Data analysis in the GMC

Dr. Helmut Fuchs

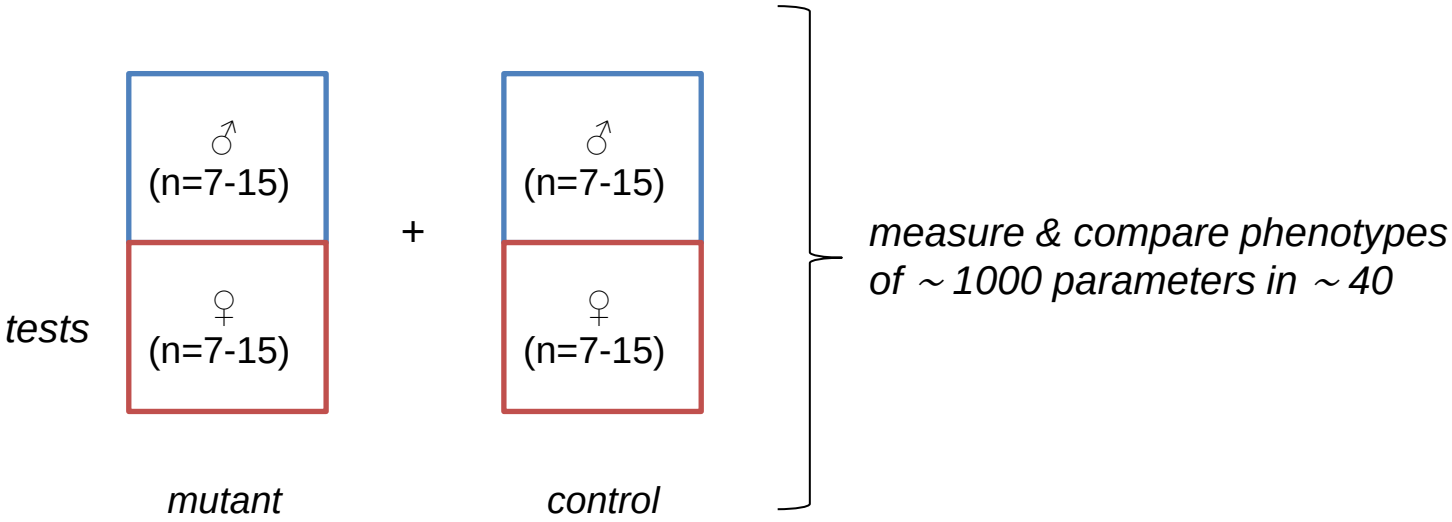
Institute of Experimental Genetics
Helmholtz Zentrum Munich

12.10.2016

Basics: What we do

Investigate genome-phenome relations by complex phenotyping of groups of mice differing in only one gene

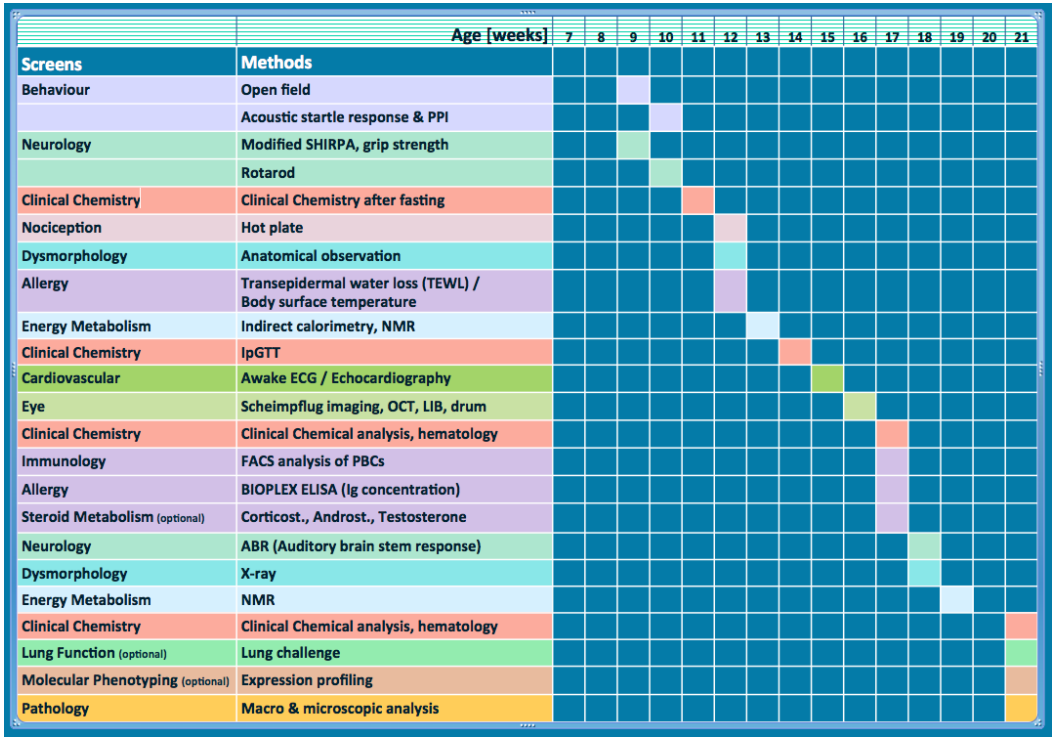
For each KO mouse line:



Controls are either *wild-type littermates of mutants* or *stock controls*

Basics: How we do it

- Standardized screen for generating hypothesis
- Each KO mouse line runs through a pre-defined workflow

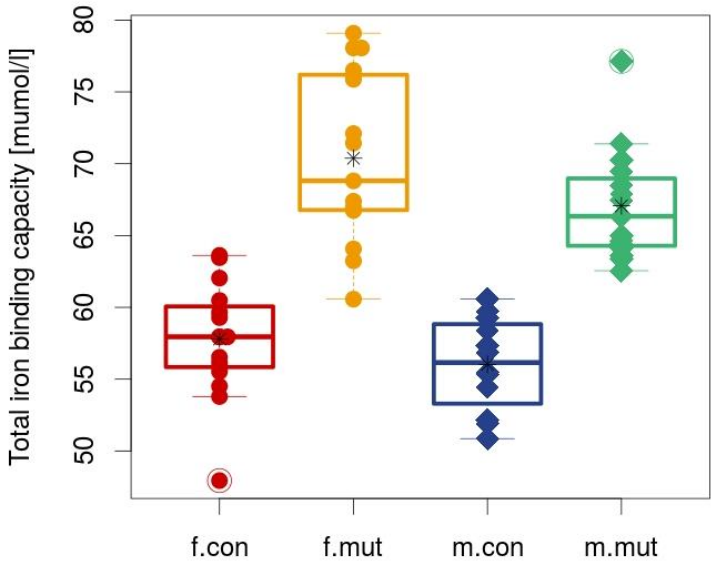


Screens	Methods	Age [weeks] 7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Behaviour	Open field															
	Acoustic startle response & PPI															
Neurology	Modified SHIRPA, grip strength															
	Rotarod															
Clinical Chemistry	Clinical Chemistry after fasting															
Nociception	Hot plate															
Dysmorphology	Anatomical observation															
Allergy	Transepidermal water loss (TEWL) / Body surface temperature															
Energy Metabolism	Indirect calorimetry, NMR															
Clinical Chemistry	IpGTT															
Cardiovascular	Awake ECG / Echocardiography															
Eye	Scheimpflug imaging, OCT, LIB, drum															
Clinical Chemistry	Clinical Chemical analysis, hematology															
Immunology	FACS analysis of PBCs															
Allergy	BIOPLEX ELISA (Ig concentration)															
Steroid Metabolism (optional)	Corticost., Androst., Testosterone															
Neurology	ABR (Auditory brain stem response)															
Dysmorphology	X-ray															
Energy Metabolism	NMR															
Clinical Chemistry	Clinical Chemical analysis, hematology															
Lung Function (optional)	Lung challenge															
Molecular Phenotyping (optional)	Expression profiling															
Pathology	Macro & microscopic analysis															

- For each KO line and each test:
  - Collect data in database
  - Run automated analysis with R-scripts
  - Interpretation of data

Examples: Boxplots

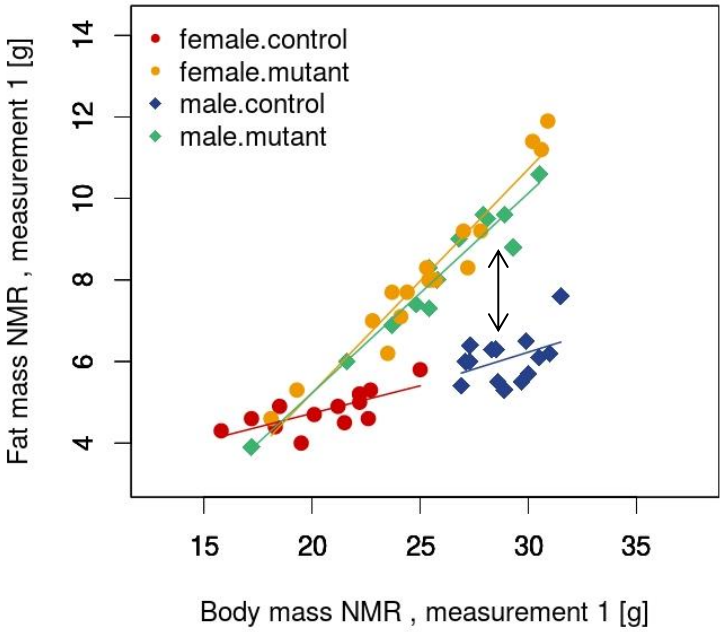
- Suitable for the analysis of one numerical variable
- Advantages:
  - Depiction of single data points → outlier identification
  - Display of the variability / spread in the data
  - Marking of statistical measures (mean, median, quartiles) gives hint on data distribution



- If you use barplots to describe the data, do *NOT use the SEM as error bars (it does not describe the spread in the data)*

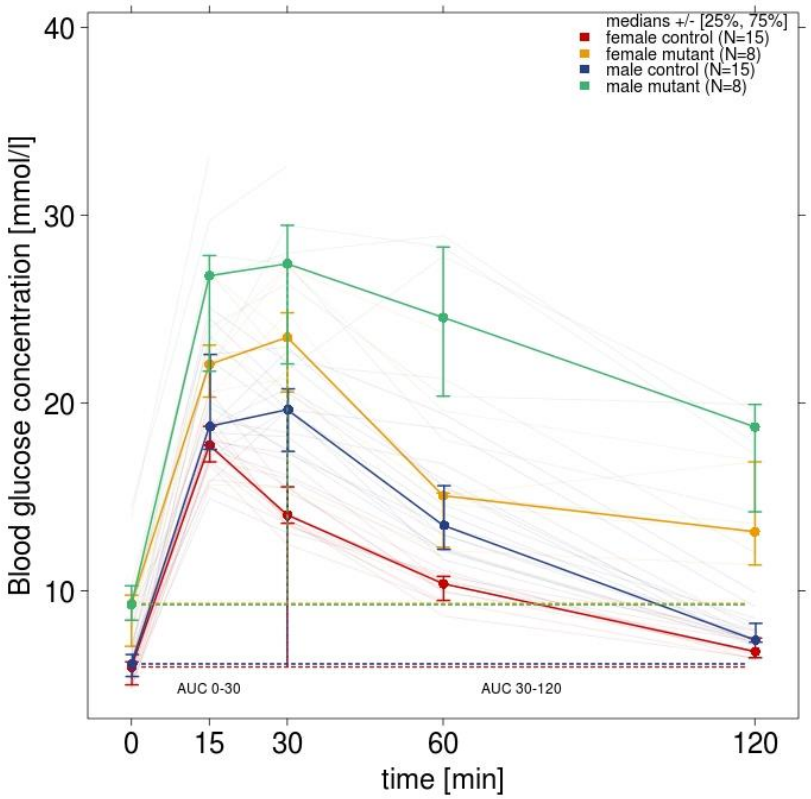
Examples: Scatterplots

- Display of correlation of two numerical variables
  → Interpretation of slope and regression line



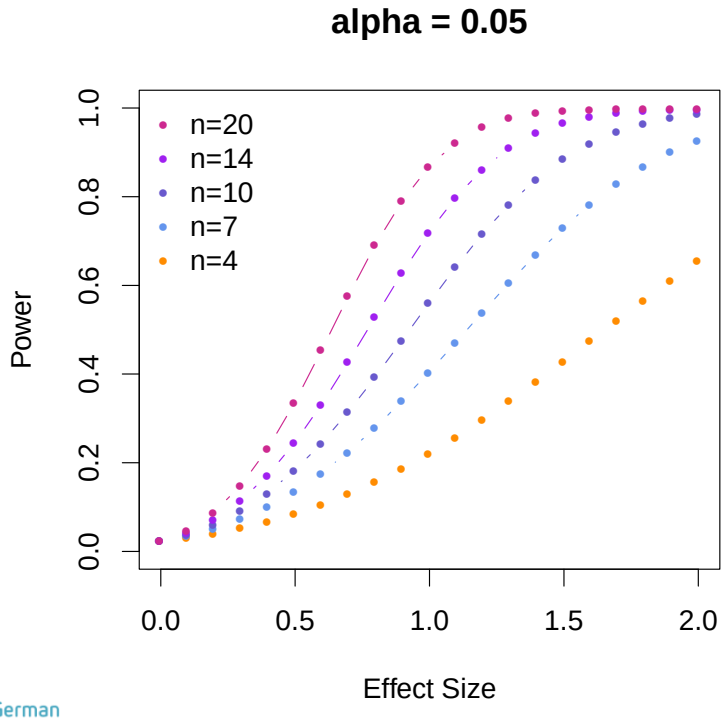
Examples: IPGTT line plot

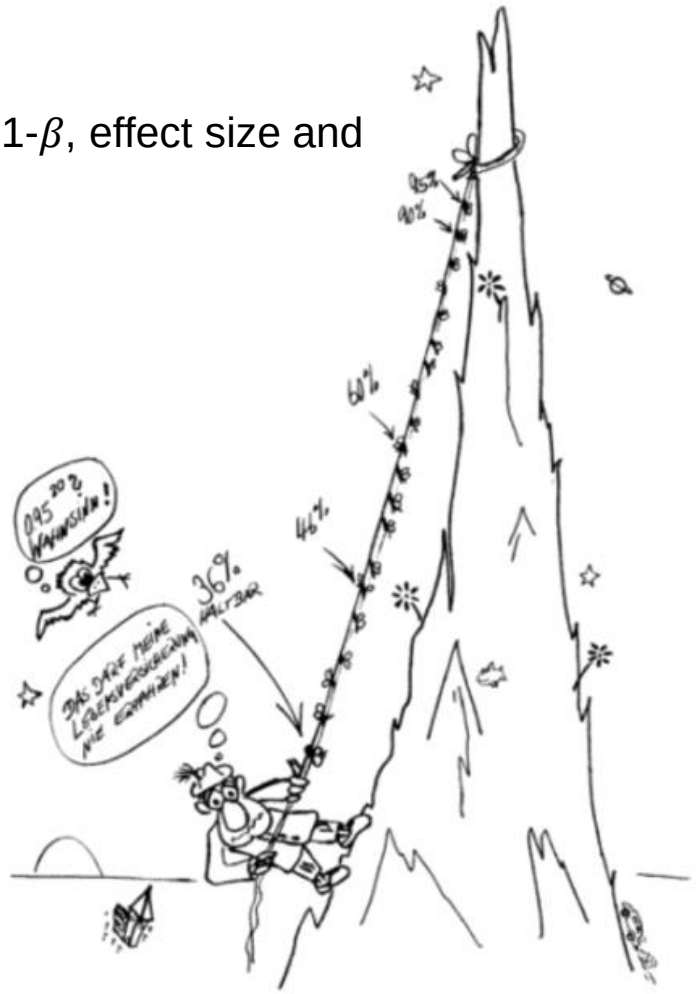
- Suitable for one variable measured over several time points
- Interpretation of the AUC of 0-30 and AUC 30-120



Important Issues

1. Descriptive vs. Inferential statistics
2. Multiple testing issue:
   - The more parameters are tested, the higher the probability of finding false positives is
3. Sample Size
   - Trade off between significance level α , power 1-β, effect size and welfare/costs





Important Issues

4. Every variable needs its own analysis depending on different assumptions
   - Normal distribution ⇔ any distribution
   - Quantitative ⇔ qualitative characteristics
   - Single time point ⇔ repeated measurements
   - Comparison of two groups ⇔several groups

Example Scatterplot:

	female		male		Linear model			
	control	mutant	control	mutant	sex	genotype	body mass	sex:genotype
	n=15	n=15	n=15	n=15				
	mean ± sd	mean ± sd	mean ± sd	mean ± sd	p-value	p-value	p-value	p-value
Body mass NMR	19.6 ± 1	19.7 ± 1	24.9 ± 1.5	25.4 ± 1.7	< 0.001	0.338	NA	0.575
Fat mass NMR	4.6 ± 0.4	4.8 ± 0.4	5.4 ± 0.4	5.6 ± 0.6	0.007	0.146	< 0.001	0.679
Lean mass NMR	12 ± 0.7	11.9 ± 0.5	16 ± 1.1	16.2 ± 1	< 0.001	0.309	< 0.001	0.578

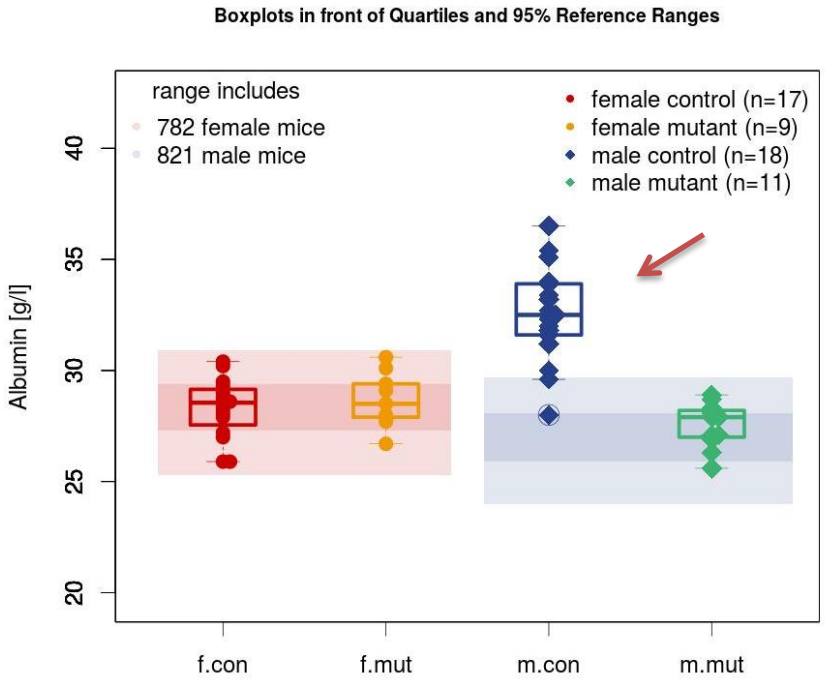
Include body weight as covariate

Important Issues

5. Metadata Influence (environment, settings ,...)
   - e.g. experimenter, daytime, age
   - If possible keep these factors constant, if not capture metadata

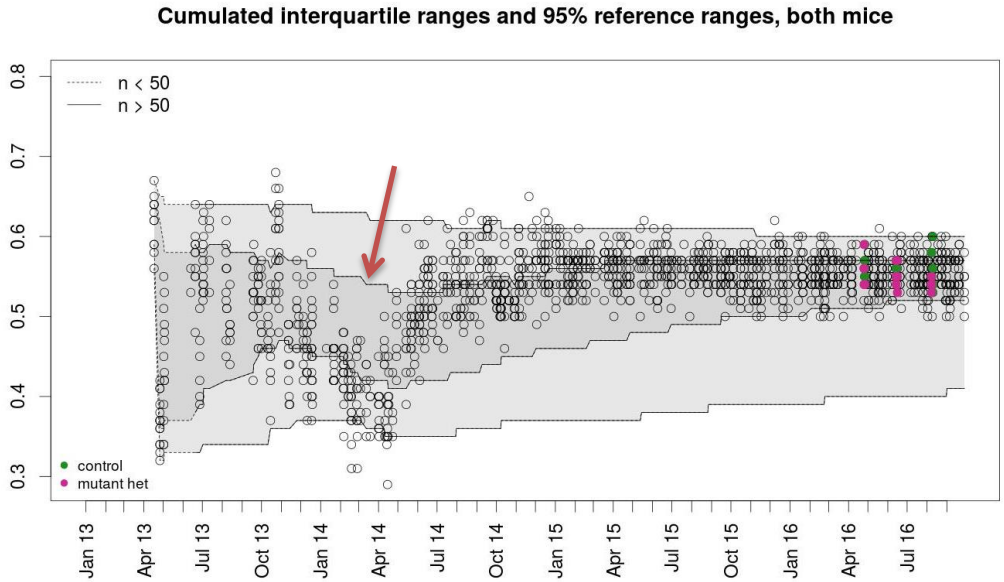
6. Biological relevance vs. significance
   - P-value is not „everything“ (arbitrary)
   - Depends strongly on sample size, multiple testing issue, correlation between variables
   - Even a non-significant results can be biological relevant → the p-value just represents the likelihood that the observed difference (or even a larger one) may have occurred by chance

Quality control



- check if controls are in a normal range

- check for metadata splits



German Mouse Clinic