

Data Integration and Knowledge Management within Oncotyrol

A. Dander, R. Gallasch, S. Pabinger, H. Fiegl, Z. Trajanoski

Oncotyrol, Center for Personalized Cancer Medicine, Innsbruck, Austria
Biocenter, Division for Bioinformatics, Innsbruck Medical University, Innsbruck, Austria

Agenda

- ▶ Oncotyrol
 - ▶ Project: Knowledge Management
 - ▶ Showcase: Breast Cancer - Metastasis Risk
 - ▶ Outlook
-



- ▶ 29 research projects
 - ▶ 5 partner universities
 - ▶ 34 industrial partners
 - ▶ Phase I: July 2008 – June 2012
 - ▶ Total project volume € 24 million
 - ▶ Main research topics:
 - ▶ Breast cancer
 - ▶ Leukemia
 - ▶ Prostate cancer
 - ▶ Phase II: July 2012 – June 2015
-



- ▶ Projects with bioinformatics requirements
 - ▶ Use high-throughput techniques
 - ▶ Mass spectrometry
 - ▶ Microarray
 - ▶ Next generation sequencing
 - ▶ Overview on public data
 - ▶ Combination of clinical, experimental and public data

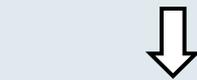
- ➔ One project for „Knowledge Management“
 - ▶ Data integration
 - ▶ Data analyses

Knowledge Management

Project: Knowledge Management



Public data



Data integration and knowledge management



in silico
Validation



**Breast Cancer
Metastasis Risk**

Gynecology
Medical University
Innsbruck

**New Markers for
Prostate Cancer**

Urology
Medical University
Innsbruck

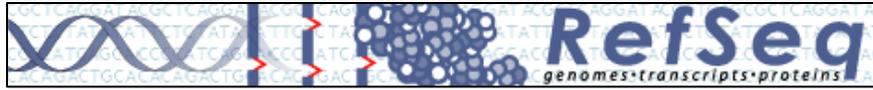
**Glucocorticoid-
responsive
Transcriptome**

Tyrolean Cancer
Research Institute



Public Databases

Taxonomy



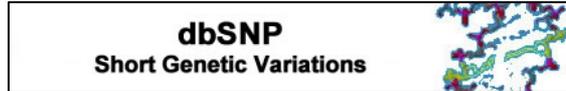
IPI - International Protein Index



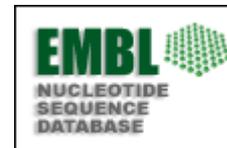
UniGene



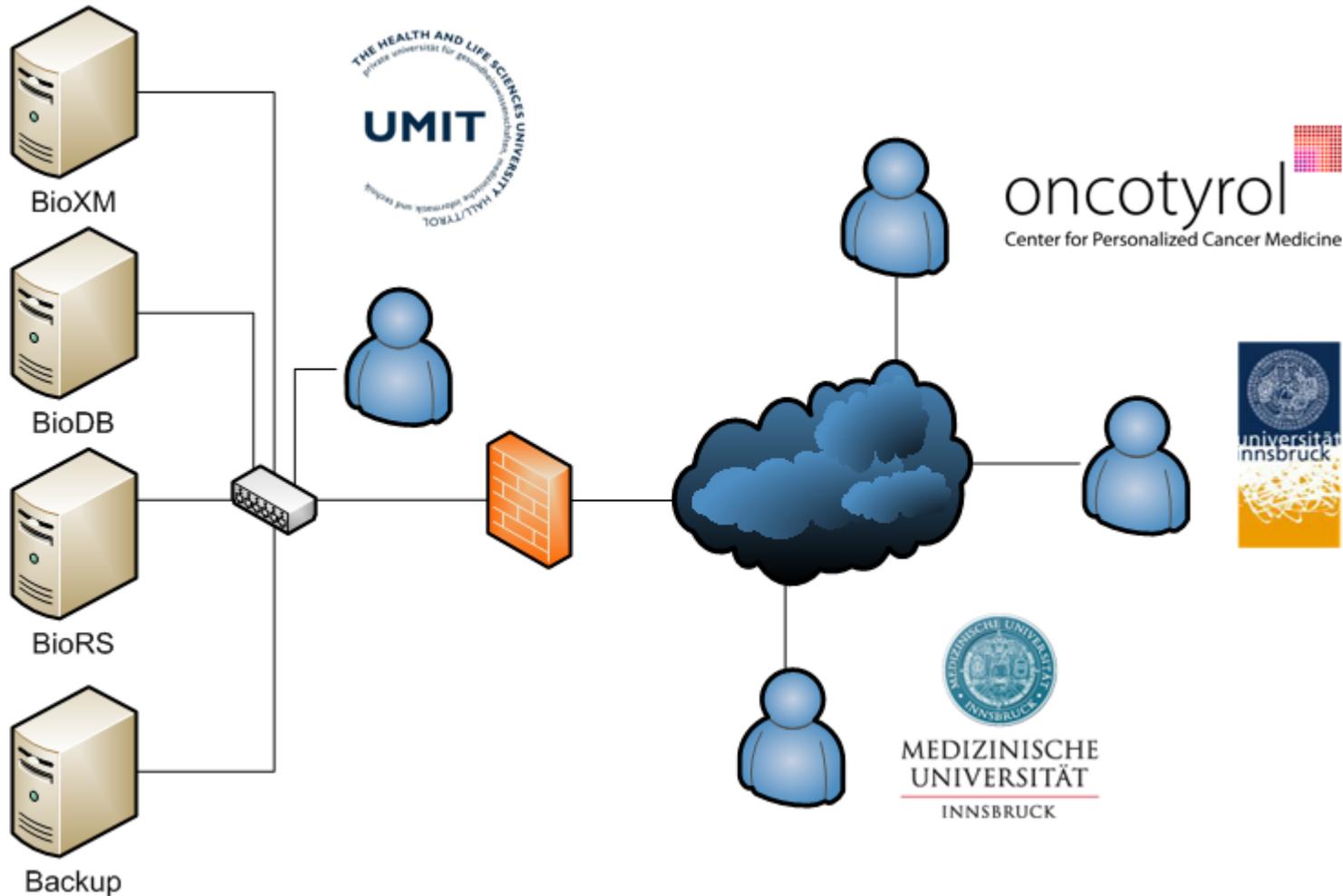
InterPro protein sequence analysis & classification



Patent Nucleotide Sequences



Network structure



Showcase: Breast Cancer – Metastasis Risk

- ▶ Breast cancer
 - ▶ All patients have to undergo extensive therapy
 - ▶ ~20% of all adjuvant treated patients develop metastasis
- ▶ Estimated therapy outcome
 - ▶ Inefficient for 20%
 - ▶ Unnecessary for 70%
 - ▶ Helpful for 10%
- ▶ Identification of a risk pattern
 - ▶ Epigenetic, genetic, metabolomic, and transcriptomic biomarkers
 - ▶ Determined in white blood cells and serum

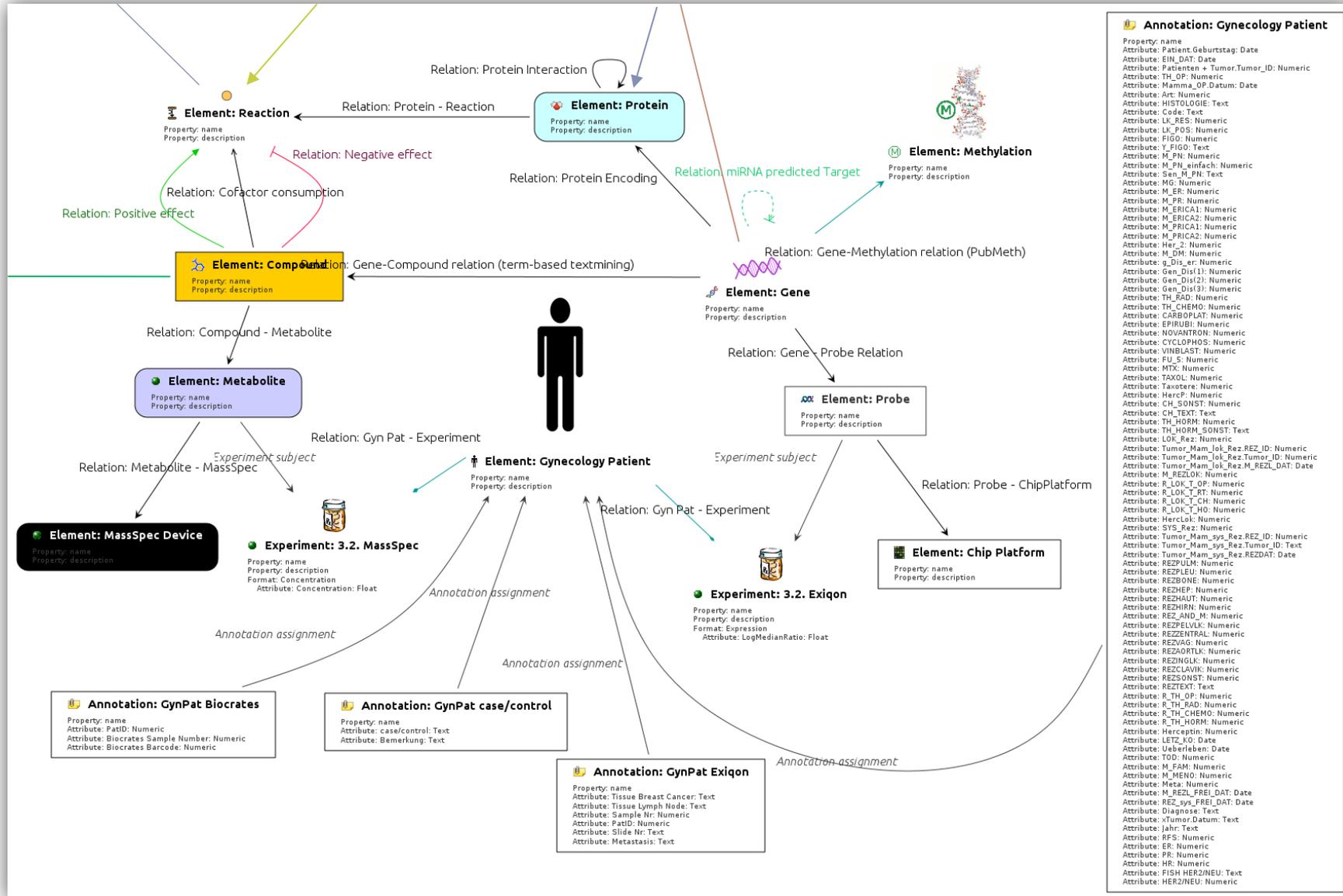
Showcase: Breast Cancer – Metastasis Risk

- ▶ Prospective Study
 - ▶ Start in 2009
 - ▶ Three different locations
 - ▶ Innsbruck, Austria
 - ▶ Salzburg, Austria
 - ▶ Brixen, Italy
 - ▶ Currently more than 1.000 samples
 - ▶ Metastasis < 5 years (➔ Phase II)

Breast Cancer – Retrospective Pilot Study

- ▶ Retrospective pilot study
 - ▶ Clinical data (n=1117)
 - ▶ 107 cases
 - ▶ 1010 controls
 - ▶ Blood serum samples
 - ▶ Metabolic data [mass spectrometry] (n=55)
 - 20 cases
 - 35 controls
 - ▶ Formalin fixed paraffin embedded tissue (FFPE)
 - ▶ miRNA data [microarray] (n=46)
 - 14 cases
 - 32 controls

Breast Cancer Pilot Study: Data Integration

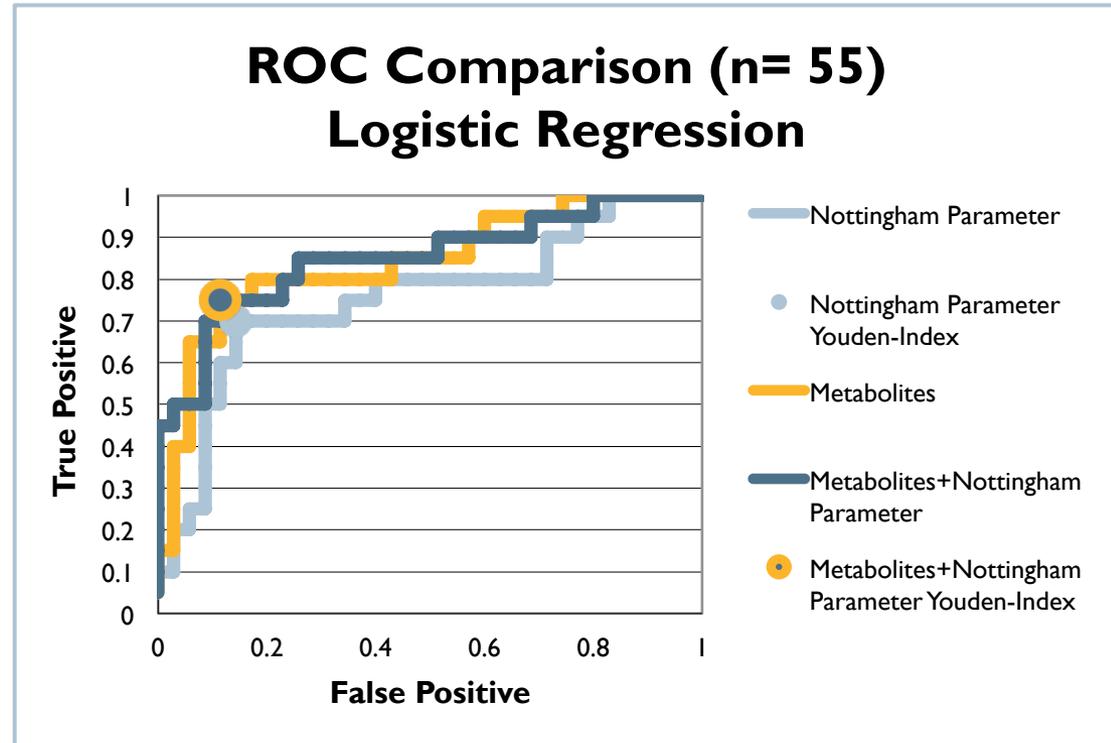


Breast Cancer Pilot Study: Data Analyses

- ▶ Formalin fixed paraffin embedded tissue (miRNA)
 - ▶ 14 cases
 - ▶ 32 controls
 - ▶ Identification of 12 differentially expressed miRNAs
 - ▶ T-test and q-value < 0.05
 - ▶ 1 confirmed candidate after RT-qPCR

Breast Cancer Pilot Study: Data Analyses

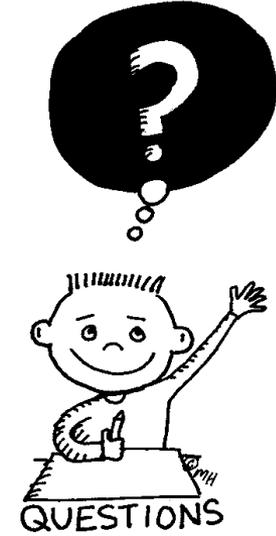
- ▶ Metabolic data from blood serum
 - ▶ 20 cases / 35 controls
 - ▶ 4 metabolites
- ▶ Clinical parameters
 - ▶ Nottingham prognostic index (NPI)
 - ▶ Size
 - ▶ Lymph nodes
 - ▶ Grade



Data	Sens	Spec	Acc	ROC Area
NPI	70%	86%	76%	0.76
Metabolic	75%	89%	80%	0.84
Combination	75%	89%	83%	0.85

Outlook

- ▶ Continued support of projects within Oncotyrol
 - ▶ Data integration
 - ▶ Public databases
 - ▶ Clinical and experimental datasets
 - ▶ Data analyses
 - ▶ Metabolomics
 - ▶ Gene expression
 - ▶ Epigenetics
 - ▶ Next generation sequencing data
 - ▶ Breast cancer metastasis risk
 - ▶ Integration of clinical and experimental data
 - ▶ Integrative analysis of overall dataset
-



Data Integration and Knowledge Management within Oncotyrol

A. Dander, R. Gallasch, S. Pabinger, H. Fiegl, Z. Trajanoski

Oncotyrol, Center for Personalized Cancer Medicine, Innsbruck, Austria
Biocenter, Division for Bioinformatics, Innsbruck Medical University, Innsbruck, Austria

Data integration challenges

- ▶ Different DBs → Different IDs
 - ▶ **Genesymbol BRCA1**: EntrezGene 672; HGNC 1100; OMIM 113705, UNIGENE Hs. 194143
- ▶ History → Synonymous names
 - ▶ >190 aliases for formaldehyde
- ▶ Different resolution in molecular DBs and experiments
 - ▶ Isomorphic molecules
- ▶ Not all information available for download
- ▶ Different data types and data access
 - ▶ Relational DBs
 - ▶ File formats
 - ▶ (CEL, XLS, CSV, ARFF, XML, GML, fastq, individual formats ...)