Seminarbaustein: Data Science mit 

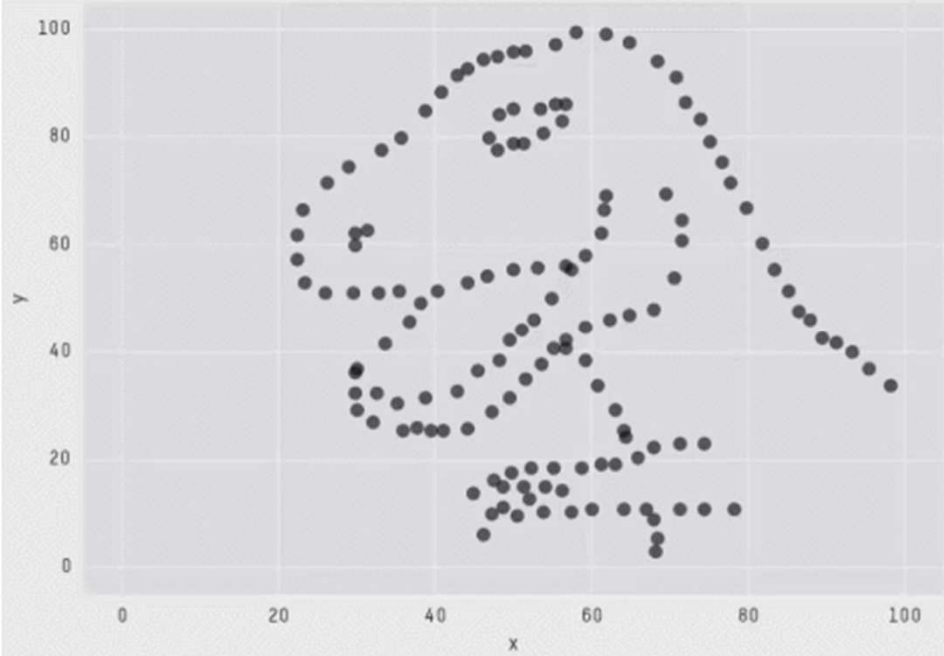
R ist das am häufigsten genutzte Werkzeug zur statistischen Datenanalyse und zur grafischen Darstellung der Ergebnisse. Es ist kostenfrei, flexibel einsetzbar und wird von einer großen Community ständig weiter entwickelt. Im Seminar erlernen Sie den Umgang damit und erhalten hilfreiche Unterstützung zur Anwendung.

Hier eine kurze Einführung.

Dozent: Dr. Andreas Busjahn

Denken mit Zahlen (aka Data Science)

Was versteckt sich in den Daten? (?Dinosaurier?)



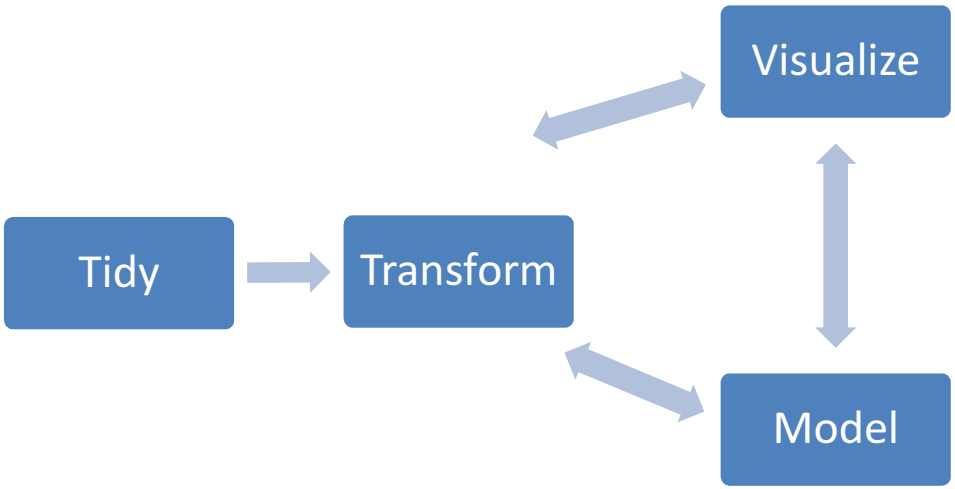
X Mean: 54.2659224
Y Mean: 47.8313999
X SD : 16.7649829
Y SD : 26.9342120
Corr. : -0.0642526

The R Project for Statistical Computing

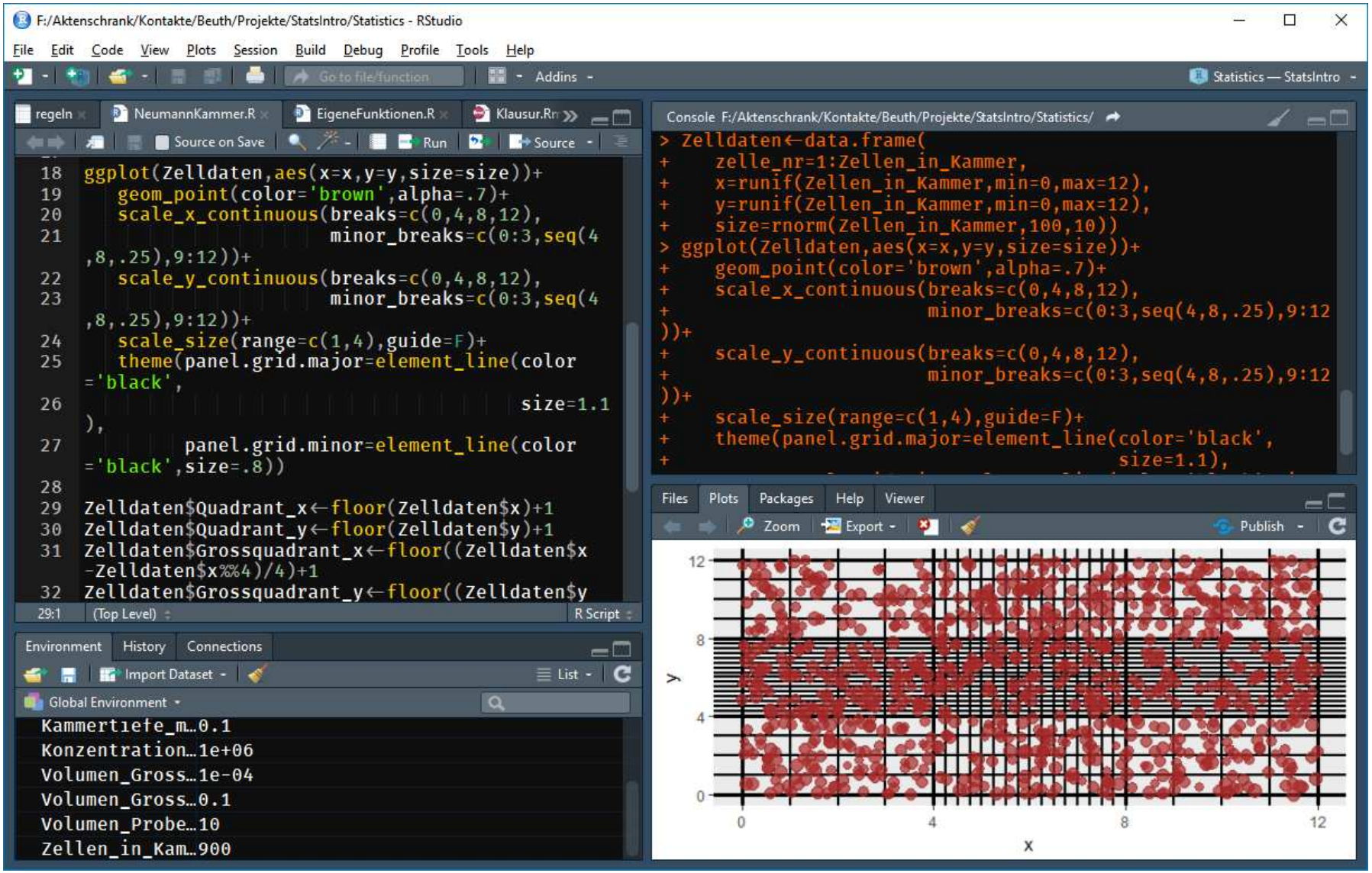
- Skriptsprache für statistische Analysen und Grafik
- Multiversum universeller und hochspezialisierter Lösungen
- Freie Software mit 13.741 Erweiterungspaketen
- Unterstützt durch IBM, Microsoft, Google, Oracle…
- Bioconductor: Teilprojekt für Bioinformatik, …omics
- Grammar of Graphics: Flexibles Grafiksystem

Workflow

- Import (1 bis 50 Zeilen)
- Datenvorbereitung (50 bis 1000 Zeilen)
- Visualisierungen (10 bis 100 Zeilen)
- Tests / Modelle (ca. 100 Zeilen, Tests: 1 bis 5)

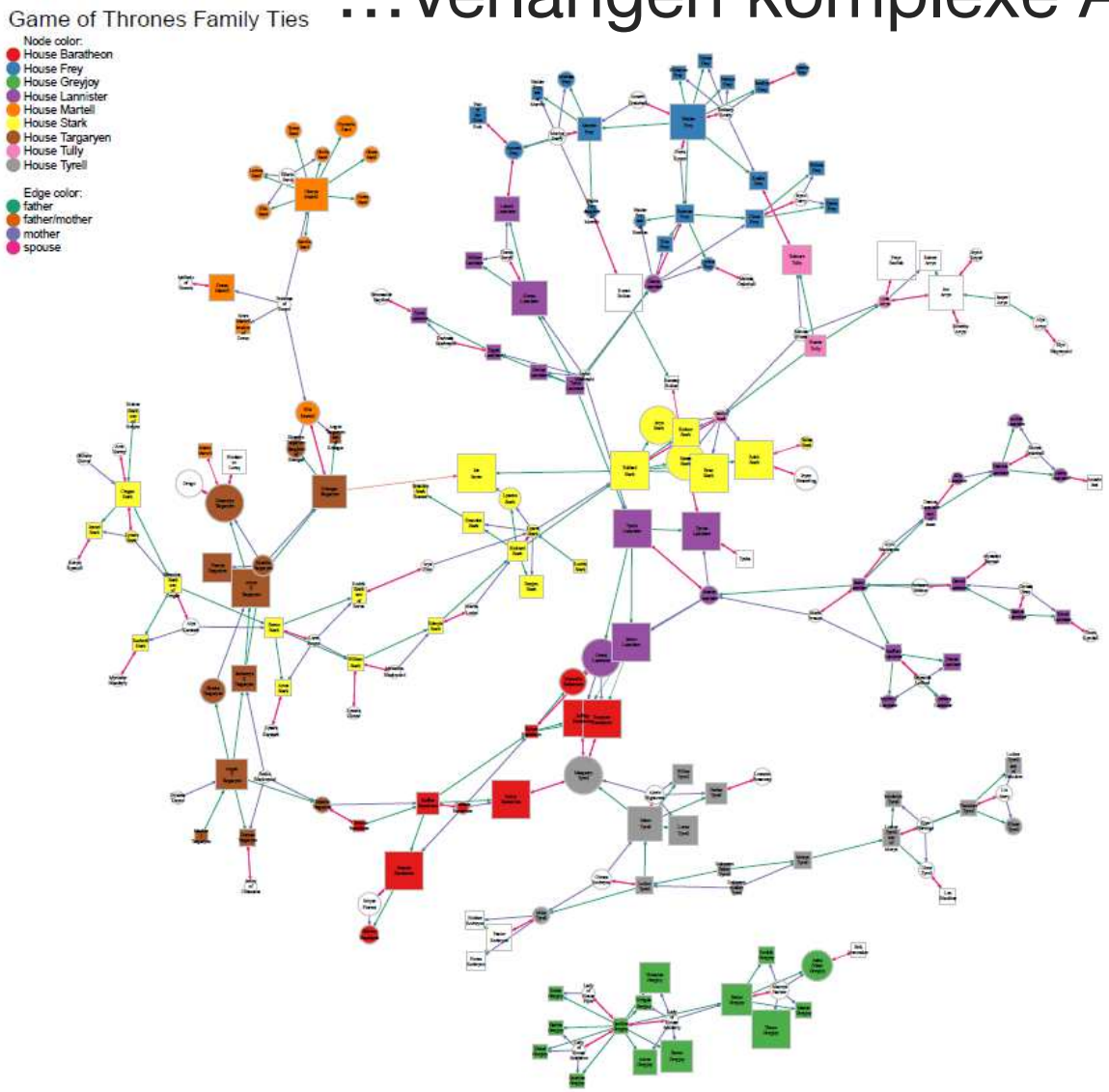


Graphische Benutzeroberfläche RStudio

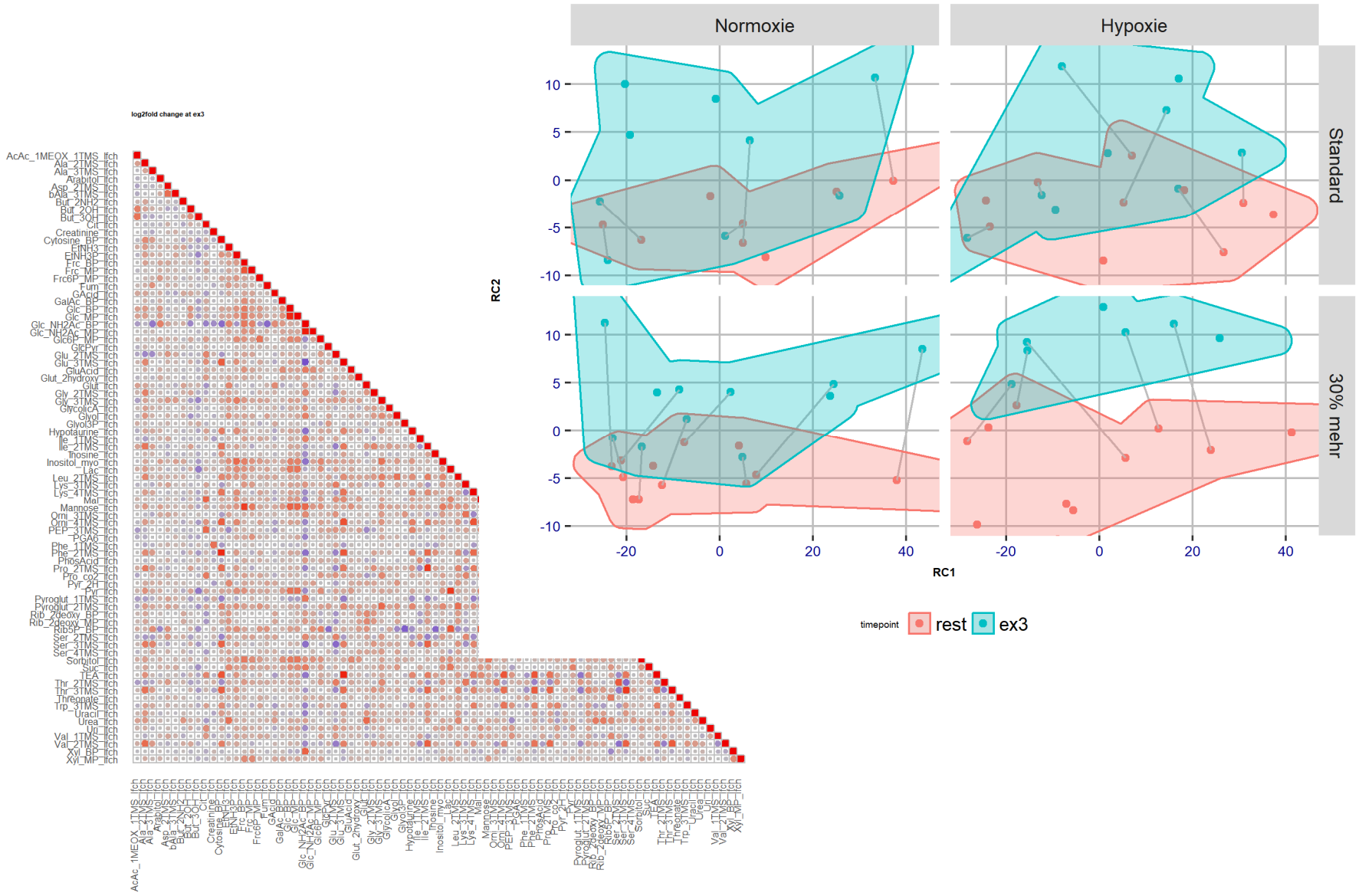


Komplexe Zusammenhänge…

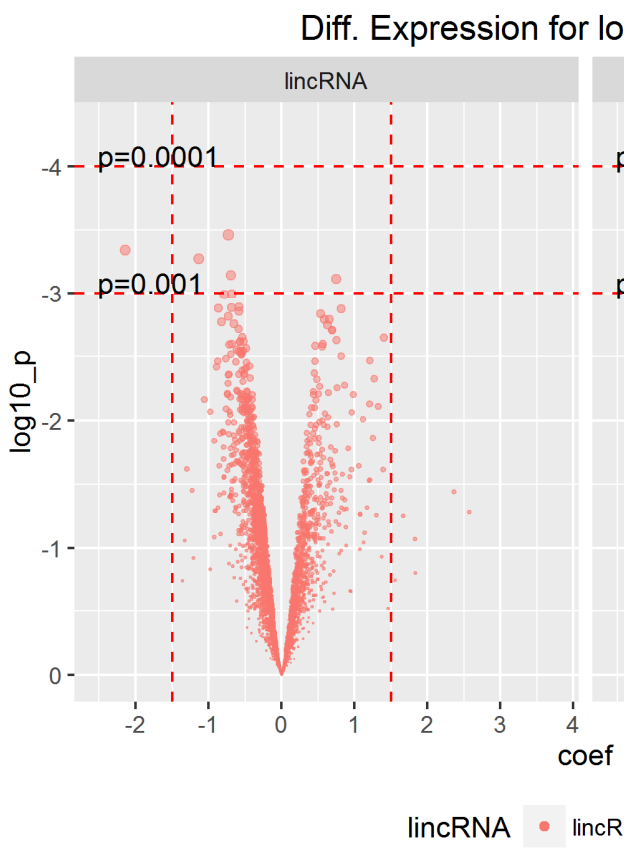
…Verlangen komplexe Analysen

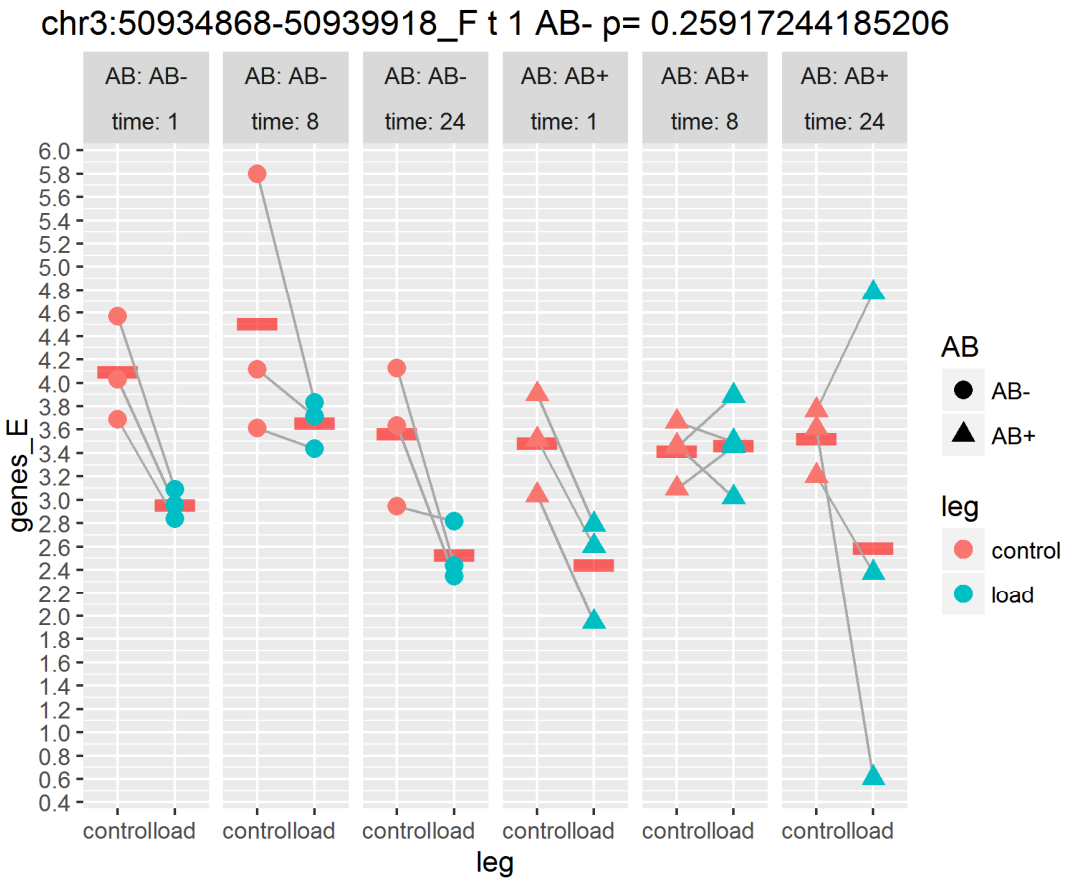


Metabolomics



Genexpression





Reproducible Research

- Jeder Schritt einer wissenschaftlichen Studie muss nachvollziehbar sein
- Angabe statistischer Methoden (PCA, OPLS…) nicht ausreichend, da hunderte Variationen
- KEIN Schritt der Analyse darf manuell erfolgen (copy/paste in Excel…)

- Lösung:
  - Scriptbasierte Analysen mit 
  - R Markdown als Reportgenerator: Text, Tabellen und Abbildungen -> LaTeX / Worddokument