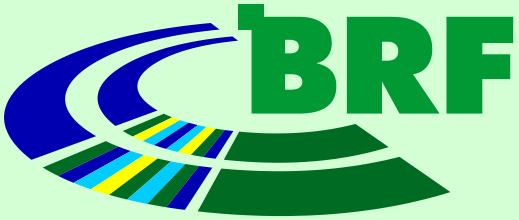
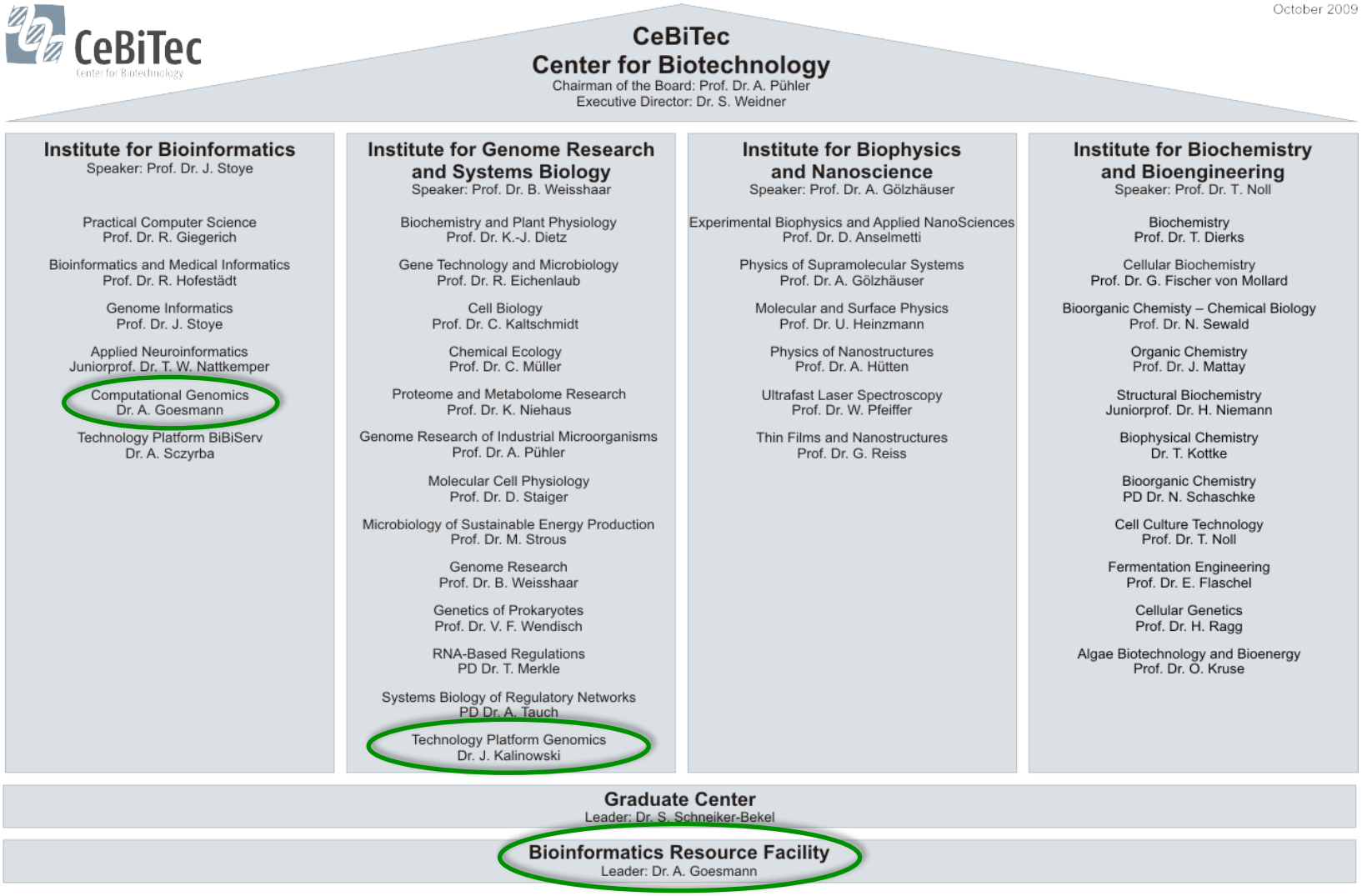


Research Data Management at the Center for Biotechnology

Alexander Goesmann
Computational Genomics
Bioinformatics Resource Facility
Center for Biotechnology
Bielefeld University
29.10.2010

The Bielefeld Center for Biotechnology



http://www.cebitec.uni-bielefeld.de

Computational Genomics at CeBiTec

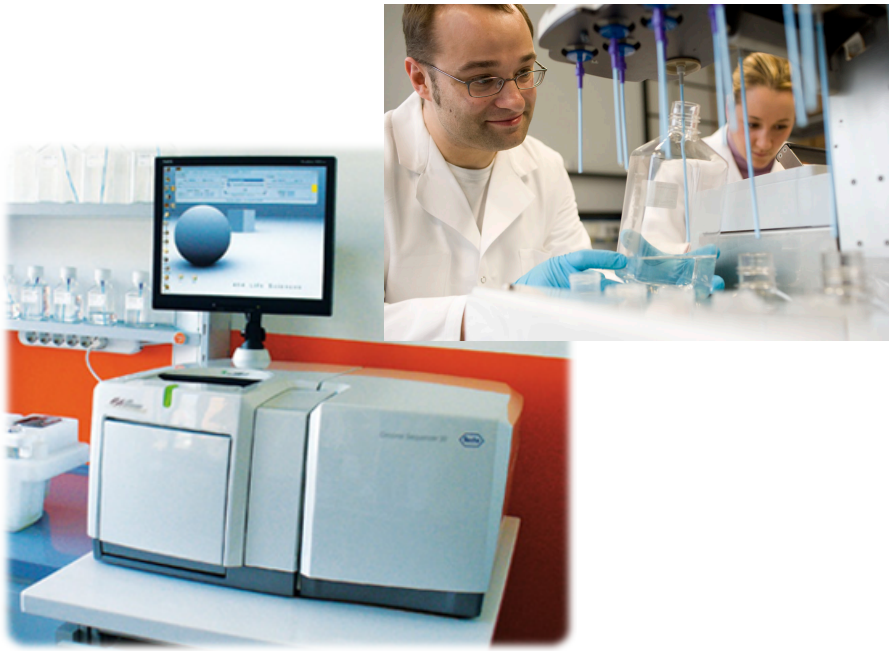
- The Bioinformatics Resource Facility (BRF) provides
  - Large scale hardware and software infrastructure for (microbial) genome research
  - Sequence analysis and genome annotation
  - Storage and analysis of gene expression data
  - Training courses

- The Junior Research Group for Computational Genomics develops new tools and software applications for
  - Multi-Omics data storage and analysis
  - Visualization
  - Data Integration

- IIT Biotech Informatics GmbH

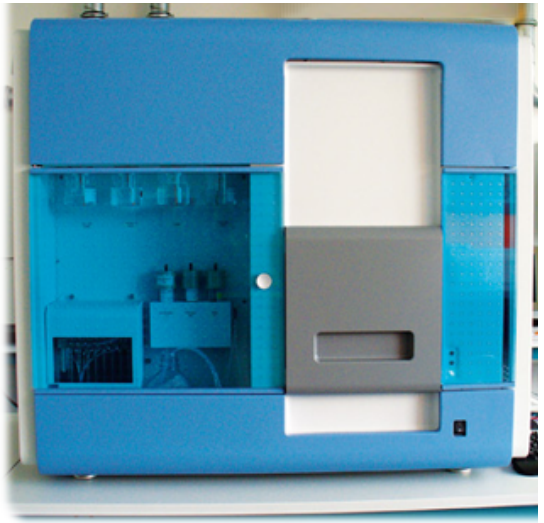
CeBiTec

High-throughput Sequencing



454 GS FLX™ Sequencer (Roche Diagnostics)

Currently 0.5 Gbp/10 hour run (400 bp reads)
~ 2 - 8 bacterial genomes or 1 yeast/fungus at 25fold coverage



Solexa™ Genetic Analyzer (Illumina)

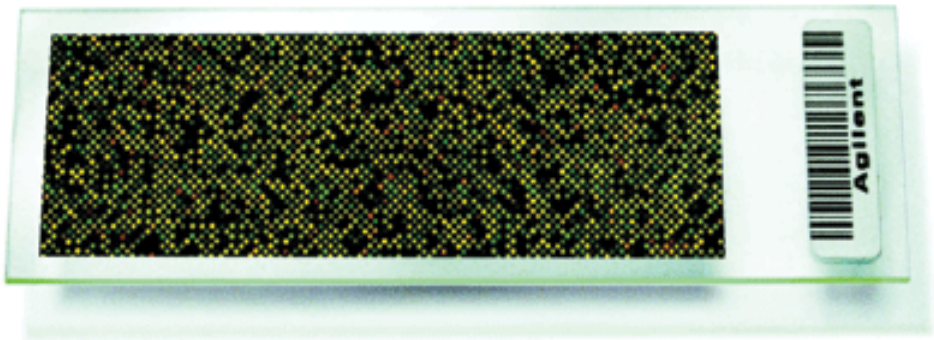
Currently 55 Gbp/10 days run (100 bp reads)
~ a mammalian genome at 15fold coverage

CeBiTec

High-throughput Transcriptomics

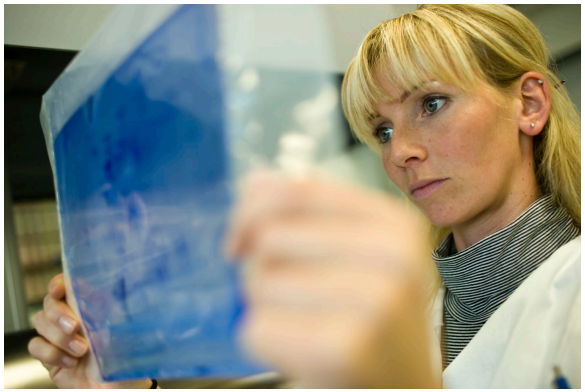
Microarray Fabrication and Analysis at the CeBiTec

- New Microarray Scanner (Agilent)
- 2μm resolution, up to 2.000.000 features/slide
- applicable for standard or microfabricated arrays (Agilent, NimbleGen)



CeBiTec

High-throughput Proteomics



Protein Gel Electrophoresis and Mass Spectrometry at the CeBiTec

New mass spectrometry equipment:

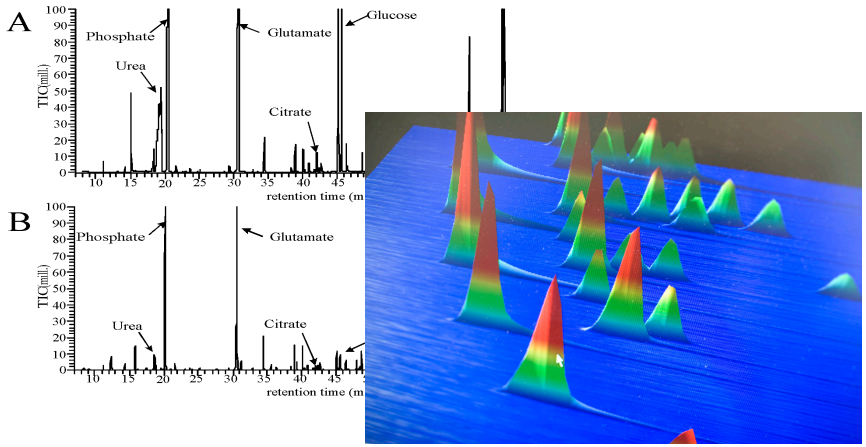
- Bruker UltrafleXtreme – MALDI-TOF/TOF
  - latest technology for LC-MALDI applications
- Bruker microTOF-QII
  - general purpose instrument for small molecules (proteomics & metabolomics)

CeBiTec

High-throughput Metabolomics

Metabolite Harvesting and Analysis by Mass Spectrometry at the CeBiTec

- maximally controlled parallel cultivation
- proprietory cell harvesting methods
- 2D gas-chromatography/mass spectroscopy
- LC-coupled mass spectrometry for hydrophilic metabolites
- HPLC separation for amino acid and keto acid analysis



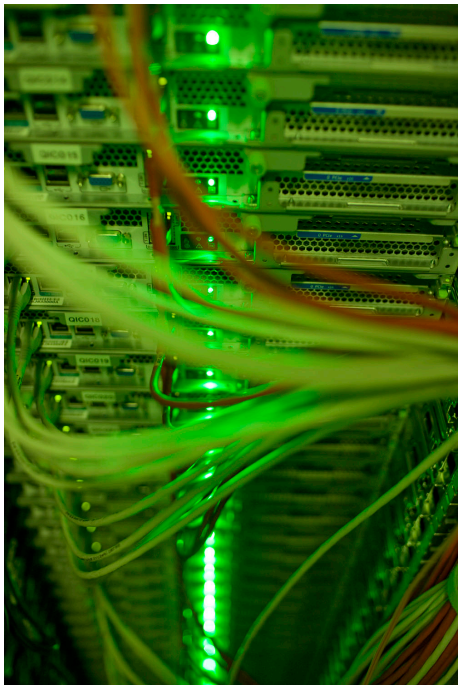
Main Tasks of the Technology Platform Bioinformatics

- Provide large scale hardware and software infrastructure for (microbial) genome research
- Sequence analysis, genome annotation, and metagenomics
- Storage and analysis of gene expression data
- Develop new tools and software applications for "multi-omics" data storage and analysis in **close collaboration** with in-house users and project partners
- Data Visualization
- Data Integration
- General Support
- Training courses

CeBiTec

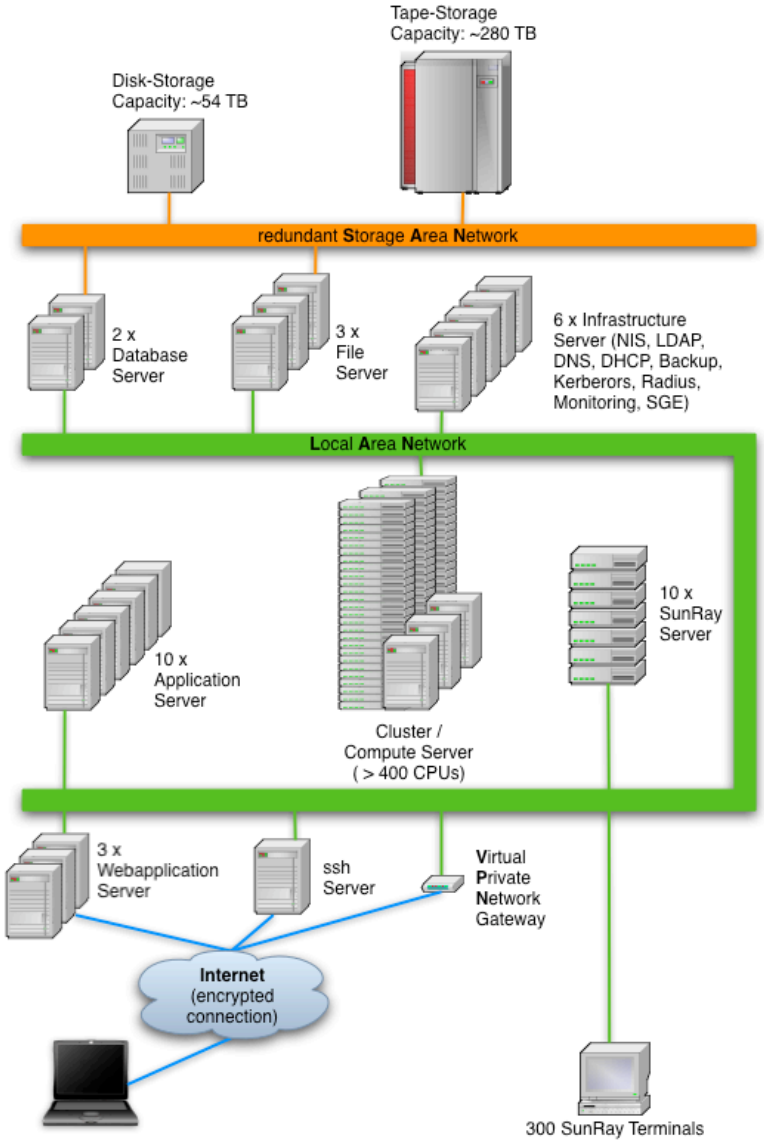
The BRF User Environment

- 4 Central SunRay Terminal-Servers
- Disk-less Clients (everywhere!)
- Standard Desktop (KDE)
- Local access to more than 300 bioinformatics tools & sequence databases
- Dedicated application servers
- Compute-Cluster



CeBiTec

The BRF Hardware Infrastructure



- ~ 100 TB disc storage
- ~ 280 TB tape storage
- ~ 800 CPUs
  (> 2200 CPU cores)
- ~ 17 TFLOPS (Rpeak)
  compute capacity
- investment of 850.000 EUR
  in 2009 / 2010

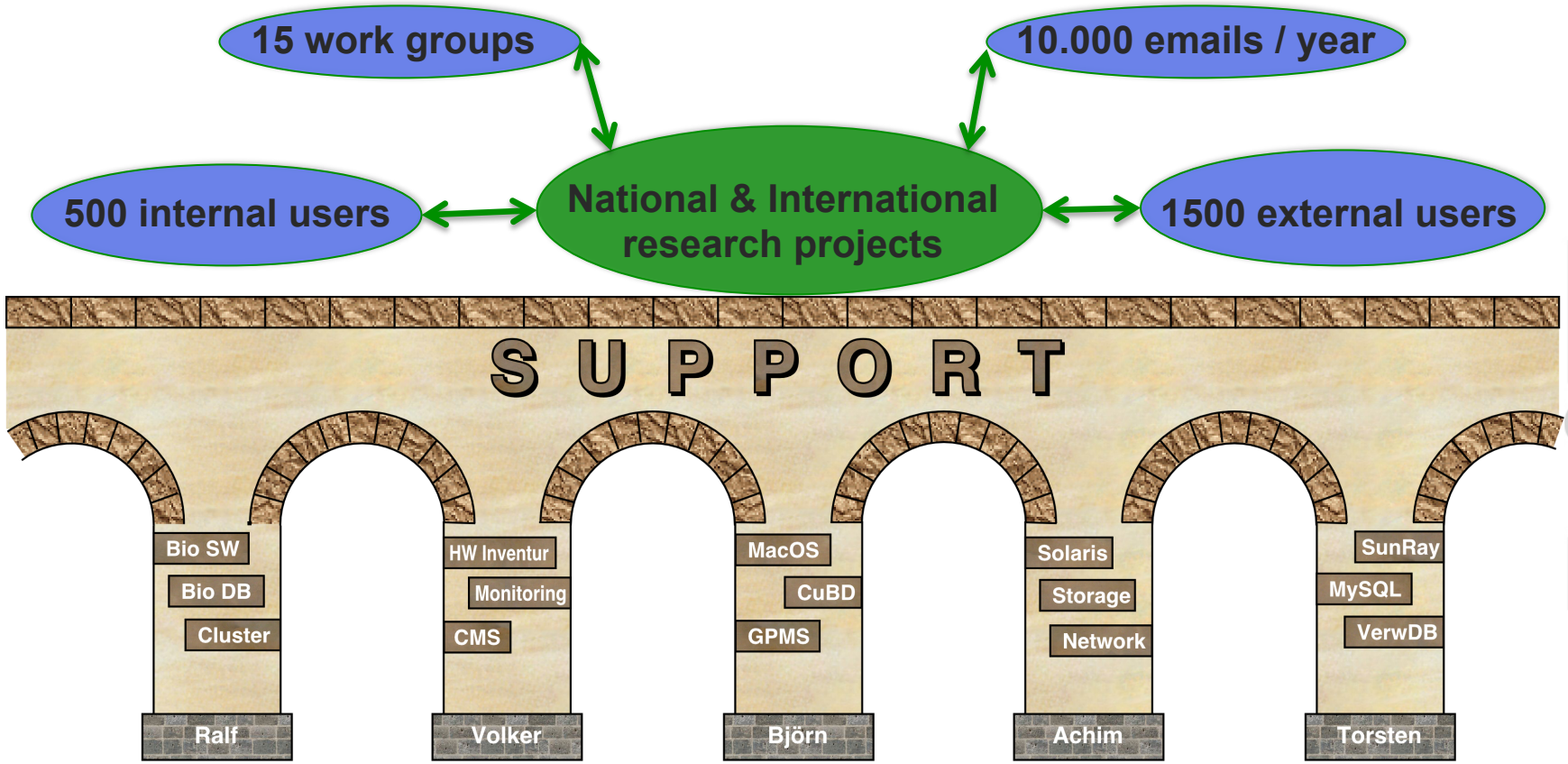
CeBiTec

Compute and Storage Resources

- 128 x SunFire V20z
- 28 x SunFire 2200 M2
- 76 x SunFire X2250
- 34 x PRIMERGY RX200S5
- 8 x SunFire V880/V890
- 7 x SunFire X4600
- 4 x Sun T2000
- 2 x SunFire X4440
- 9 x PRIMERGY RX600S5
- Sun StorEdge 351x and J4500 JBOD
- 2 x Sun StorEdge L700
- 1 x Quantum ATL P7000
- 1 x DeCypher SeqCruncher

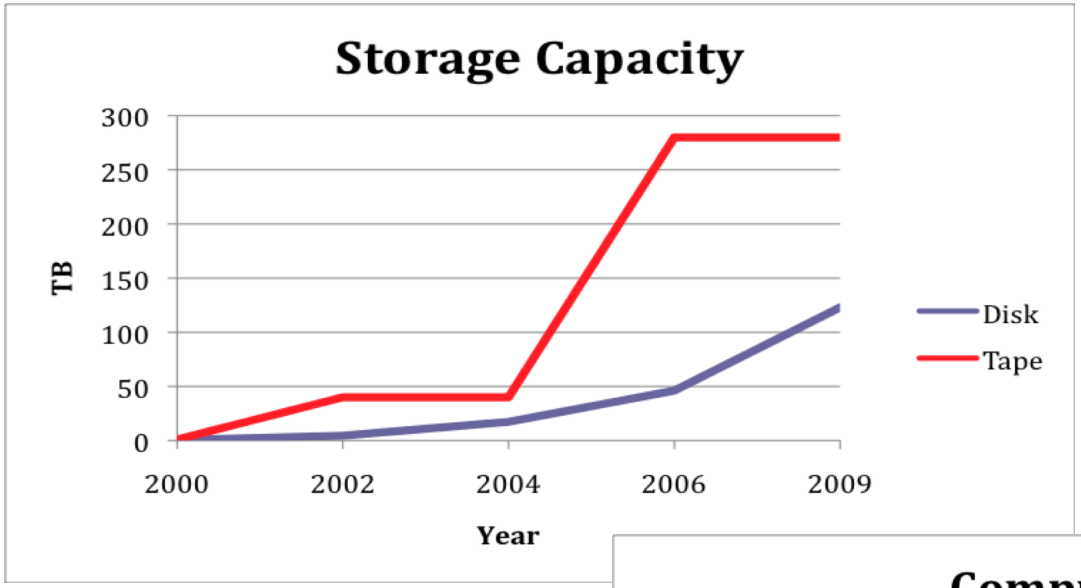
CeBiTec

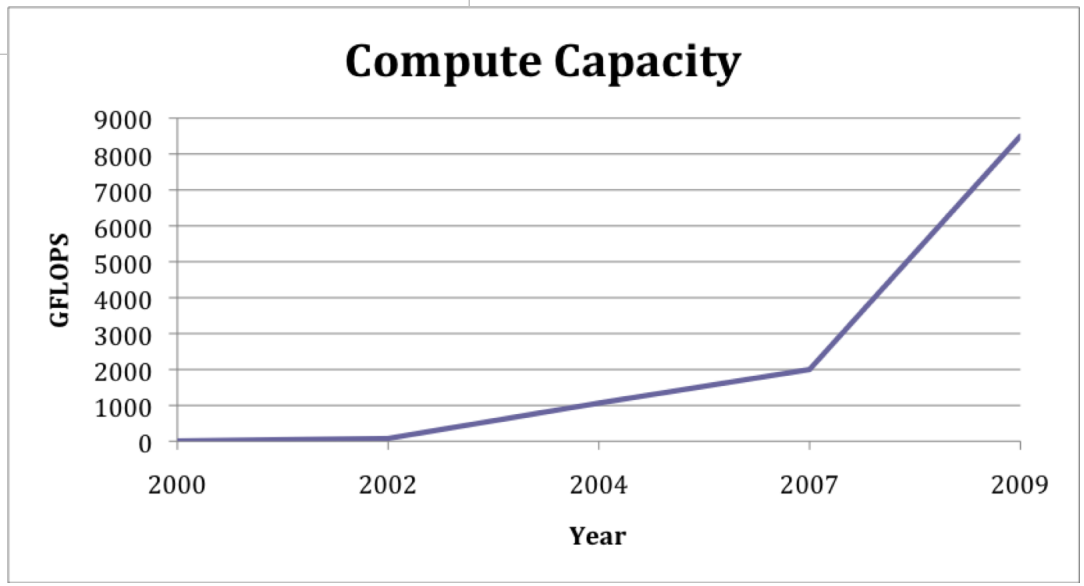
BRF System Administration



CeBiTec

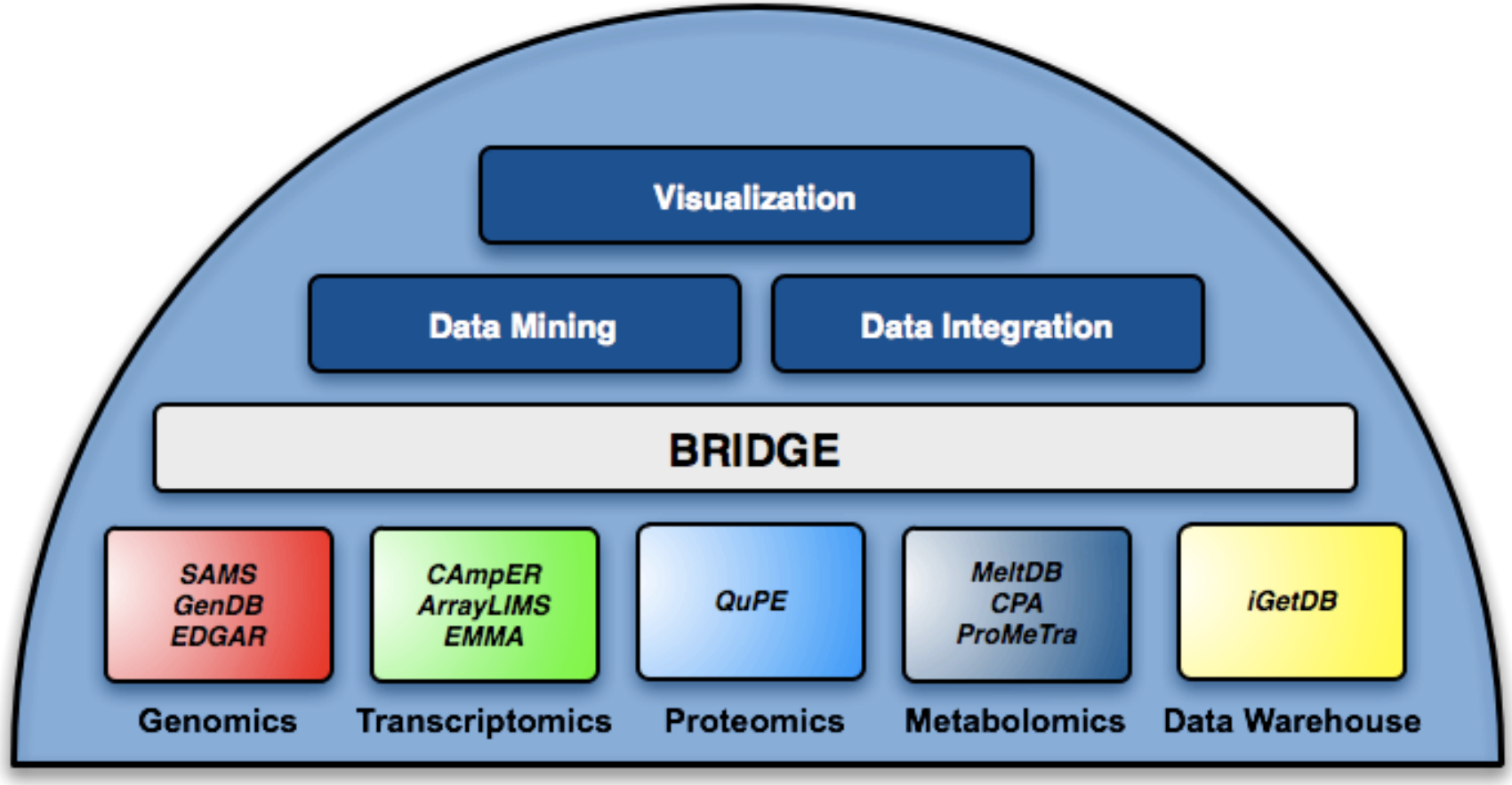
Development of Resources





CeBiTec

The BRF Software Suite



Bekel et al., 2009
Meyer et al., 2003
Blom et al., 2009

Dondrup et al., 2009

Albaum et al., 2009

Neuweger et al., 2008
Oehm et al., 2008
Neuweger et al., 2009

Becker et al., 2009
Henckel et al., 2009

30 publications from 2008 and 2009

CeBiTec

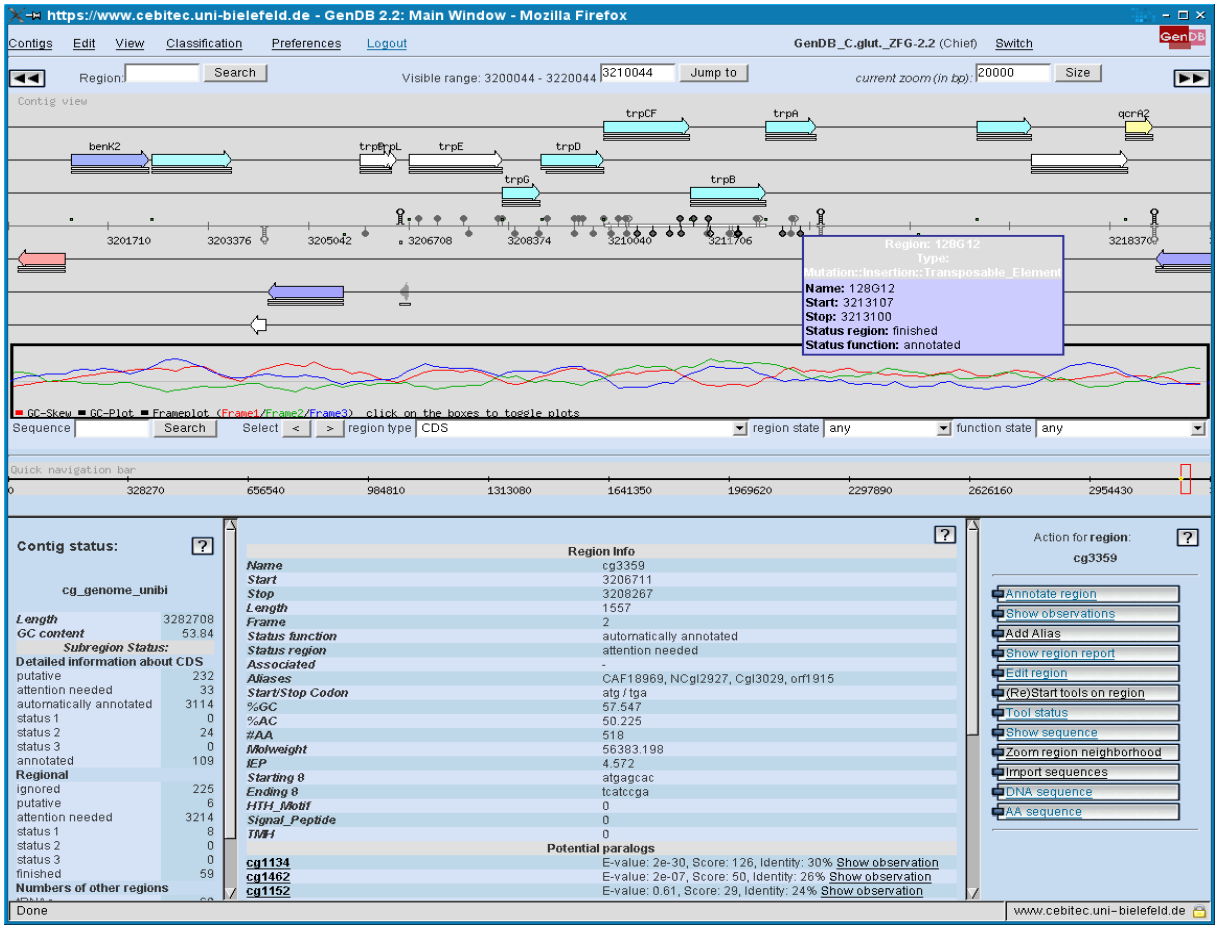
Genome analysis with SAMS, GenDB, Carmen and EDGAR

CeBiTec

Ultrafast Genome Sequencing and Annotation

- Analysis of short contigs (< 1000 bp) & single reads using SAMS
- Long contigs are linked and imported into GenDB
- **Complete automatic annotation for 3 MB in 4 h**
- High quality gene prediction: Reganor, Gismo
- Consistent function annotation: Metanor
- **Annotation based on reference genome**
- Distributed manual annotation via web interface
- Functional Classification: KEGG, COG, GO
- API can be used to easily implement new analysis scripts

GenDB – Distributed genome annotation

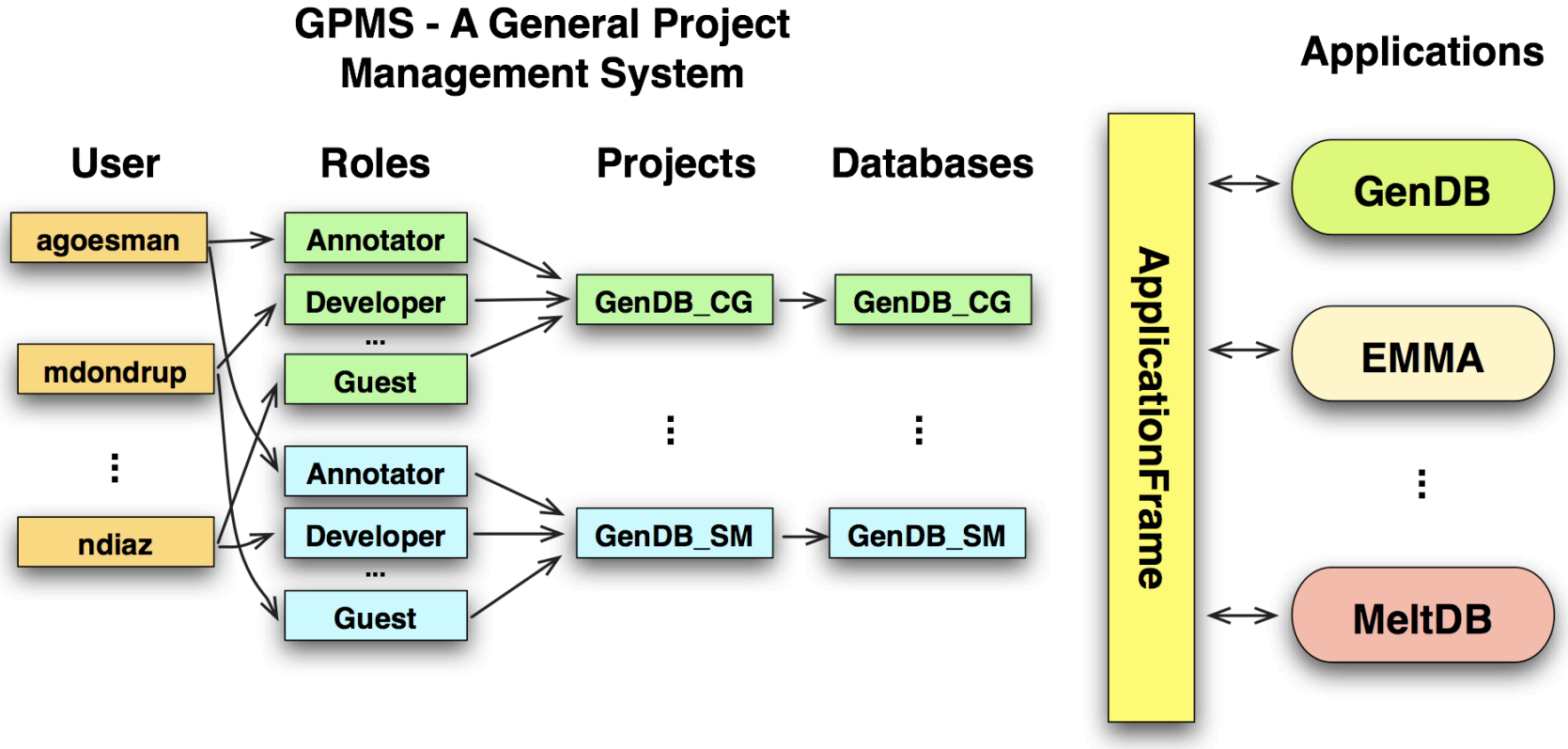


- Automatic & manual annotation
- CDS, tRNA, rRNA, IS elements, oligos, mutations, operons, terminators, …
- KEGG, COG, GO
- genome maintenance and re-annotation

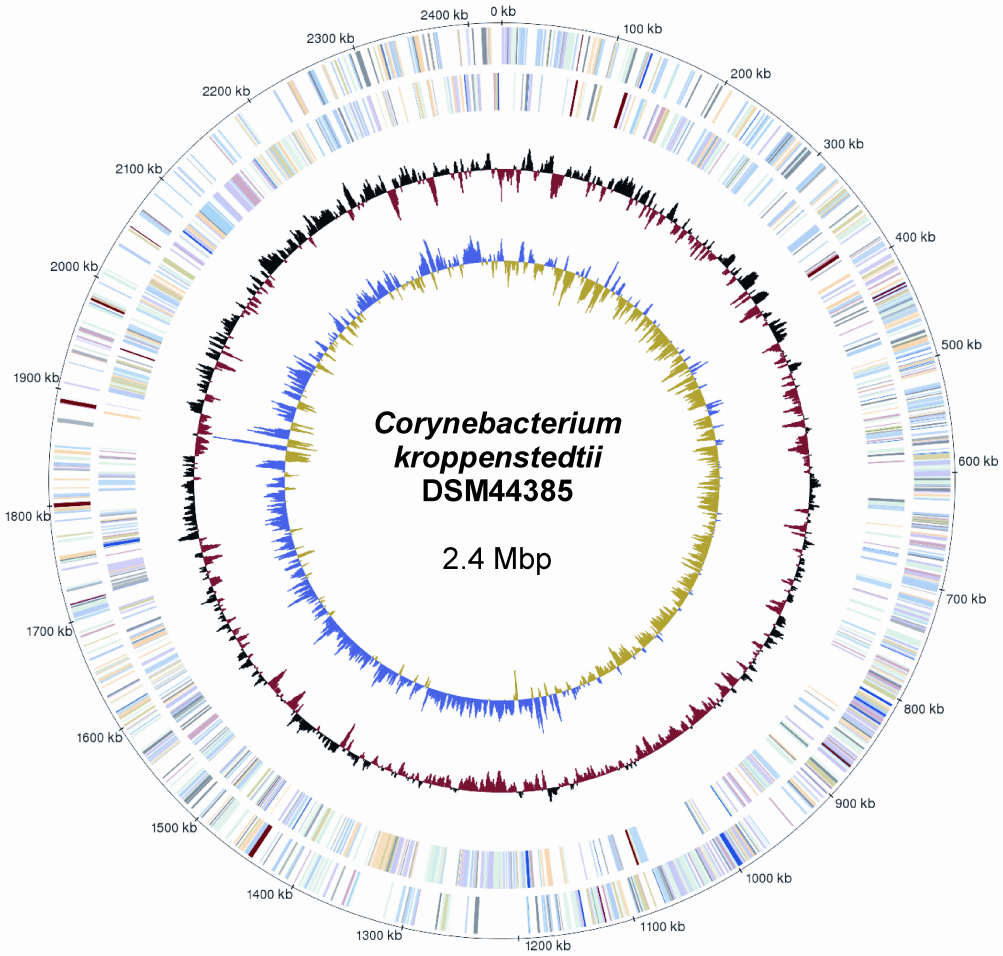
Manual genome annotation by distributed teams via web interface.

F. Meyer, A. Goesmann, A. C. McHardy, D. Bartels, T. Bekel, J. Clausen, J. Kalinowski, B. Linke, O. Rupp, R. Giegerich, A. Pühler (2003)
GenDB--an open source genome annotation system for prokaryote genomes.
Nucleic Acids Res 31(8): 2187-95.

GenDB access control via GPMS



Ultrafast genome sequencing & annotation



Overview of the *C. kroppenstedtii* DSM44385 pyrosequencing project

FLX sequencing runs	1
Shotgun reads	560,248
Detected bases	110,018,974
Mean read length	196 bp
Assembled contigs	**6**
(including 1x 16S-23S-5S rDNA)	
Size of assembled contigs	
850,812 478,533	
546,376 400,026	
152,811 5,784 (rrn consensus)	
Assembled bases	2,434,342
Mean G+C content	57.5 %
Predicted coding sequences	2119
Coding density	88.1 %
Average gene length	1016 bp
Average intergenic region	163 bp
Ribosomal RNAs	4x (16S-23S-5S)
Transfer RNAs	46

CeBiTec

One run – Done!

- training
  - end of September, 2007
- paper submitted
  - December, 2007
- paper accepted
  - March, 2008

G Model
BIOTEC 4948 1–9

ARTICLE IN PRESS

Journal of Biotechnology xxx (2008) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biotechnology

journal homepage: www.elsevier.com/locate/jbiotec

ELSEVIER

Ultrafast pyrosequencing of *Corynebacterium kroppenstedtii* DSM44385 revealed insights into the physiology of a lipophilic corynebacterium that lacks mycolic acids

Andreas Tauch[a,*], Jessica Schneider[a,b], Rafael Szczepanowski[a], Alexandra Tilker[c], Prisca Viehoever[d], Karl-Heinz Gartemann[e], Walter Arnold[c], Jochen Blom[b], Karina Brinkrolf[a,f], Iris Brune[a], Susanne Götker[a], Bernd Weisshaar[d], Alexander Goesmann[b], Marcus Dröge[g], Alfred Pühler[b]

one training → one run → one genome → one publication

In silico reconstruction of metabolic pathways

Automatic and manual annotation using GenDB after sequencing and assembly

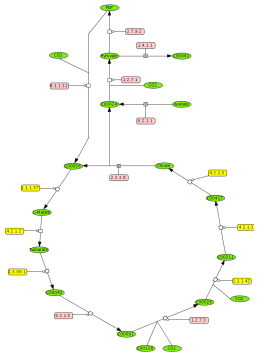
Curated model mapping

De novo model generation

C
A
R
M
E
N

New genome information will be mapped on a curated template model based on bidirectional best BLAST hits

De novo reconstruction of metabolic networks will be done based on EC numbers from GenDB; reaction and position data taken from KEGG



SBML file

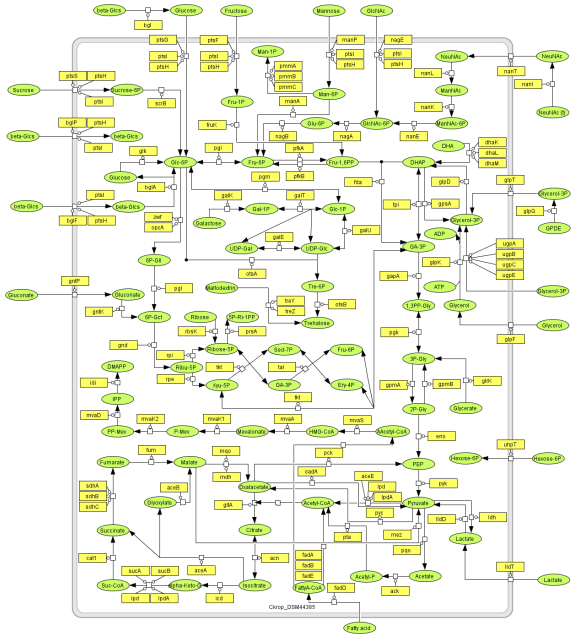
Organism-specific model visualised by the CellDesigner software

CeBiTec

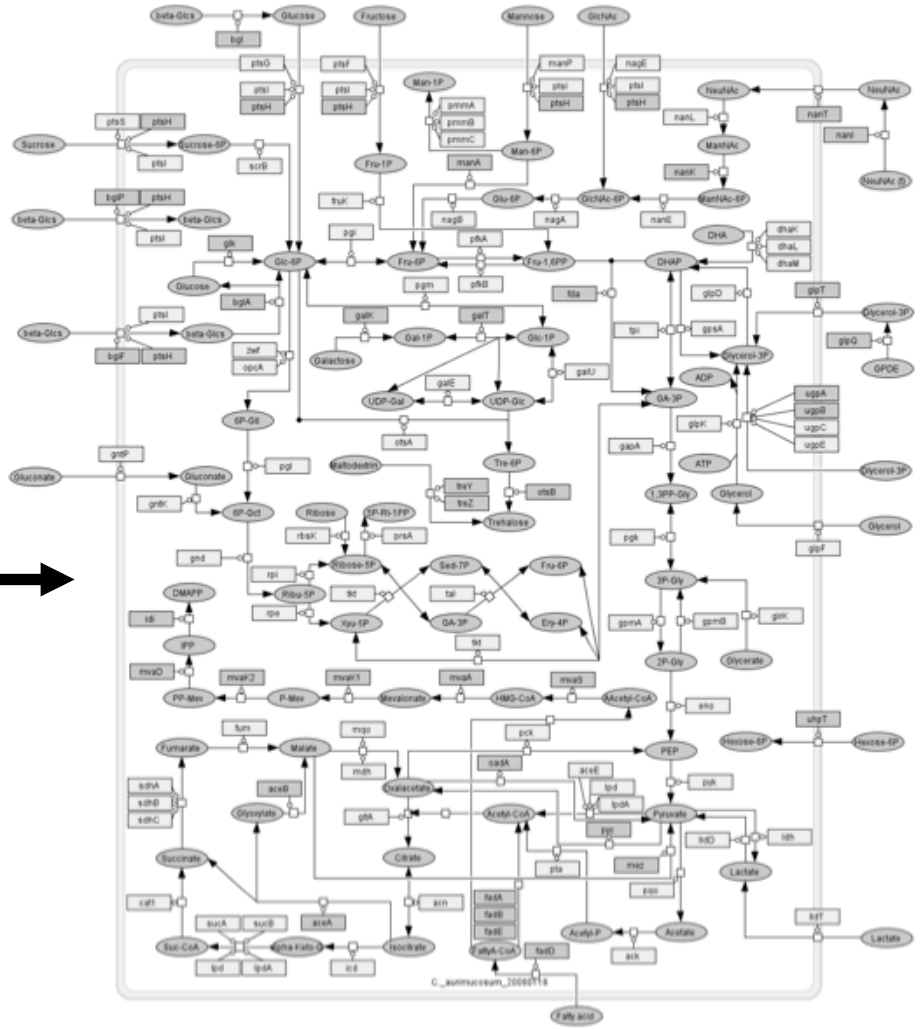
In silico reconstruction by curated model mapping

C. aurimucosum
genome sequence

+



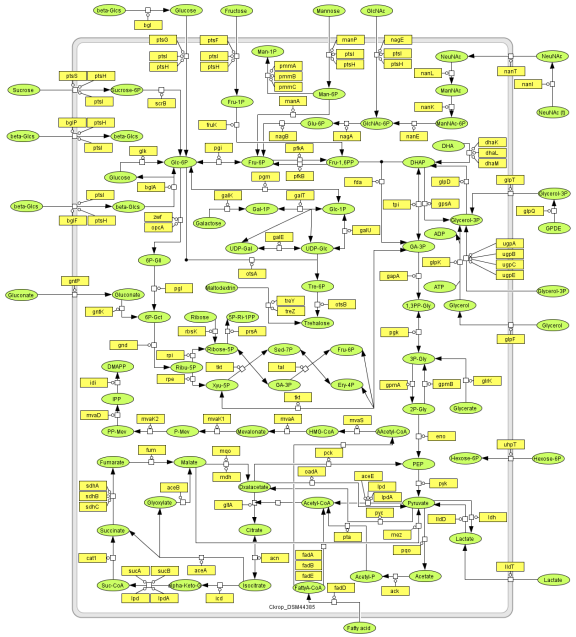
Curated model of *C. kroppenstedtii*



In silico reconstruction by curated model mapping

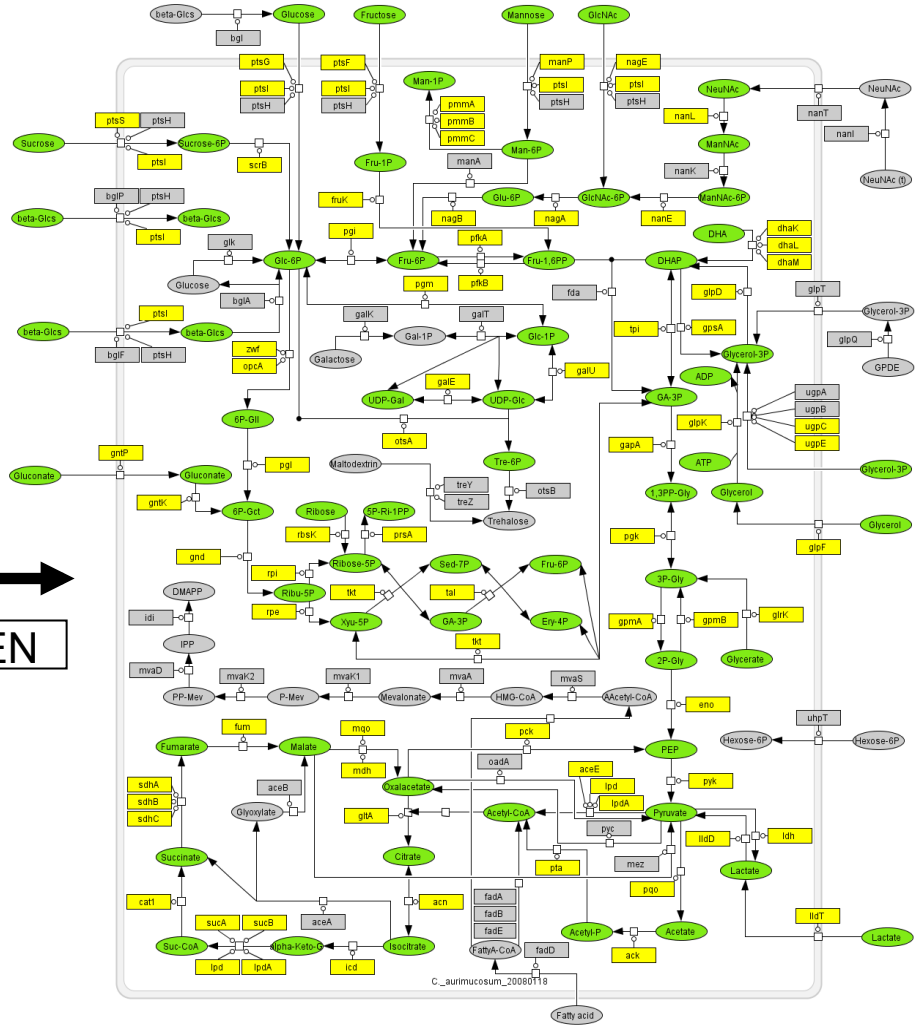
C. aurimucosum
genome sequence

+



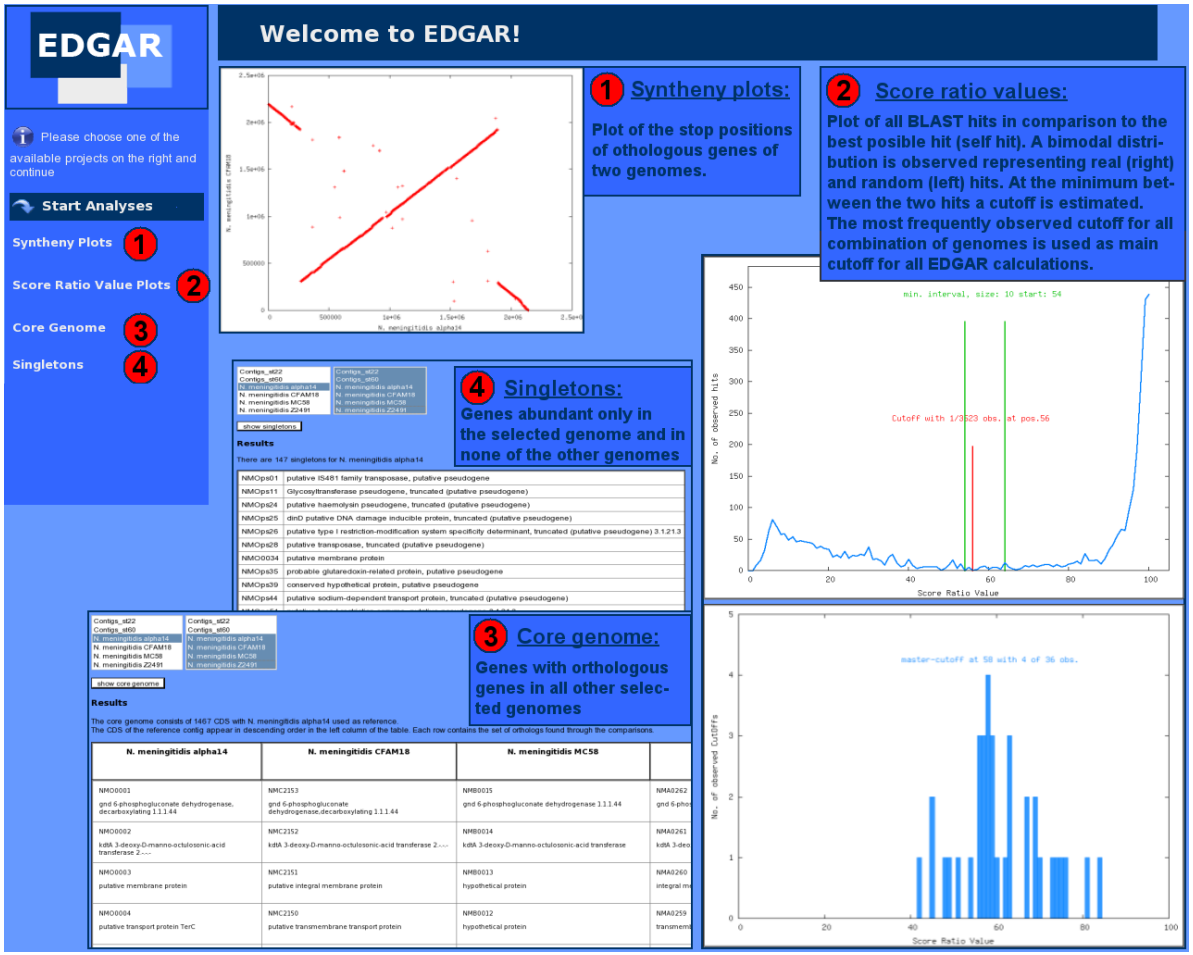
Curated model of *C. kroppenstedtii*

CARMEN



Carbohydrate metabolism of *C. aurimucosum*

Core genome and pan genome analysis with EDGAR

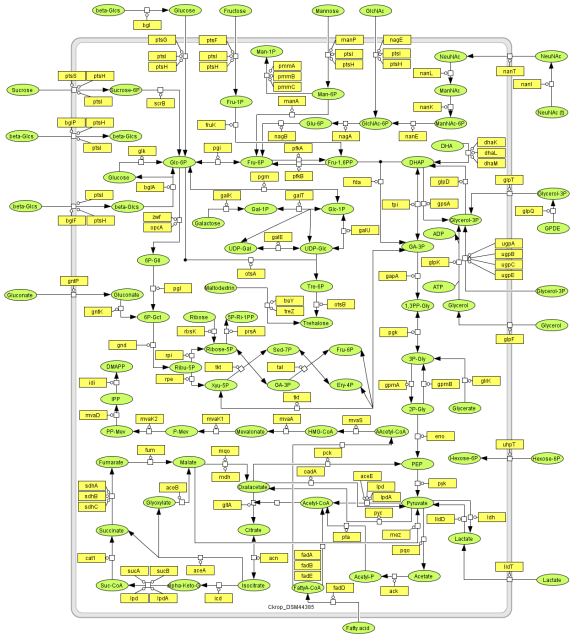


http://edgar.cebitec.uni-bielefeld.de

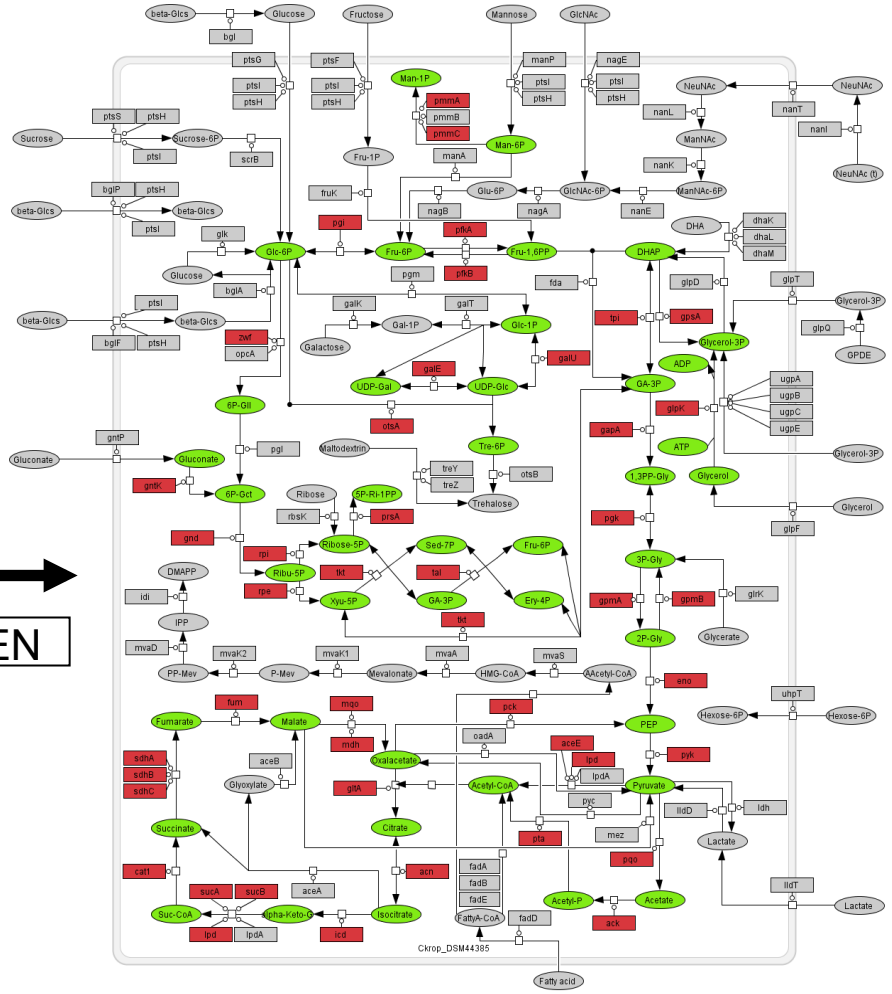
- based on Blast score ratios
- precomputed data for 582 genomes across 75 genus groups
- Venn diagrams
- Synteny plots
- Core genome
- Singletons
- Pan genome
- Comparative Viewer
- Phylogenetic Tree

EDGAR: A software framework for the comparative analysis of prokaryotic genomes.
J. Blom, S.P. Albaum, D. Doppmeier, A. Pühler, F.J. Vorhölter, M. Zakrzewski, A. Goesmann
BMC Bioinformatics. 2009 May 20;10(1):154.

CARMEN and EDGAR – Core genome mapping

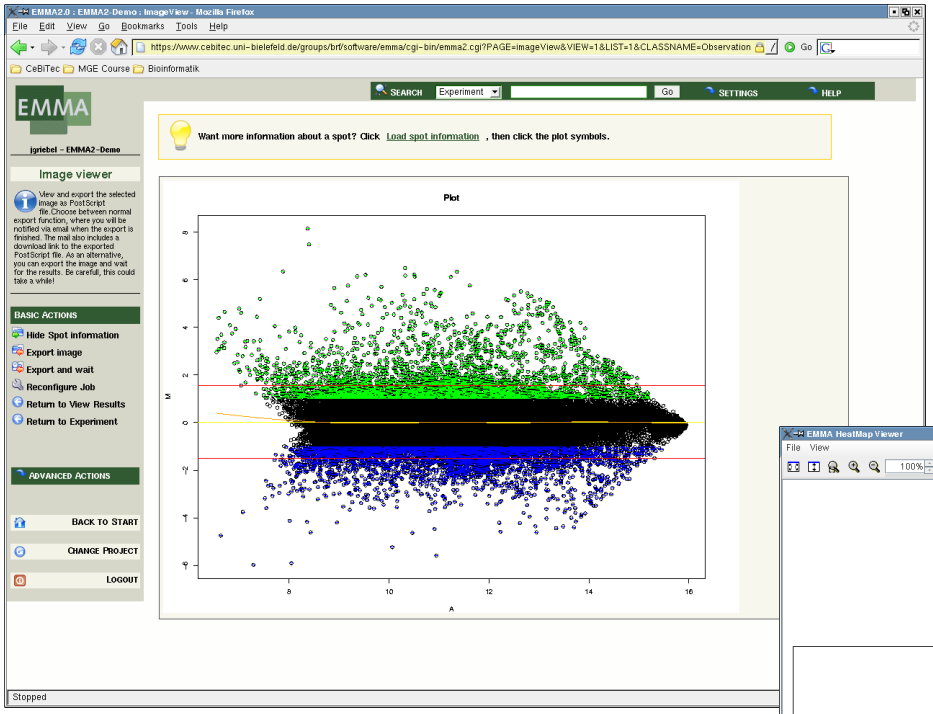


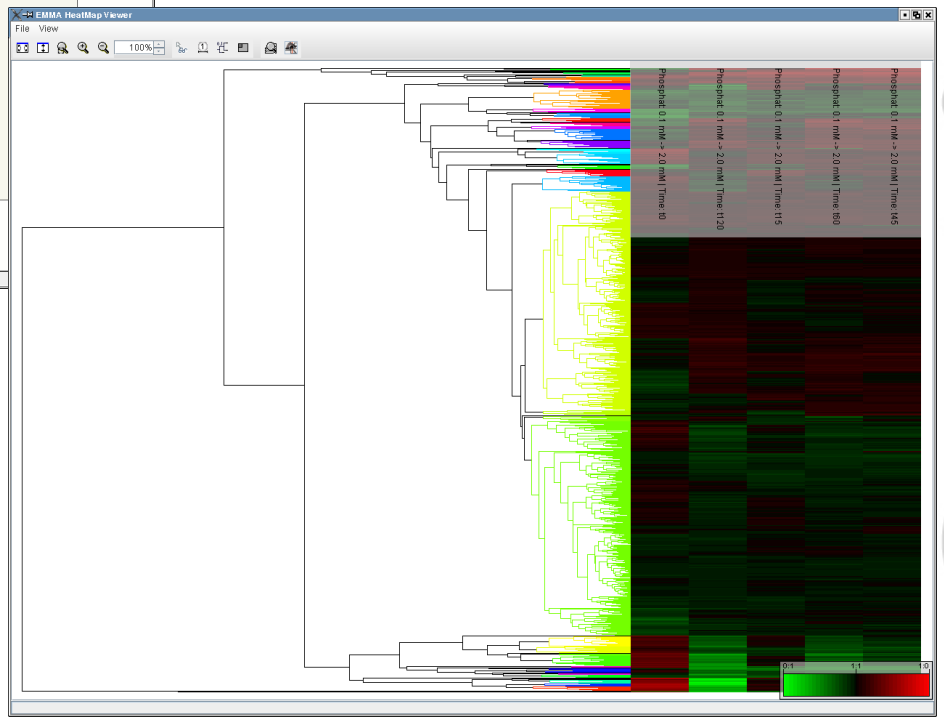
CARMEN



Core genome of *C. kroppenstedtii*, *C. diphtheriae*, *C. aurimucosum*, *C. jeikeium*, *C. urealyticum* and *C. glutamicum* based on EDGAR

EMMA 2 – High-throughput Transcriptomics Software



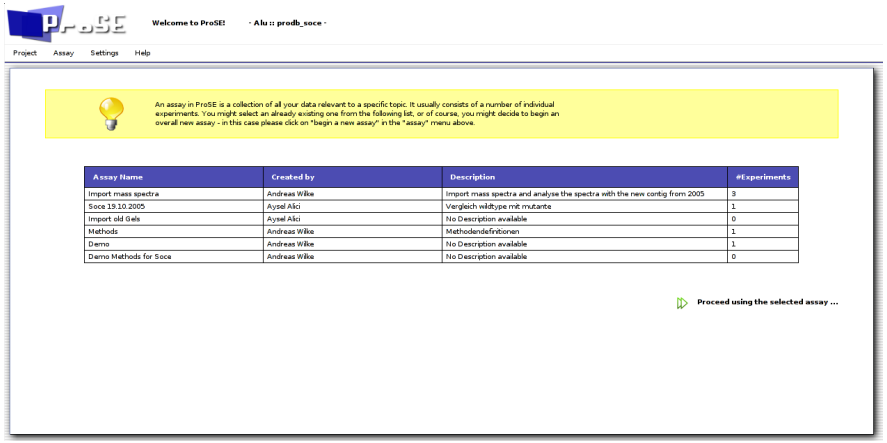


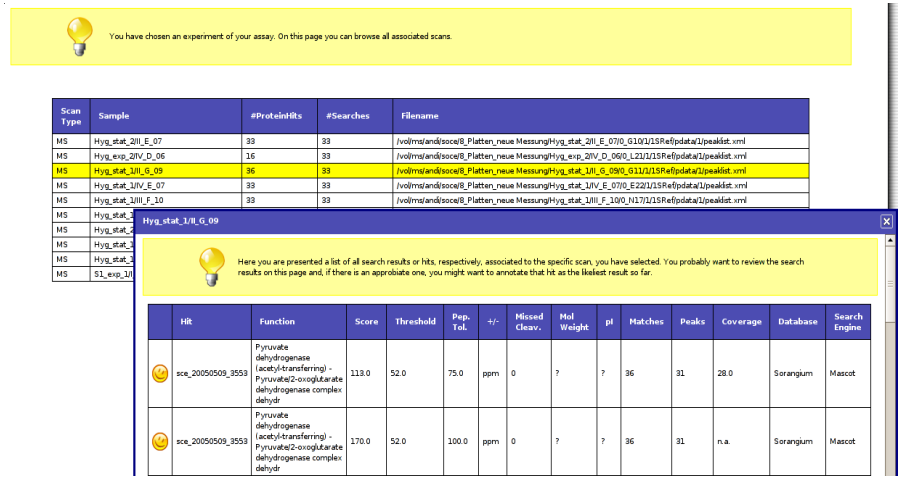
- MAGE compatible
- Separate LIMS (ArrayLIMS)
- Completely web-based

- Configurable tool pipelines
- Interactive heatmap & cluster browser
- Fine grained access control

Dondrup *et al.*, BMC Bioinformatics, 2009

QuPE





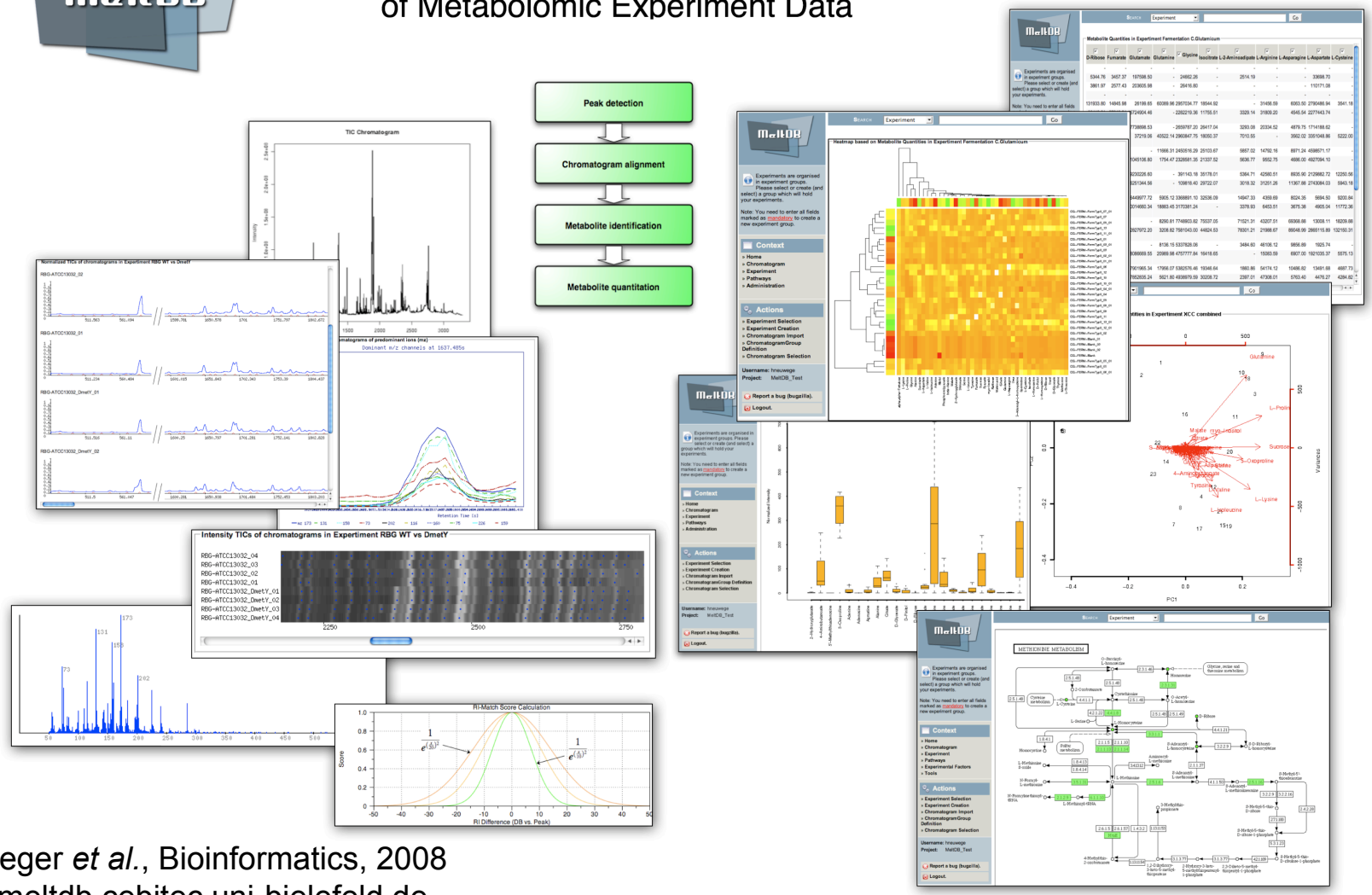
Albaum *et al.*, Bioinformatics, 2009

- Successor of ProDB, an open source system for high-throughput proteomics
- Developed in collaboration with users from Greifswald & Bochum within the BMBF QuantPro project
- Focus on automatic and manual annotation of mass spectra
- Java/AJAX frontend (Echo2)
- Reuse Importers/Exporters
- Simple but intelligent LIMS system, reduced to essential information
- Integration of algorithms for the quantitative proteome analysis

CeBiTec

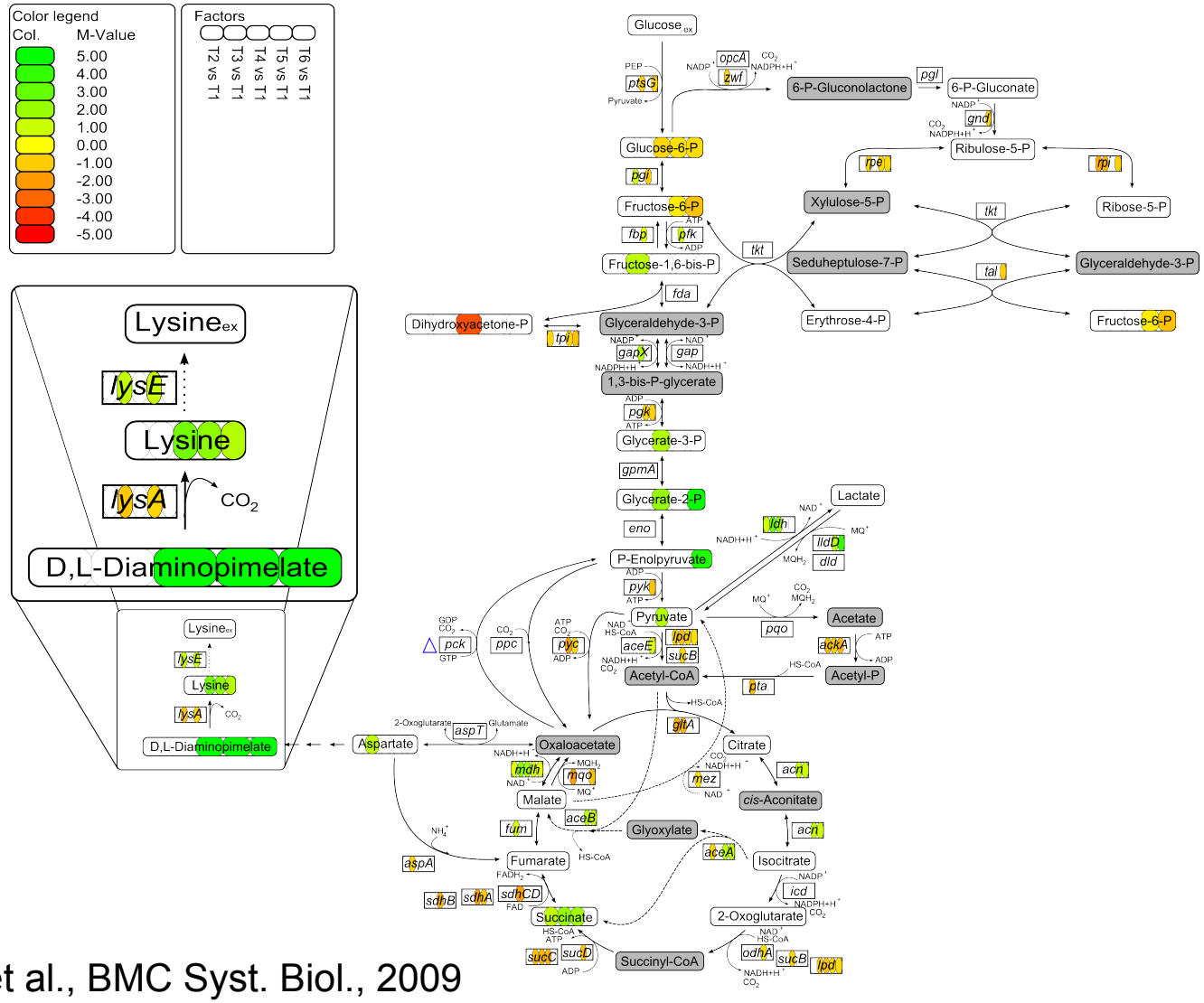
MeltDB

A Framework for the Analysis and Integration of Metabolomic Experiment Data



Neuweger *et al.*, Bioinformatics, 2008
http://meltdb.cebitec.uni-bielefeld.de

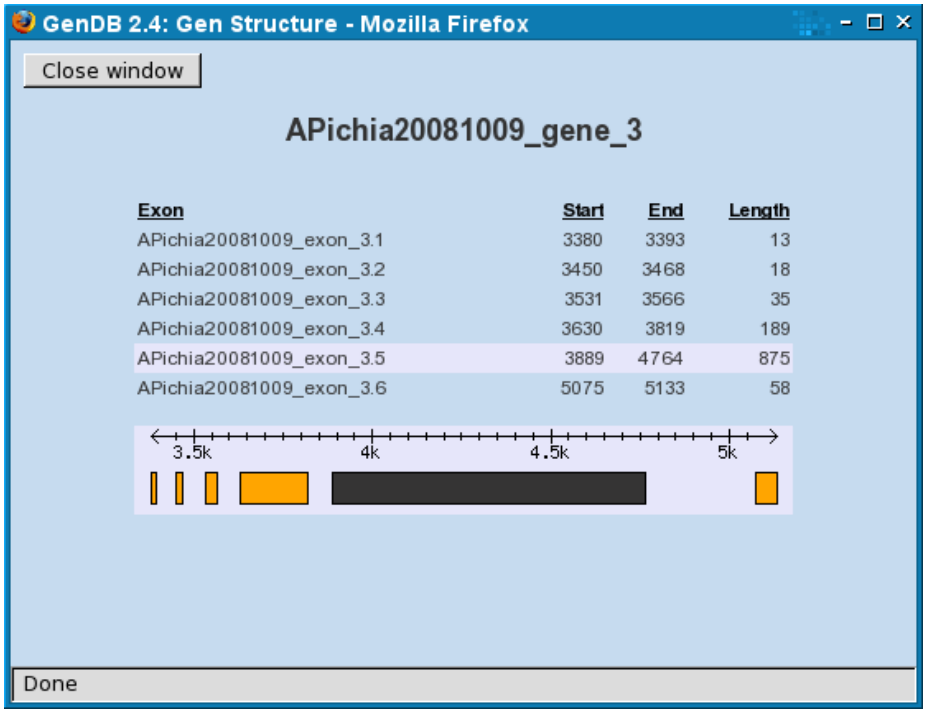
ProMeTra: Visualization of –omics data sets

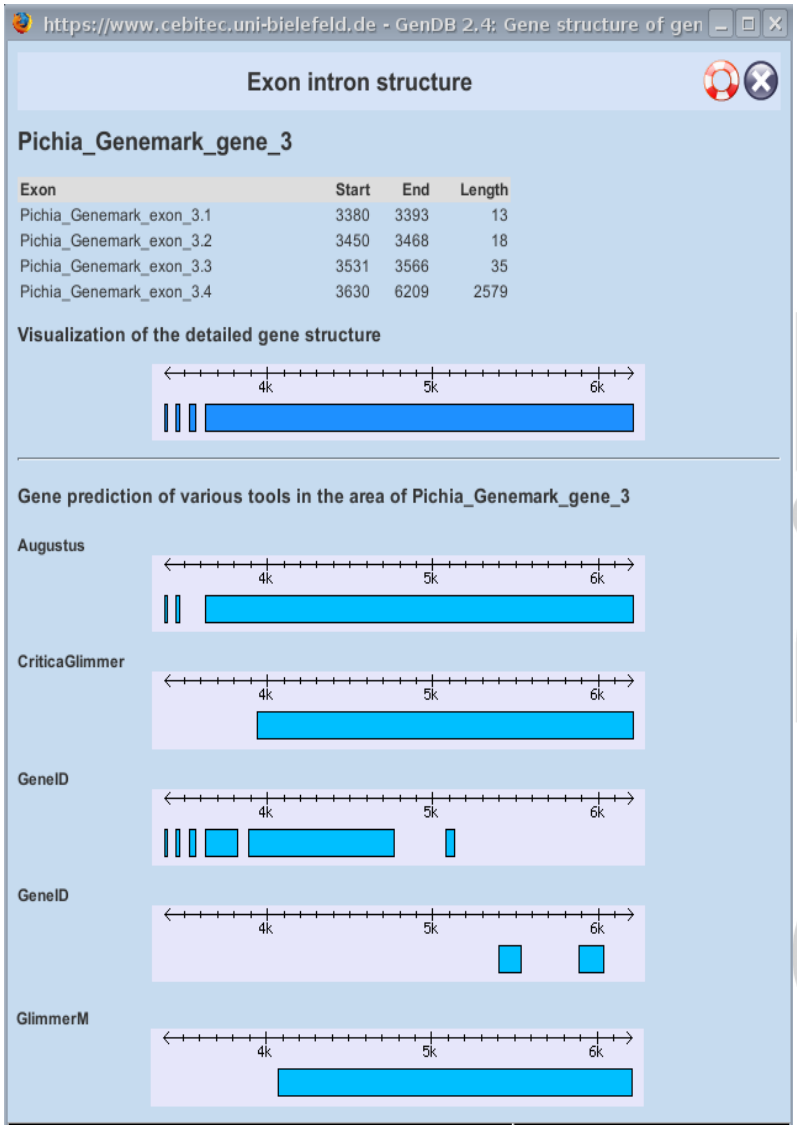


Neuweger et al., BMC Syst. Biol., 2009
http://prometra.cebitec.uni-bielefeld.de

Work in progress: Support for eukaryotic genomes

- Extend GenDB data model for storing eukaryotic genes
- Evaluate & integrate eukaryotic gene prediction tools
- Adapt annotation pipeline for eukaryotic genomes





Work in progress: Read Mapping on GPUs

Mapping of massive amounts of short read data (Solexa,454) by using modern graphic cards (GPUs) to speed up read matching against reference genomes:

- SARUMAN – **S**emiglobal **A**lignment of short **R**eads using C**U**DA and Needle**man**-Wunsch
- Exact algorithm, no heuristic
- Allows for deletions, insertions and substitutions

- Sample runtimes:

Organism	Genome size	# Reads	Runtime
S. meliloti	3.6MB	6.4M	2 minutes
M. marisnigri JR1	2.4MB	76 M	44 minutes

CeBiTec

MetaSAMS & other new tools for Metagenomics

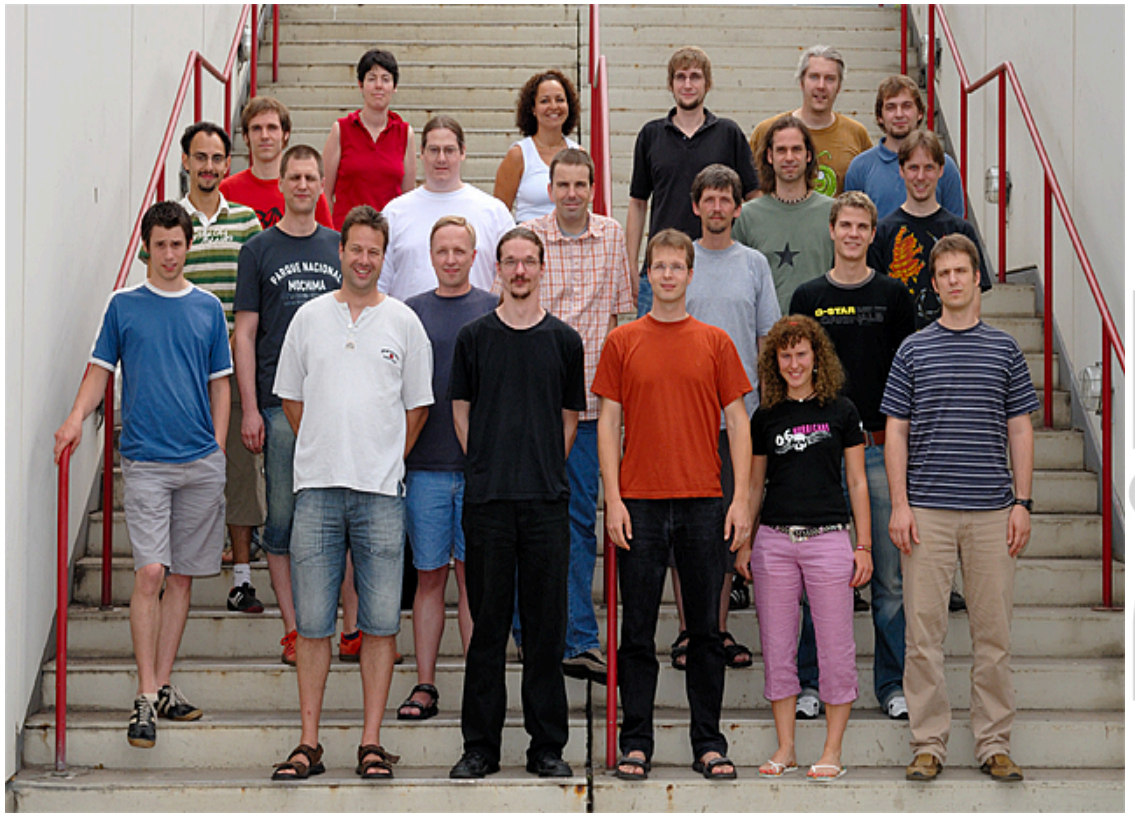
Several new bioinformatics tools are currently developed to build a comprehensive software platform for the taxonomic and functional analysis of large metagenome data sets including:

- CARMA *(Krause et al., 2008)*
- Web-CARMA *(Gerlach et al., 2009)*
- TACOA *(Diaz et al., 2009)*
- MetaSAMS *(Bekel et al., 2009)*
- Metaphor

CeBiTec

People

- Stefan Albaum
- Thomas Bekel
- Regina Bisdorf
- Jochen Blom
- Tobias Jakobi
- Sebastian Jaenicke
- Lukas Jelonek
- Sebastian Jünemann
- Burkhard Linke
- Dr. Heiko Neuweger
- Oliver Rupp
- Jessica Schneider
- Martha Zakrzewski
- Student Programmers



Group Leader: Dr. Alexander Goesmann

System Administrators:
Björn Fischer, Torsten Kasch, Achim Neumann,
Ralf Nolte, Rainer Orth, Volker Tölle

The End!

Thanks for your attention!

- Contact: agoesman@CeBiTec.Uni-Bielefeld.DE
- Homepage: http://www.cebitec.uni-bielefeld.de/brf
- Software: http://www.cebitec.uni-bielefeld.de/brf/software/brfsoftware.html

CeBiTec