considered experiments. Topics are: • Biological background • SNP-Data, introduction into genetic epidemiology • Univariate Analyses • Multiple Testing

	• Statistical Learning • Penalized regression, Random Forest • Estimation of prediction errors • Validation, good statistical practice
Prüfung	Prüfungsleistung (benotet): -Klausur: 120 min Wiederholungsklausur innerhalb von 6 Monaten nach der regulären Klausur. Details werden zu Beginn des Moduls bekannt gegeben. In der Klausur weisen die Studierenden nach, inwieweit sie die vorgestellten Modelle und Methoden verstanden haben, komprimiert wiedergeben sowie auf konkrete Daten anwenden und auf verwandte Problemstellungen übertragen können. In der Klausur werden Aufgaben gestellt, die z.B. die Erläuterung und Anwendung von Prädiktionsmethoden und Methoden zum multiplen Testen, die Klassifikation mit hochdimensionalen molekularen Daten, die Interpretation von R-Code-Auszügen und die Diskussion von Graphik-R-Outputs erfordern.
Examination	Examination requirements (graded): - written exam: 120 min A makeup exam will be offered within six months after the regular examination, details will be announced at the beginning of the course. Within the written exam, students demonstrate that they understand the presented models and methods, that they can reproduce and apply them as well as that they can apply these to real data and to transfer and to extend models and methods to similar problems. The written exam consist of questions , which, for instance, require the description and application of prediction methods and methods for multiple testing, the classification of high-dimensional molecular data, the interpretation of R code snippets and the discussion of graphical R output.
Literatur/Literature	
Medienformen	Beamer-Präsentation, Tafelpräsentation, Handout
Media	slides show, blackboard presentation, handouts
Lehr- und Lernmethode	Vorlesung, Übungen. Das Modul besteht aus einer Vorlesung und Rechnerübungen mit dem statistischen Programm R (als Tutorübungen). In der Übung werden Übungsblätter zur

	Anwendung der in der Vorlesung vorgestellten Methoden auf reelle Datensätze mit dem statistischen Programm R von den Studierenden unter Anleitung des Übungsleiters bearbeitet. Der Übungsleiter erklärt und kommentiert anschließend eine Musterlösung.
Teaching and Learning Methods	Lecture, tutorials In the tutorials the students work on exercise sheets on the application of the methods presented in the lecture to real datasets with the statistical program R under the supervision of the instructor. At the end the instructor explains and comments a possible solution.
Turnus	Wintersemester
Modulverantwortlicher	Prof. Dr. Anne-Laure Boulesteix
Dozenten	Prof. Dr. Anne-Laure Boulesteix