UNIVERSITÄT
LEIPZIG

Medizinische Fakultät

LIFE molecular data

New molecular data sets in LIFE-Adult and respective research programmes

LIFE-Tag

Leipzig, 12th May 2026

Prof. Dr. Markus Scholz (IMISE)

imise.

Background

- High **molecular-genetic research focus** of LIFE study center
- Molecular assessments are **expensive**
- Initial funding allowed only limited number of samples / techniques
- **Step-by-step expansion of data sets** / measurements by additional **third party funding** or **industry cooperation**
- AG Genstat responsible for
  - Planning of measurement campaign (selection of cases / techniques)
  - Pre-processing, QC analysis
  - Integration + Harmonization
  - Annotation
  - Management of the data
  - Pipeline building for reproducible research

Motivation - SYMPATH project

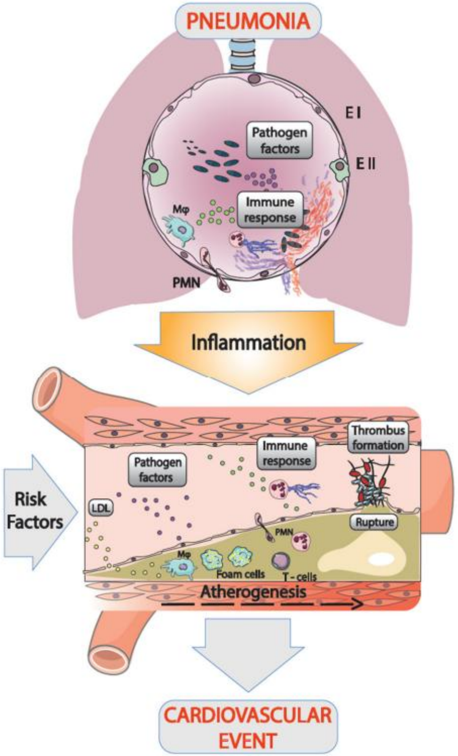
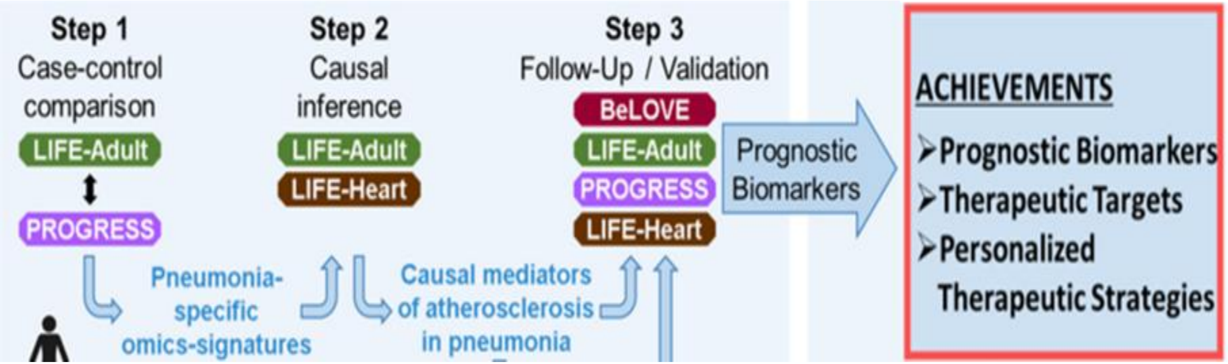


Figure 1 - Pathomechanistic relations between pneumonia, atherosclerosis and cardiovascular events

Identify molecular mediators of acute pulmonary infection towards cardiovascular disease



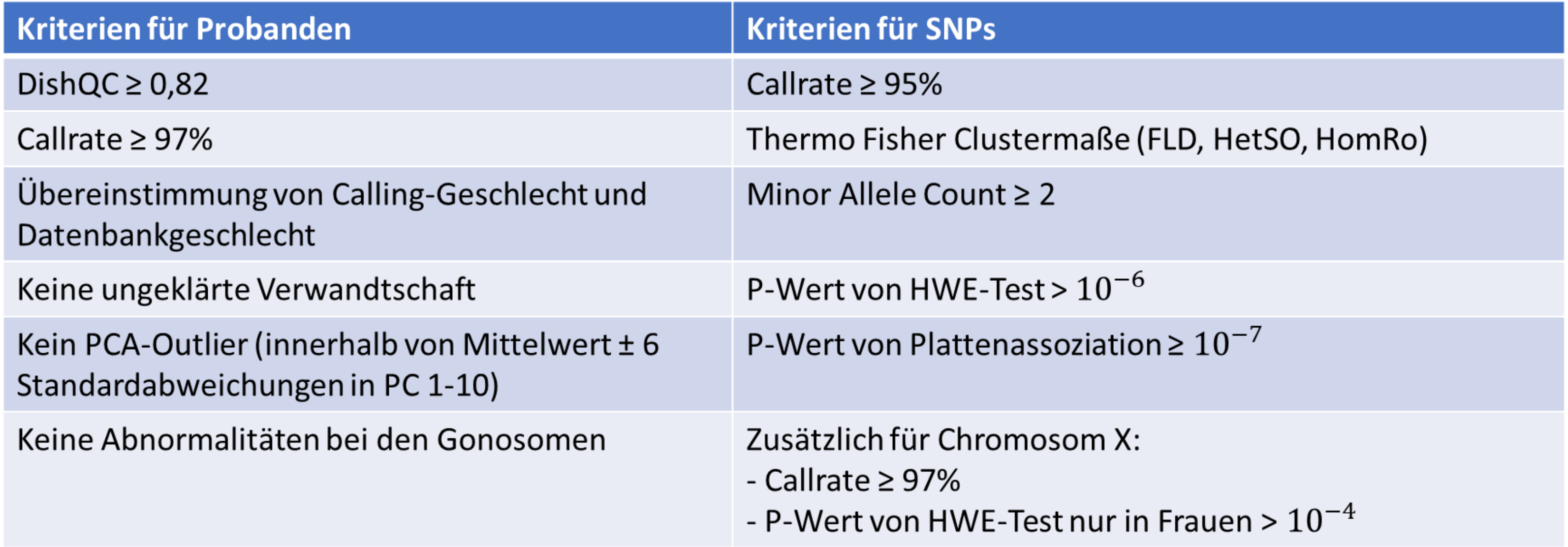
Changes due to infection

Differences between CAD case/control → Causal inference

Genetics

- Current LIFE-Adult SNP platform: **Thermo Fisher Axiom CEU1**
- Round 3 with LIFE and IMISE funds → **N~7,700**
- Round 4 → Completion with SYMPATH third party funds + IMISE budget

Quality control criteria



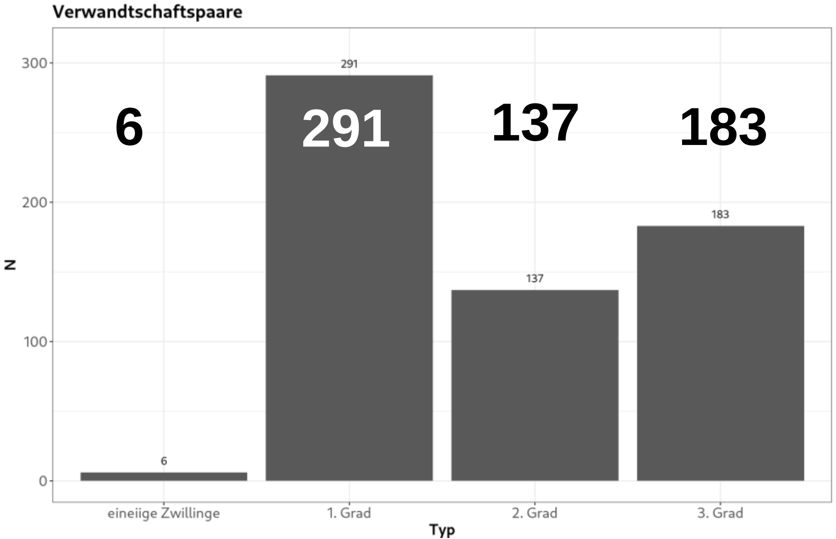
Kriterien für Probanden	Kriterien für SNPs
DishQC ≥ 0,82	Callrate ≥ 95%
Callrate ≥ 97%	Thermo Fisher Clustermaße (FLD, HetSO, HomRo)
Übereinstimmung von Calling-Geschlecht und Datenbankgeschlecht	Minor Allele Count ≥ 2
Keine ungeklärte Verwandtschaft	P-Wert von HWE-Test > 10^{-6}
Kein PCA-Outlier (innerhalb von Mittelwert ± 6 Standardabweichungen in PC 1-10)	P-Wert von Plattenassoziation ≥ 10^{-7}
Keine Abnormalitäten bei den Gonosomen	Zusätzlich für Chromosom X: - Callrate ≥ 97% - P-Wert von HWE-Test nur in Frauen > 10^{-4}

Sample drop out
N=109 of 10,098

Feature (SNP) drop out
M~35k of ~590k

Some QC results

Relatedness



6

291

137

183

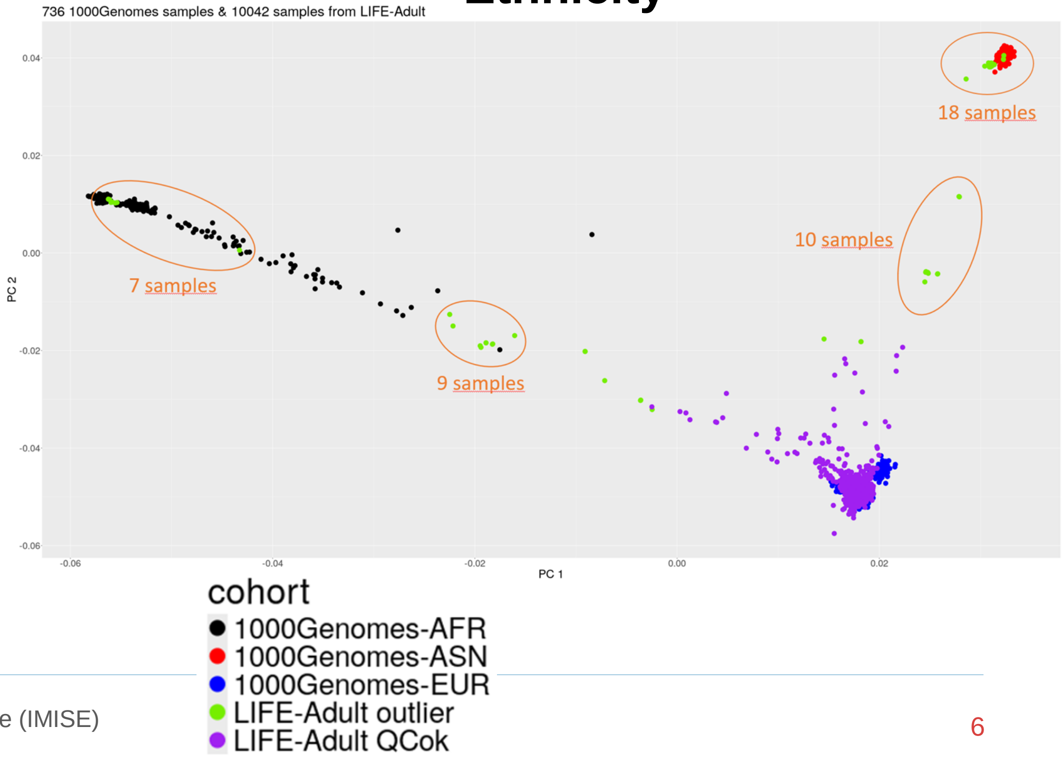
Monocygotic twins

1st degree relatives

2nd degree relatives

3rd degree relatives

Ethnicity



18 samples

10 samples

7 samples

9 samples

Next step - Imputation

Axiom CEU 1 SNP-Array mit 587.352 SNPs

Diverse Calling-QC-Schritte

QC-Datensatz: 537.903 SNPs auf Autosomen
13.241 SNPs auf Chr. X

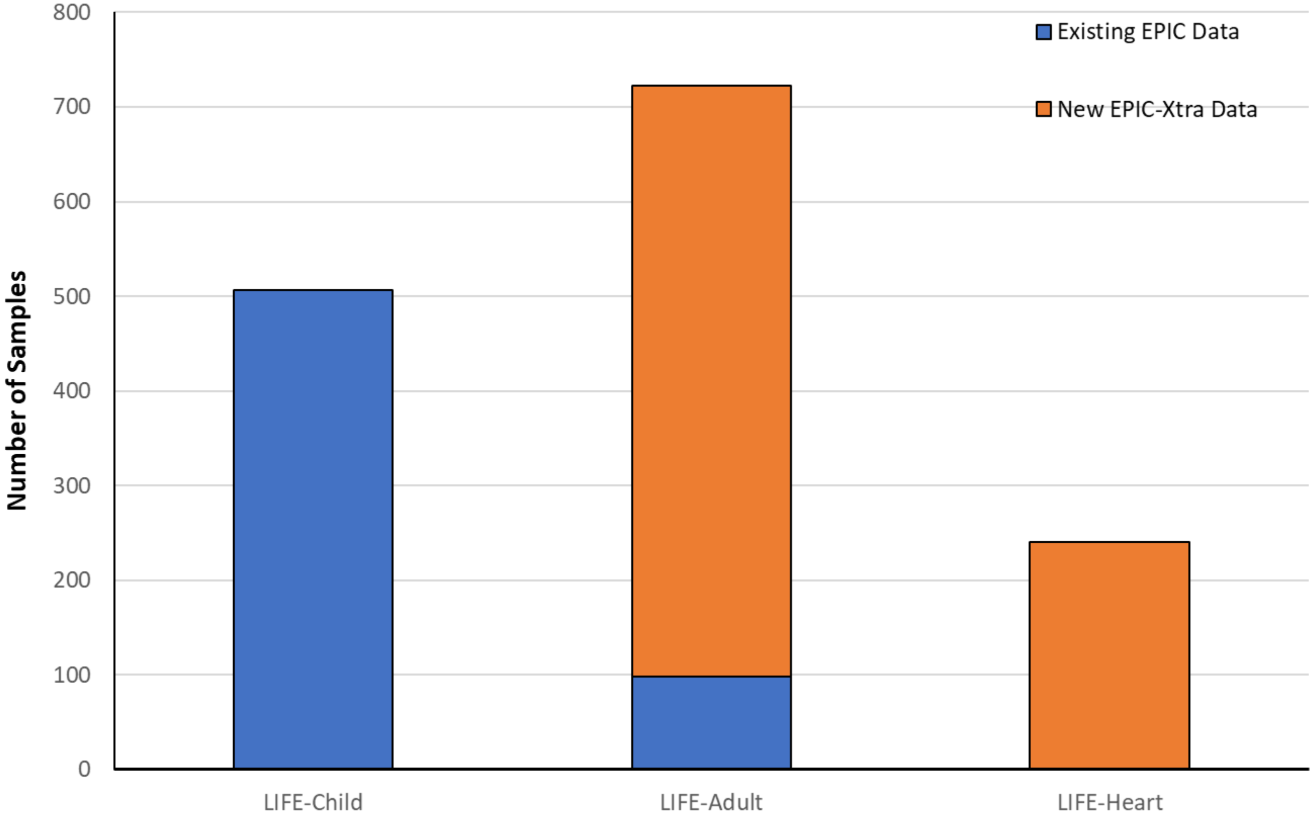
Helmholtz Imputation Server

Reference panels

Imputationsreferenz	SNPs	Gesamtanzahl SNPs
HRC	Autosomen: 11.898.222 Chr. X: 403.662	12.301.884
1000 Genomes phase 3	Autosomen: 16.834.137 Chr. X: 573.731	17.407.868

Final SNP numbers for analysis

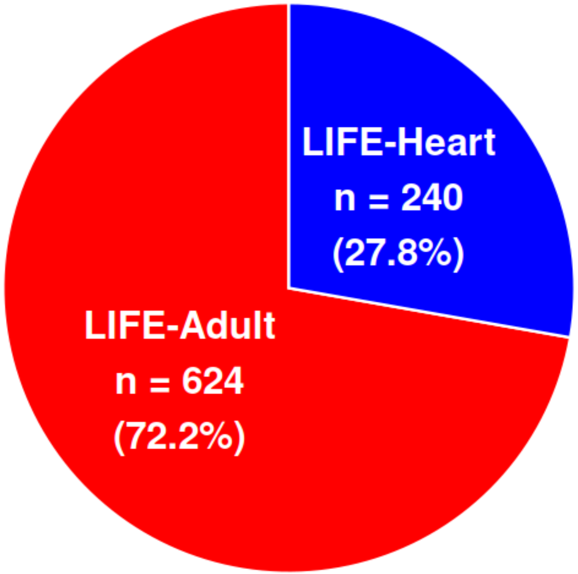
DNA Methylation



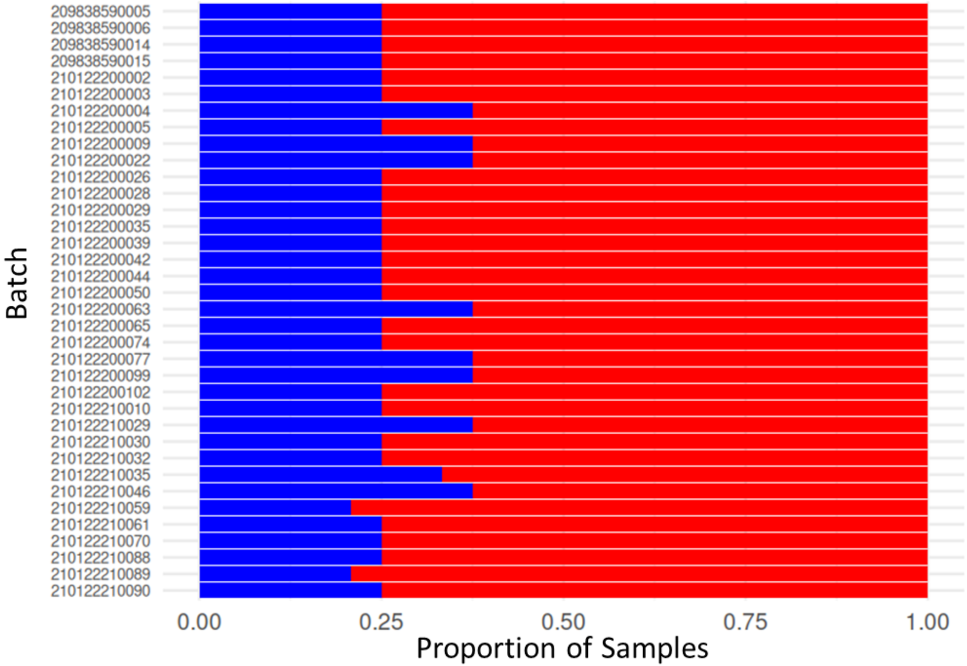
Funded by IMISE budget + SYMPATH

DNA Methylation

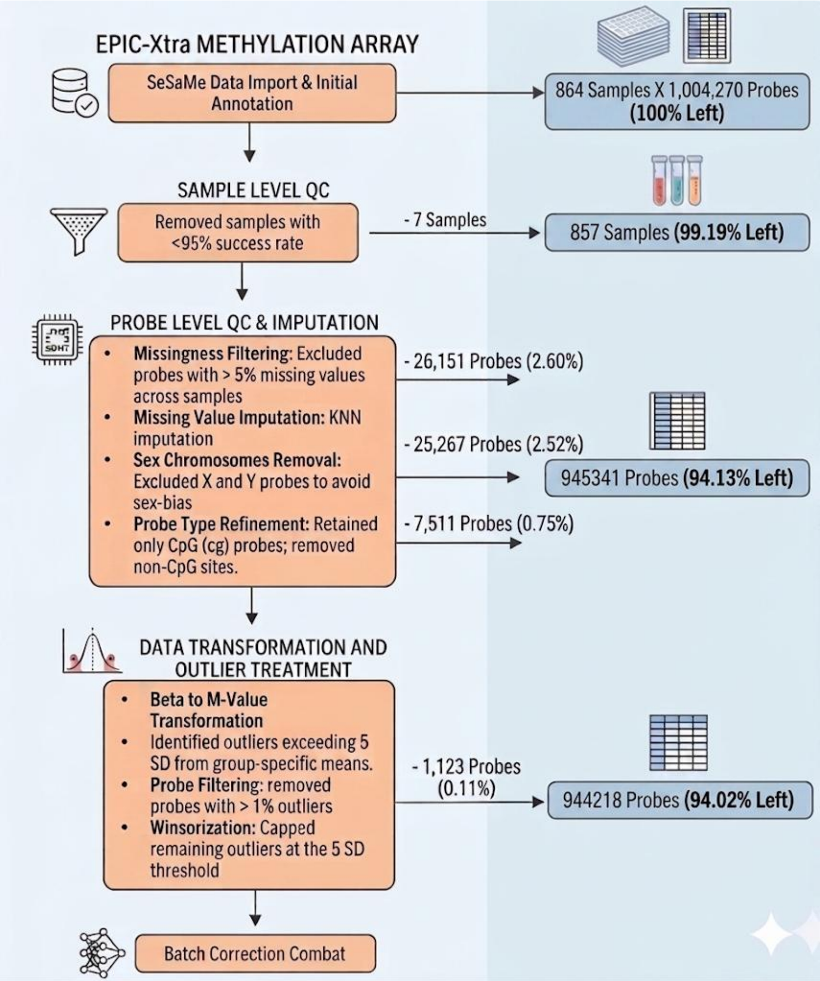
Total Sample Distribution



Sample Distribution Across Batches



Pre-processing



EPIC-Xtra METHYLATION ARRAY

SeSaMe Data Import & Initial Annotation

864 Samples X 1,004,270 Probes **(100% Left)**

SAMPLE LEVEL QC

Removed samples with <95% success rate

- 7 Samples

857 Samples **(99.19% Left)**

PROBE LEVEL QC & IMPUTATION

- **Missingness Filtering:** Excluded probes with > 5% missing values across samples
- **Missing Value Imputation:** KNN imputation
- **Sex Chromosomes Removal:** Excluded X and Y probes to avoid sex-bias
- **Probe Type Refinement:** Retained only CpG (cg) probes; removed non-CpG sites.

- 26,151 Probes (2.60%)

- 25,267 Probes (2.52%)

- 7,511 Probes (0.75%)

945341 Probes **(94.13% Left)**

DATA TRANSFORMATION AND OUTLIER TREATMENT

- **Beta to M-Value Transformation**
- Identified outliers exceeding 5 SD from group-specific means.
- **Probe Filtering:** removed probes with > 1% outliers
- **Winsorization:** Capped remaining outliers at the 5 SD threshold

- 1,123 Probes (0.11%)

944218 Probes **(94.02% Left)**

Batch Correction Combat

Initial Data:

864 Samples, 1,044,270 Probes

After Preprocessing:

857 Samples, 944,218 Probes

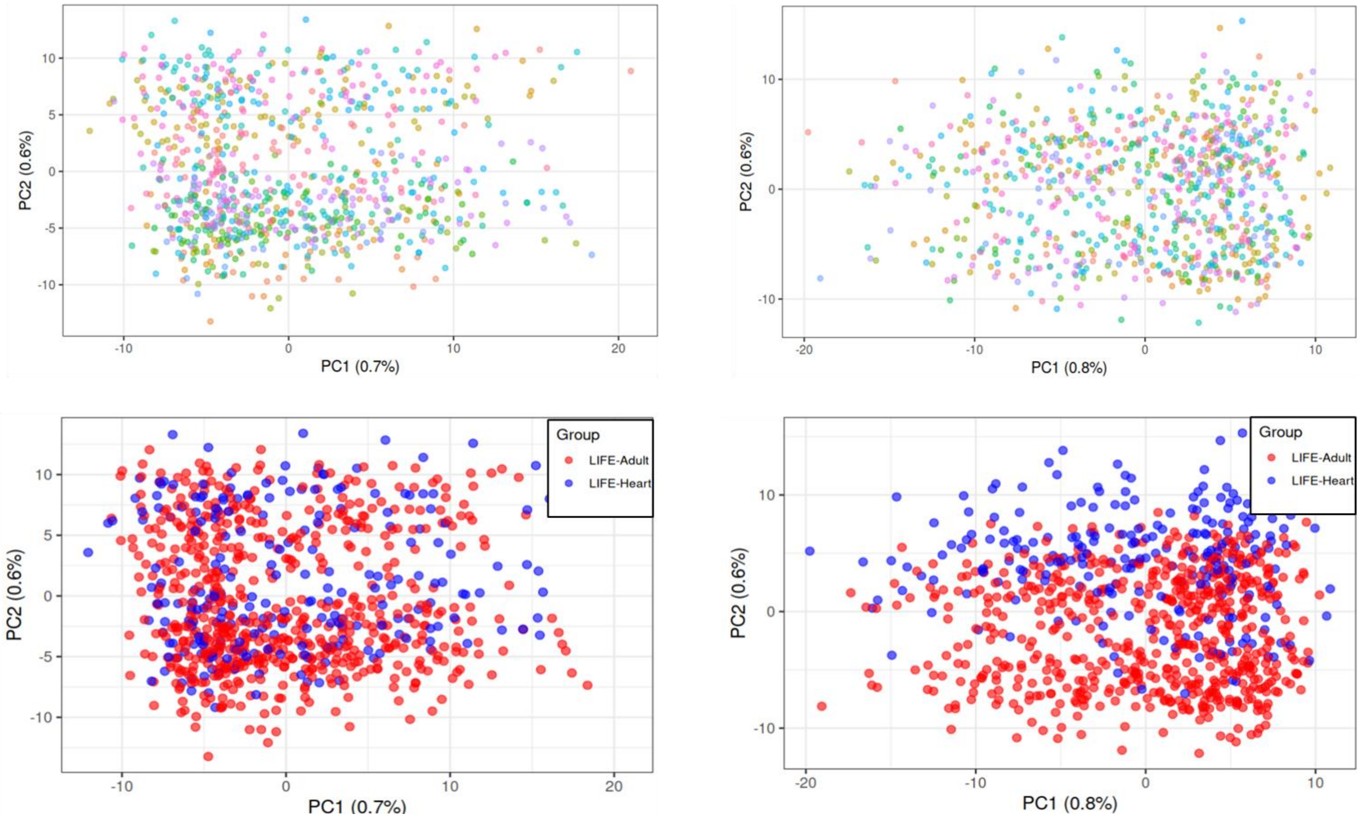
Batch correction

Before Batch Correction

After Batch Correction

Colored by Batches

Colored by Groups

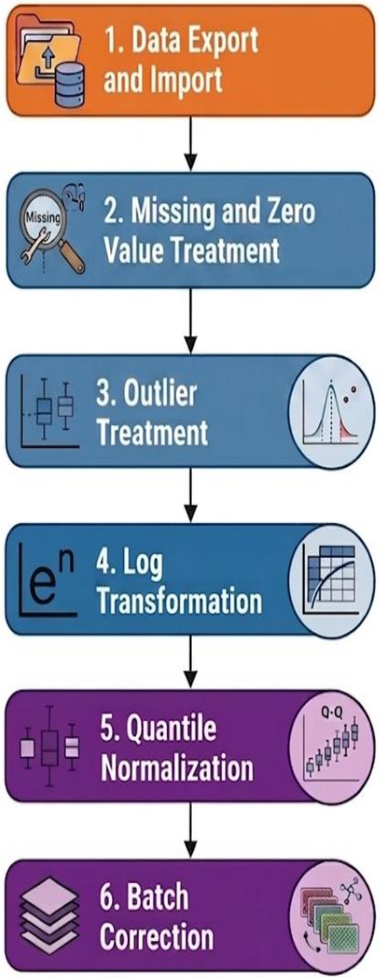


Better Adult / Heart contrasts after batch correction

Metabolomics

- Sphingolipids (funded by SYMPATH)
  - EDTA plasma
  - Measured by Kleuser lab Berlin
  - 1306 samples, 400/400 from LIFE-Adult/Heart
  - 46 analytes
- Biocrates (funded by SYMPATH)
  - EDTA plasma
  - Measured by ILM, LC-MS
  - 1284 samples, ~400/400 from LIFE-Adult/heart
  - 1437 analytes
- Metabolon (company collaboration)
  - Serum, plasma, urine
  - N=6000 planned

Pre-processing Biocrates



1. Data Export and Import

2. Missing and Zero Value Treatment

3. Outlier Treatment

4. Log Transformation

5. Quantile Normalization

6. Batch Correction

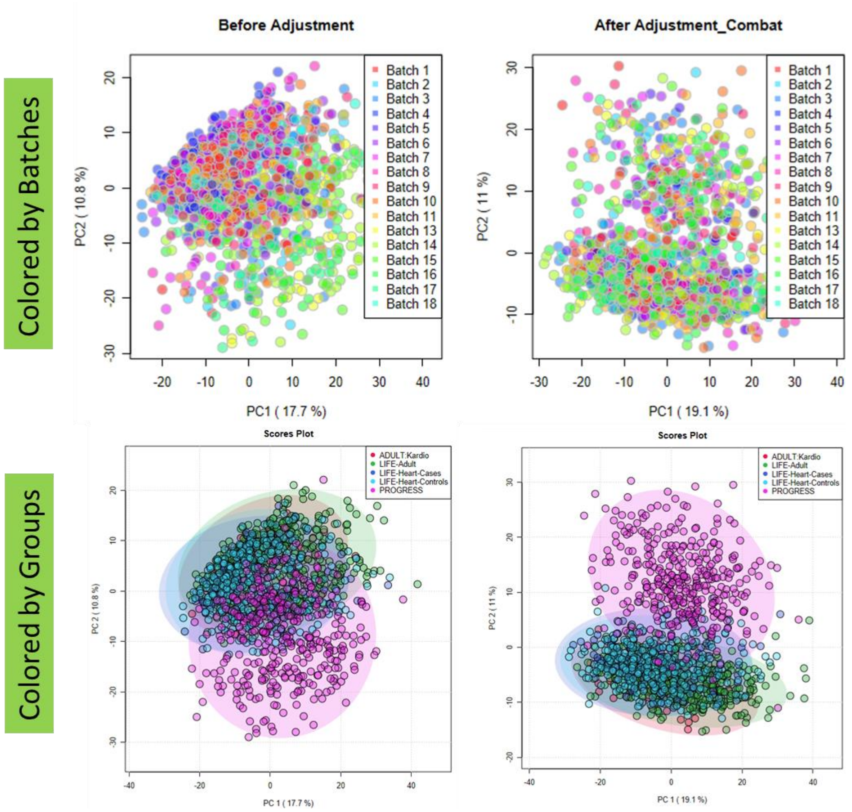
Diagram Key

🟧 Biocrates WebIDQ Software

🟪 MetaboAnalyst

🟦 Local pipeline in R

Batch Correction



UNIVERSITÄT
LEIPZIG

Medizinische Fakultät

THANK YOU VERY MUCH!
QUESTIONS?

markus.scholz@imise.uni-leipzig.de
https://www.genstat.imise.uni-leipzig.de/
@GenStatLeipzig