Import und Repräsentation hochdimensionaler (Omics-)Daten

Christian Knell

Lehrstuhl für Medizinische Informatik (FAU Erlangen-Nürnberg)

05.08.2016

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG
MEDIZINISCHE FAKULTÄT

TranSMART-Tutorial

- TranSMART for Beginners:

  A Practical Hands-On-Training

- Datum: 28.08.2016 von 14:00 – 17:15 Uhr

- Inhalt: Einführung, ETL, Analysen, RESTful API

- Weitere Informationen unter:

  http://www.hec2016.eu/tutorials.html

Was sind hochdimensionale Daten?

- Anzahl der Beobachtungen (n) wesentlich kleiner als die Anzahl der Variablen (p)

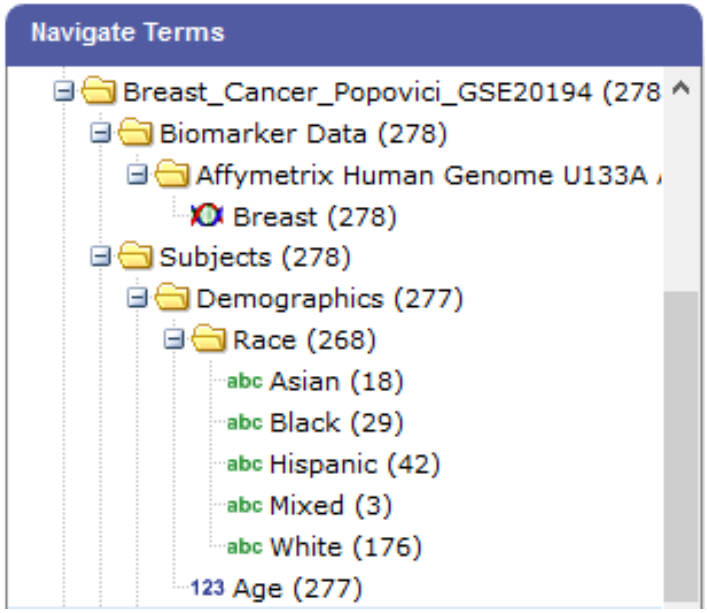
n<<p

- Medizin: speziell Omics-Bereich
  - Genom
  - Transkriptom
  - Proteom
  - Metabolom

TranSMART Omics-Integration

- TranSMART unterscheidet:
  - Kategorische / diskrete Daten (abc): Geschlecht
  - Numerische / kontinuierliche Daten (123): Alter
  - Hochdimensionale Daten ( )
- Darstellung der Daten

  in hierarchischem i2b2-Baum



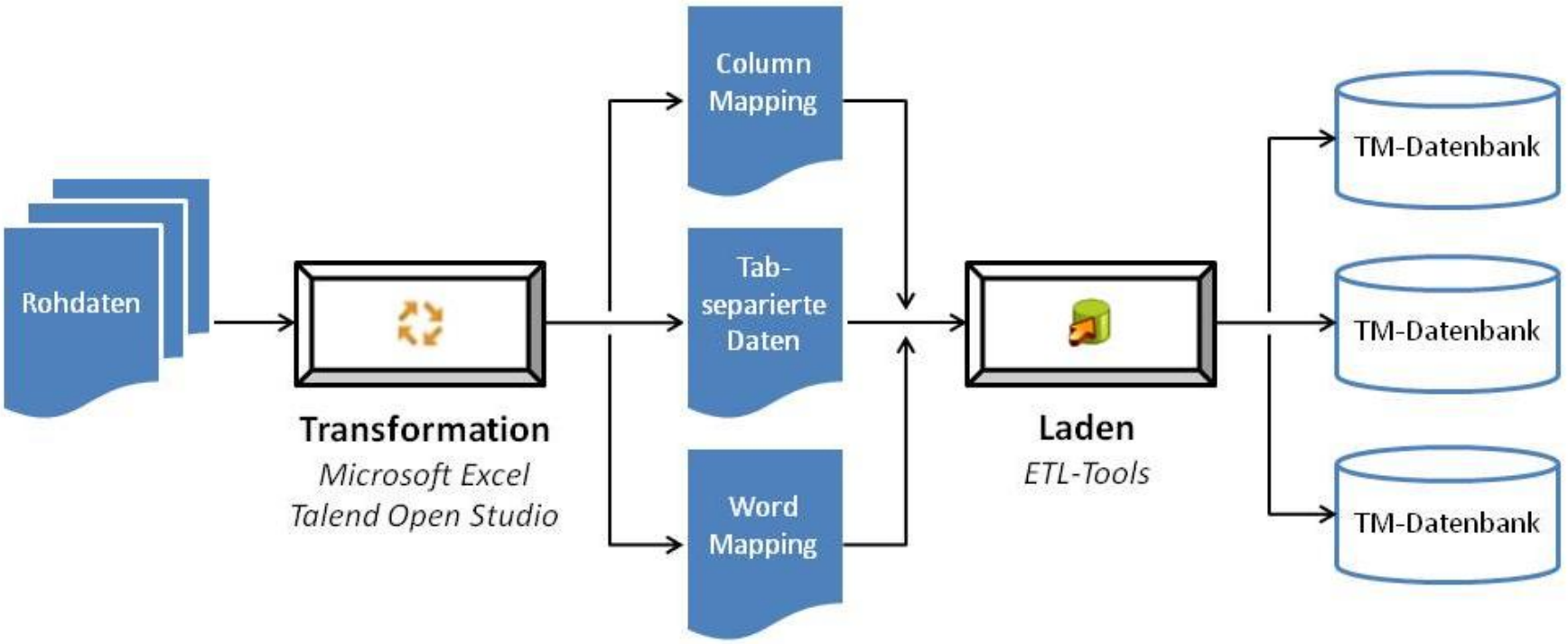
Import hochdimensionaler Daten: ETL-Tools

- Import mittels spezieller ETL-Tools:

ETL-Tool	Technische Basis	Verwendung
Integrated Curation Environment (ICE)	Java + Kettle Jobs + Stored Procedures	GUI
tMDataLoader	Groovy + Stored Procedures	Kommandozeile
transmart-data	Kettle Jobs + Stored Procedures	Kommandozeile
transmart-batch	Spring Batch + Groovy	Kommandozeile

- TranSMART 2.0 setzt auf transmart-batch

Import hochdimensionaler Daten: Übersicht



Import hochdimensionaler Daten: Übersicht

Datentyp / ETL-Tool	ICE	tMDataLoader	transmart-batch	transmart-data
aCGH / CNV		✓	✓	✓
Metabolom		✓	✓	✓
miRNA	✓	✓	✓	✓
mRNA	✓	✓	✓	✓
Protein	✓	✓	✓	
RBM	✓	✓		✓
RNAseq	✓	✓	✓	✓
SNP	✓	✓		
VCF		✓		✓

SANOFI THOMSON REUTERS thehyve

Beispiel-Import

- Import von mRNA-Daten (`GSE8581` // `GPL570`)
- ETL-Tool: transmart-batch
- Benötigte Dateien:
  - ANNOTATION_FILE
  - ANNOTATION_PARAMETER_FILE
  - RAW_DATA_FILE
  - SUJECT_SAMPLE_MAP_FILE
  - PARAMETER_FILE

Beispiel-Import: ANNOTATION_FILE

- Enthält den verwendeten Genchip
- Inhalt (`GPL570.tsv`):

GPL_ID	PROBE_ID	GENE_SYMBOL	GENE_ID	ORGANISM
GPL570	1007_s_at	DDR1	780	Homo Sapiens
GPL570	1053_at	RFC2	5982	Homo Sapiens
GPL570	117_at	HSPA6	3310	Homo Sapiens
GPL570	121_at	PAX8	7849	Homo Sapiens

- Verfügbar unter:

  http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi

Beispiel-Import: ANNOTATION_PARAMETER_FILE

- Festlegung von Meta-Informationen
- Inhalt (`annotation.params`):

```
PLATFORM=GPL570
TITLE="Affymetrix Human Genome U133A 2.0 Array"
ORGANISM="Homo Sapiens"
ANNOTATIONS_FILE=GPL570.tsv
```

Beispiel-Import: RAW_DATA_FILE

- Enthält die gemessenen Genexpressionen
- Inhalt (`GSE8581.txt`):

PROBE_ID	GSM210006	GSM210007	GSM210008	GSM210009
1007_s_at	179.15	367.436	271.773	411.581
1053_at	35.5867	60.1003	50.2729	62.2023
117_at	33.9371	31.3962	34.251	59.2324
121_at	105.912	140.325	166.999	138.719

Beispiel-Import: SUBJECT_SAMPLE_MAP_FILE

- Zuordnung der einzelnen Proben zu einem Patienten
- Inhalt (`mapping.txt`):

STUDY_ID	SUBJECT_ID	SAMPLE_ID	CATEGORY_CD
GSE8581	12345	GSM210006	Biomarker_Data+mRNA
GSE8581	12345	GSM210007	Biomarker_Data+mRNA
GSE8581	12346	GSM210008	Biomarker_Data+mRNA
GSE8581	12346	GSM210009	Biomarker_Data+mRNA

- Des Weiteren:
  - SITE_ID, PLATFORM, TISSUETYPE, ATTR1, ATTR2, SOURCE_CD

Beispiel-Import: PARAMETER_FILE

- Festlegung von Meta-Informationen
- Inhalt (`expression.params`):

```
DATA_FILE_PREFIX=GSE8581.txt
MAP_FILENAME=mapping.txt
```

- Des Weiteren:
  - DATA_TYPE, LOG_BASE, …

Beispiel-Import: Ladevorgang I

- Benötigte Ordnerstruktur:

```
ANNOTATION_NAME (GPL570)
-- ANNOTATION_PARAMETER_FILE (annotation.params)
-- ANNOTATION_FILE (GPL570.tsv)

STUDY_NAME (GSE8581)
-- PARAMETER_FILE (expression.params)
-- expression
---- RAW_DATA_FILE (GSE8581.txt)
---- SUBJECT_SAMPLE_MAP_FILE (mapping.txt)
```

Beispiel-Import: Ladevorgang II

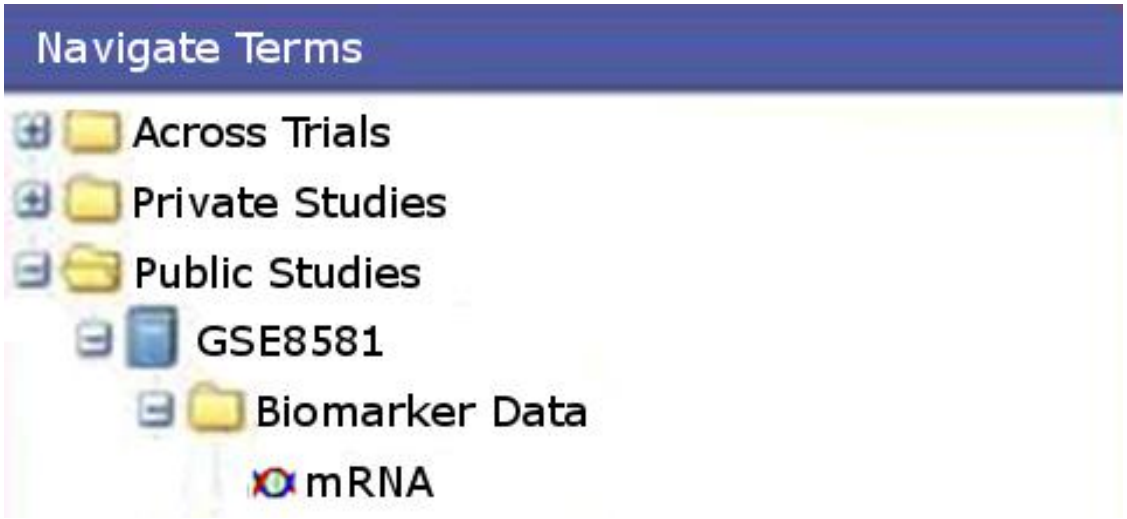
- Durchführung mittels:

```
~/transmart/transmart-batch/transmart-batch-capsule.jar
-p /path/to/ANNOTATION_NAME/annotation.params

~/transmart/transmart-batch/transmart-batch-capsule.jar
-p /path/to/STUDY_NAME/expression.params
```

Beispiel-Import: Ladevorgang III

- Ergebnis:



Analysen: Übersicht

- 17 eingebaute Analysemöglichkeiten

aCGH Survival Analysis	Heatmap	Marker Selection
Box Plot with ANOVA	Hierarchical Clustering	PCA
Correlation Analysis	K-Means Clustering	Scatter Plot with Linear Regression
Frequency Plot for aCGH	Line Graph	Survival Analysis
Group Test for aCGH	Logistic Regression	Table with Fisher Test
Group Test for RNASeq		Waterfall

Analysen: Übersicht

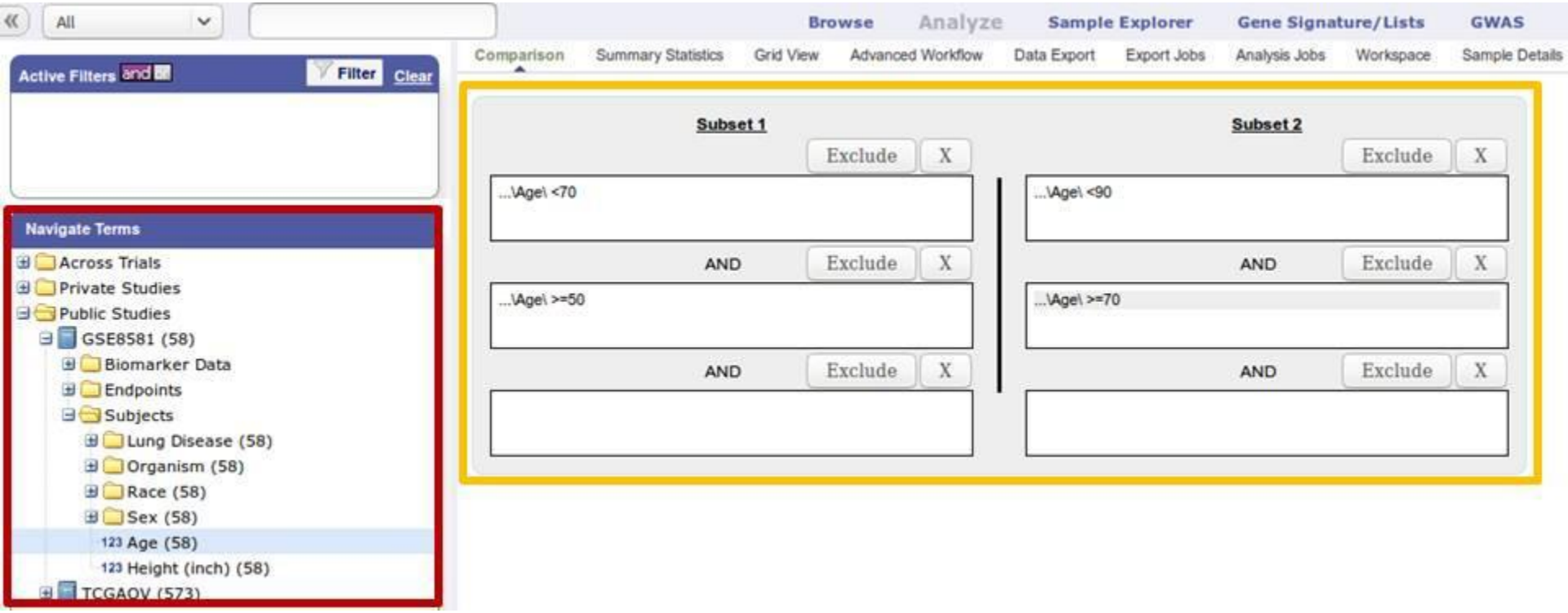
- 17 eingebaute Analysemöglichkeiten

aCGH Survival Analysis	Heatmap	Marker Selection
Box Plot with ANOVA	Hierarchical Clustering	PCA
~~Correlation Analysis~~	K-Means Clustering	Scatter Plot with Linear Regression
Frequency Plot for aCGH	Line Graph	Survival Analysis
Group Test for aCGH	Logistic Regression	Table with Fisher Test
Group Test for RNASeq		~~Waterfall~~

- 15 Analysemöglichkeiten mit HDD-Support

Analysen: Durchführung I

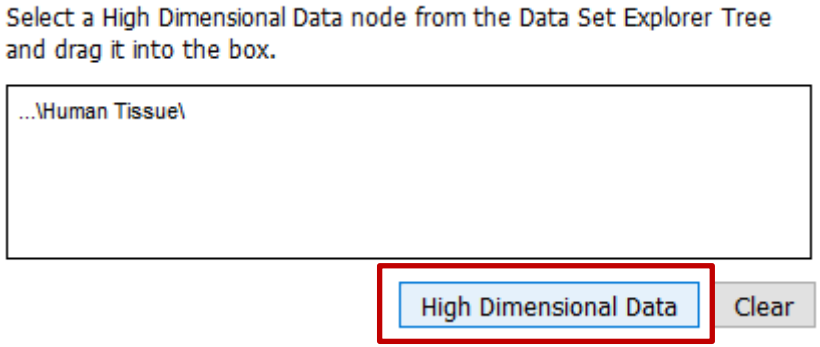
- Kohortenbildung

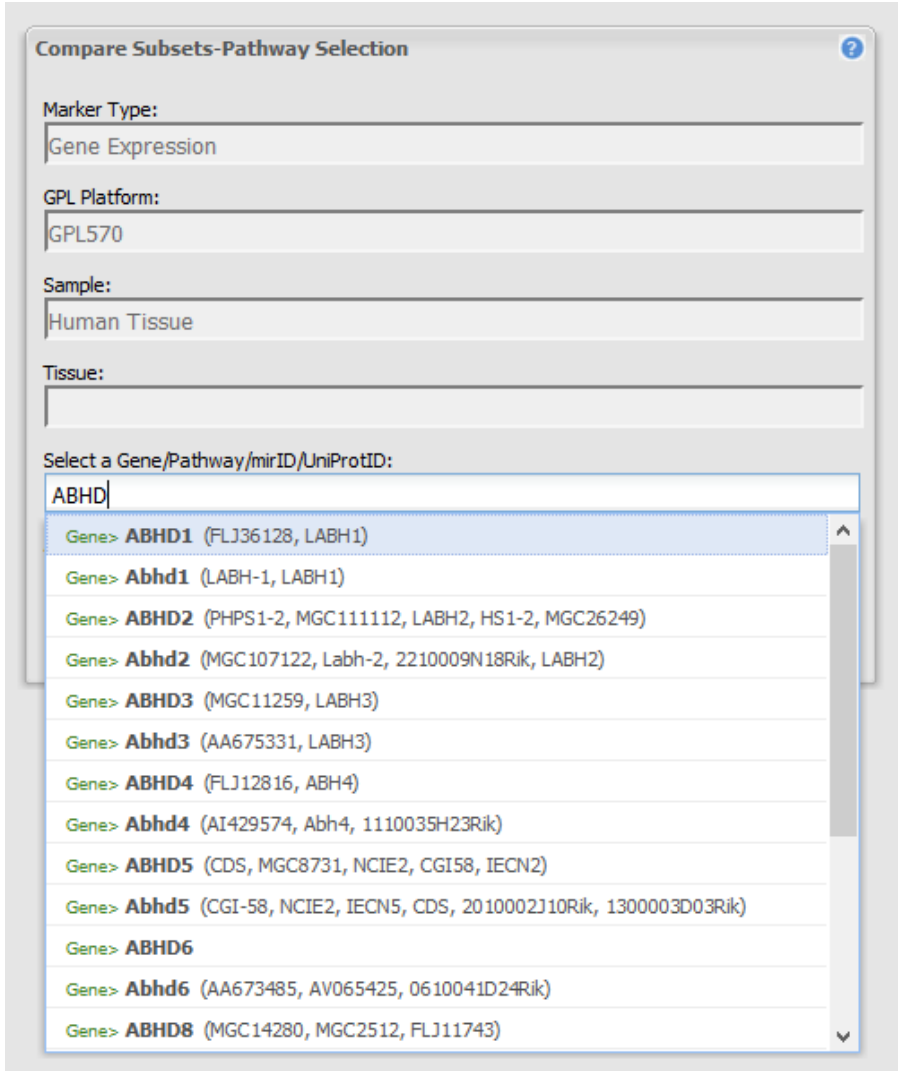


- Keine HDD möglich

Analysen: Durchführung II

- Filterung
  - GeneID
  - miRNA-ID
  - UniProtID





Vielen Dank für Ihre Aufmerksamkeit!

- Hochdimensionalität: n<<p
- Import-Tools:
  - Integrated Curation Environment (ICE)
  - tMDataLoader
  - transmart-batch
  - transmart-data
- 15 HDD-Analysen
  - Kohortenbildung ohne HDD
  - Filterung der Daten mittels GeneID, miRNA-ID oder UniProtID