

# 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



**CeBiTec**  
Center for Biotechnology  
Chairman of the Board: Prof. Dr. A. Pühler  
Executive Director: Dr. S. Weidner

October 2009

### Institute for Bioinformatics

Speaker: Prof. Dr. J. Stoye

Practical Computer Science  
Prof. Dr. R. Giegerich

Bioinformatics and Medical Informatics  
Prof. Dr. R. Hofestädt

Genome Informatics  
Prof. Dr. J. Stoye

Applied Neuroinformatics  
Juniorprof. Dr. T. W. Nattkemper

Computational Genomics  
Dr. A. Goesmann

Technology Platform BiBiServ  
Dr. A. Sczryba

### Institute for Genome Research and Systems Biology

Speaker: Prof. Dr. B. Weisshaar

Biochemistry and Plant Physiology  
Prof. Dr. K.-J. Dietz

Gene Technology and Microbiology  
Prof. Dr. R. Eichenlaub

Cell Biology  
Prof. Dr. C. Kaltschmidt

Chemical Ecology  
Prof. Dr. C. Müller

Proteome and Metabolome Research  
Prof. Dr. K. Niehaus

Genome Research of Industrial Microorganisms  
Prof. Dr. A. Pühler

Molecular Cell Physiology  
Prof. Dr. D. Staiger

Microbiology of Sustainable Energy Production  
Prof. Dr. M. Strous

Genome Research  
Prof. Dr. B. Weisshaar

Genetics of Prokaryotes  
Prof. Dr. V. F. Wendisch

RNA-Based Regulations  
PD Dr. T. Merkle

Systems Biology of Regulatory Networks  
PD Dr. A. Tauch

Technology Platform Genomics  
Dr. J. Kalinowski

### Institute for Biophysics and Nanoscience

Speaker: Prof. Dr. A. Göltzhäuser

Experimental Biophysics and Applied NanoSciences  
Prof. Dr. D. Anselmetti

Physics of Supramolecular Systems  
Prof. Dr. A. Göltzhäuser

Molecular and Surface Physics  
Prof. Dr. U. Heinzmann

Physics of Nanostructures  
Prof. Dr. A. Hüttel

Ultrafast Laser Spectroscopy  
Prof. Dr. W. Pfeiffer

Thin Films and Nanostructures  
Prof. Dr. G. Reiss

### Institute for Biochemistry and Bioengineering

Speaker: Prof. Dr. T. Noll

Biochemistry  
Prof. Dr. T. Dierks

Cellular Biochemistry  
Prof. Dr. G. Fischer von Mollard

Bioorganic Chemistry – Chemical Biology  
Prof. Dr. N. Sewald

Organic Chemistry  
Prof. Dr. J. Mattay

Structural Biochemistry  
Juniorprof. Dr. H. Niemann

Biophysical Chemistry  
Dr. T. Kottke

Bioorganic Chemistry  
PD Dr. N. Schaschke

Cell Culture Technology  
Prof. Dr. T. Noll

Fermentation Engineering  
Prof. Dr. E. Flaschel

Cellular Genetics  
Prof. Dr. H. Ragg

Algae Biotechnology and Bioenergy  
Prof. Dr. O. Kruse

### Graduate Center

Leader: Dr. S. Schneiker-Bekel

### Bioinformatics Resource Facility

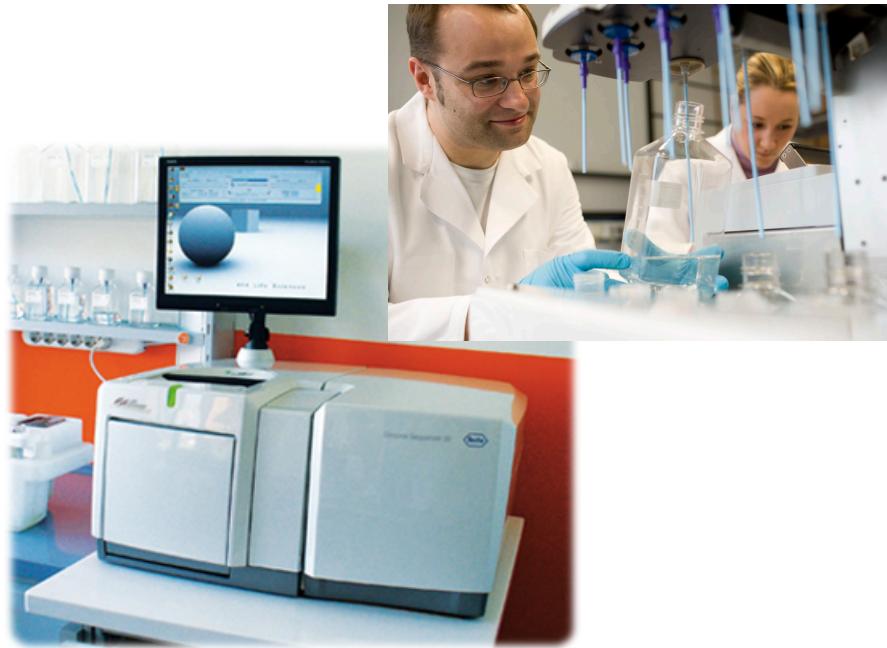
Leader: Dr. A. Goesmann

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

## Computational Genomics at CeBiTec

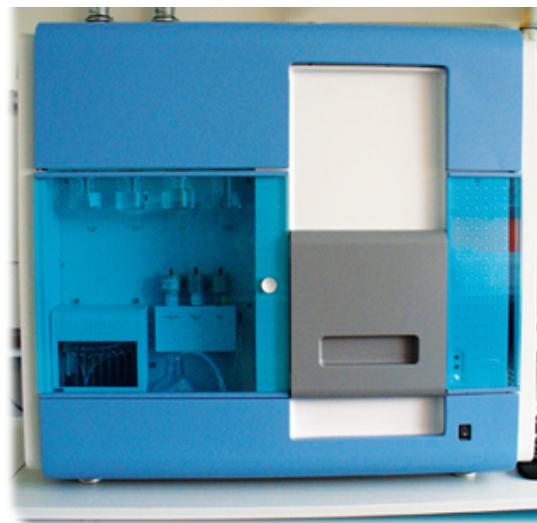
- The Bioinformatics Resource Facility (BRF) provides
  - Large scale hardware and software infrastructure for (microbial) genome research
  - Sequence analysis and genome annotation
  - Storage and analysis of gene expression data
  - Training courses
- The Junior Research Group for Computational Genomics develops new tools and software applications for
  - Multi-Omics data storage and analysis
  - Visualization
  - Data Integration
- IIT Biotech Informatics GmbH

## 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



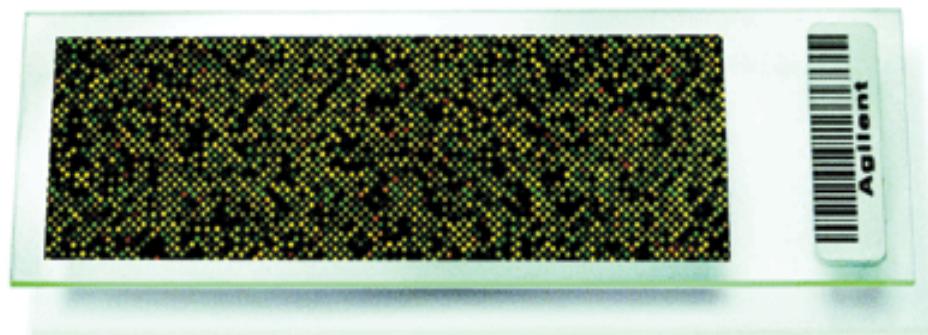
**Solexa™ Genetic Analyzer  
(Illumina)**

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

## High-throughput Transcriptomics

### Microarray Fabrication and Analysis at the CeBiTec

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

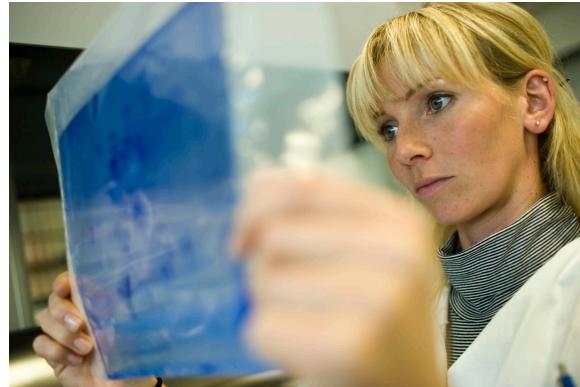


## High-throughput Proteomics

### Protein Gel Electrophoresis and Mass Spectrometry at the CeBiTec

New mass spectrometry equipment:

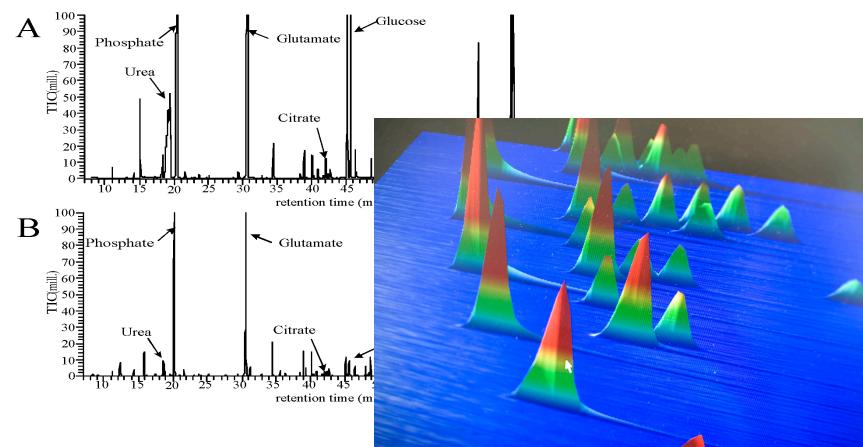
- Bruker UltrafleXtreme – MALDI-TOF/TOF
  - latest technology for LC-MALDI applications
- Bruker microTOF-QII
  - general purpose instrument for small molecules (proteomics & metabolomics)



## High-throughput Metabolomics

### Metabolite Harvesting and Analysis by Mass Spectrometry at the CeBiTec

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



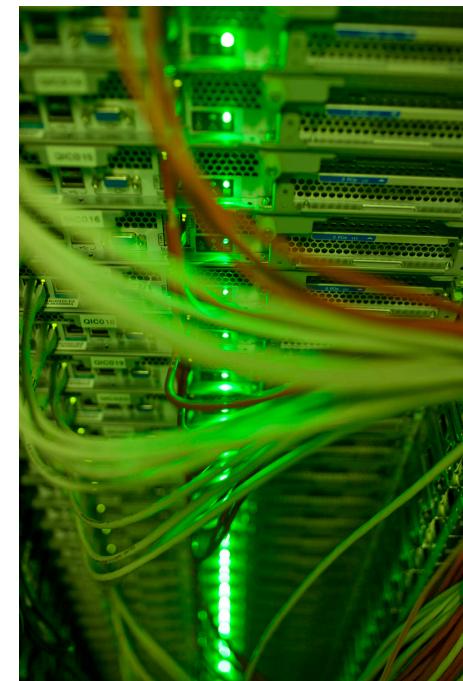
## Main Tasks of the Technology Platform Bioinformatics

- 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

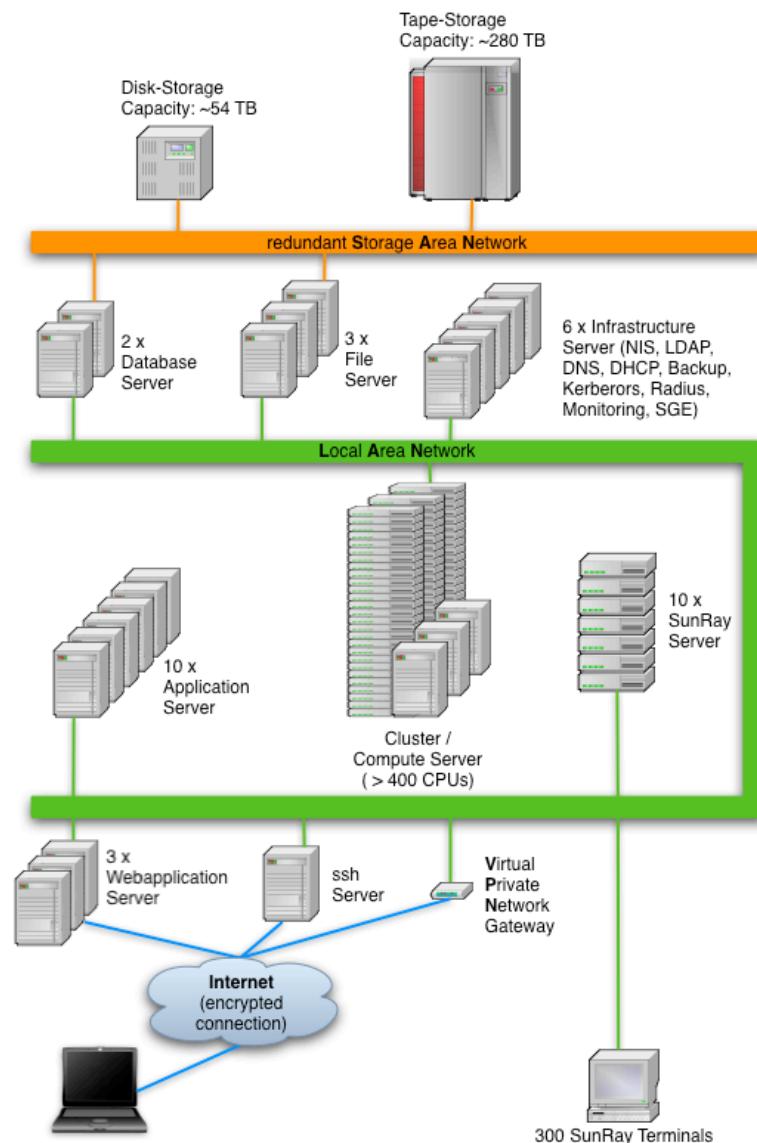
A large, semi-transparent watermark or background text is visible on the right side of the slide, reading "CeBiTec" vertically.

## 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



## The BRF Hardware Infrastructure



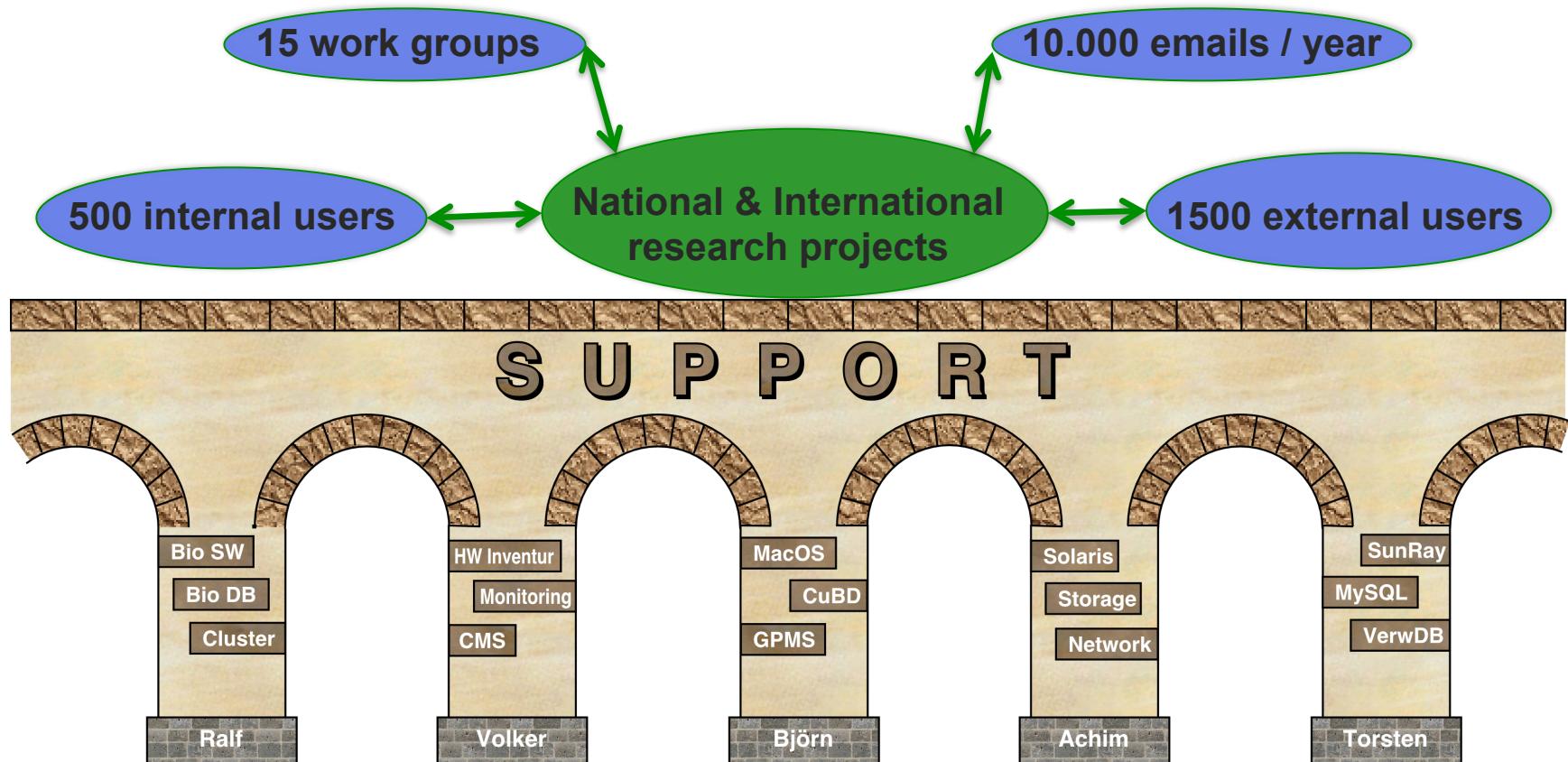
- ~ 100 TB disc storage
- ~ 280 TB tape storage
- ~ 800 CPUs (> 2200 CPU cores)
- ~ 17 TFLOPS (Rpeak) compute capacity
- investment of 850.000 EUR in 2009 / 2010

# Compute and Storage Resources

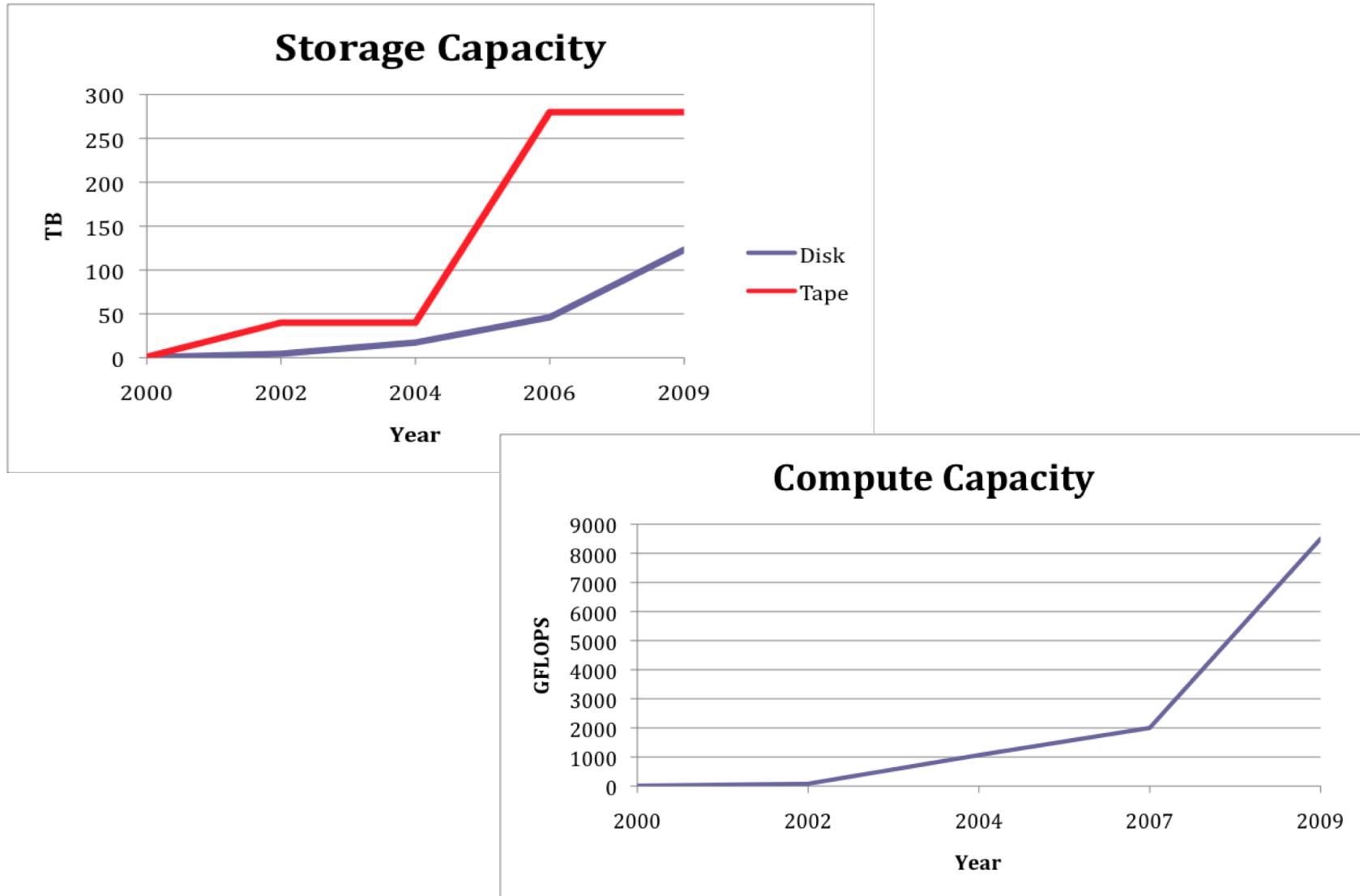


- 128 x SunFire V20z
- 28 x SunFire 2200 M2
- 76 x SunFire X2250
- 34 x PRIMERGY RX200S5
- 8 x SunFire V880/V890
- 7 x SunFire X4600
- 4 x Sun T2000
- 2 x SunFire X4440
- 9 x PRIMERGY RX600S5
- Sun StorEdge 351x and J4500 JBOD
- 2 x Sun StorEdge L700
- 1 x Quantum ATL P7000
- 1 x DeCypher SeqCruncher

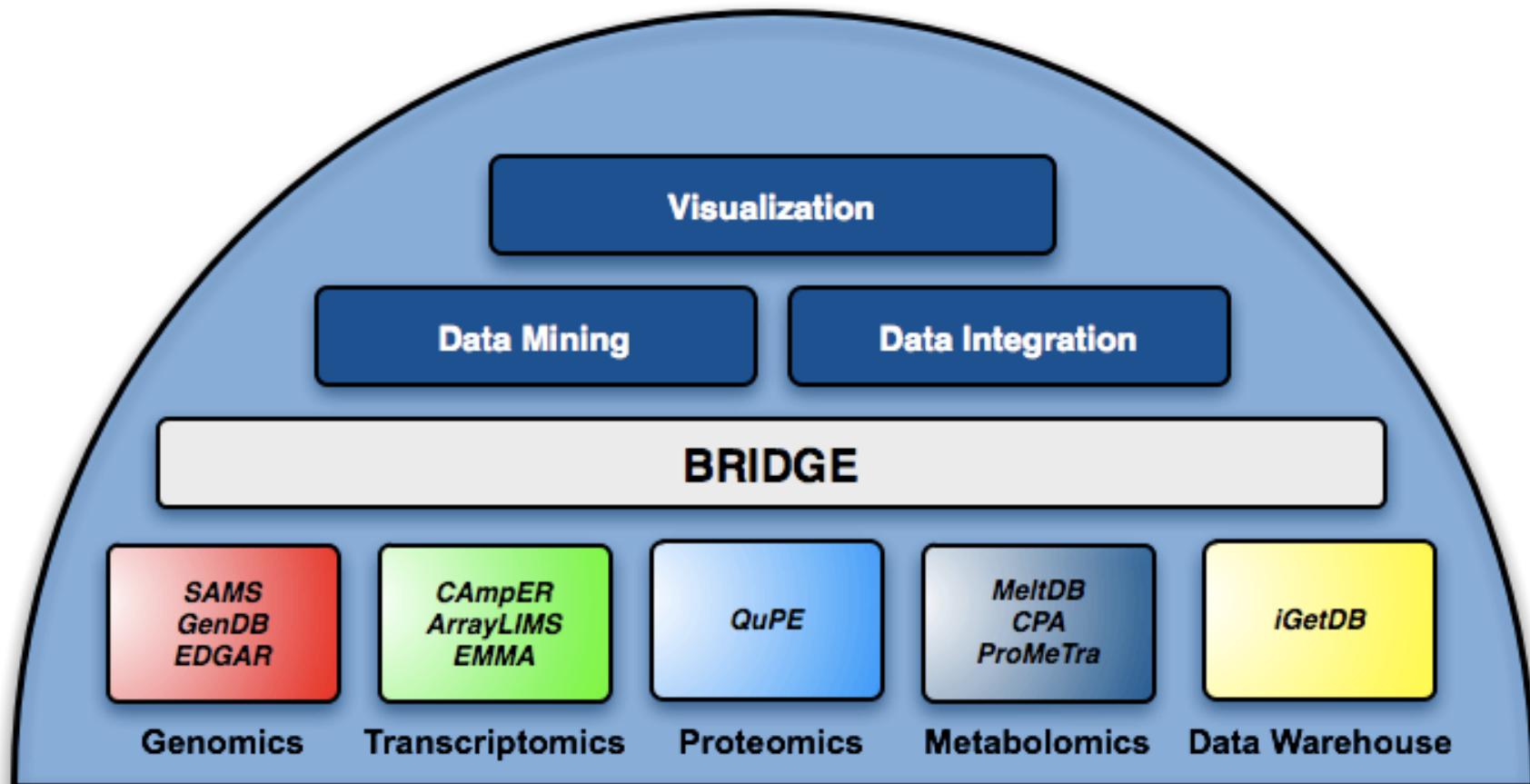
## BRF System Administration



## Development of Resources



## 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

## Genome analysis with SAMS, GenDB, Carmen and EDGAR

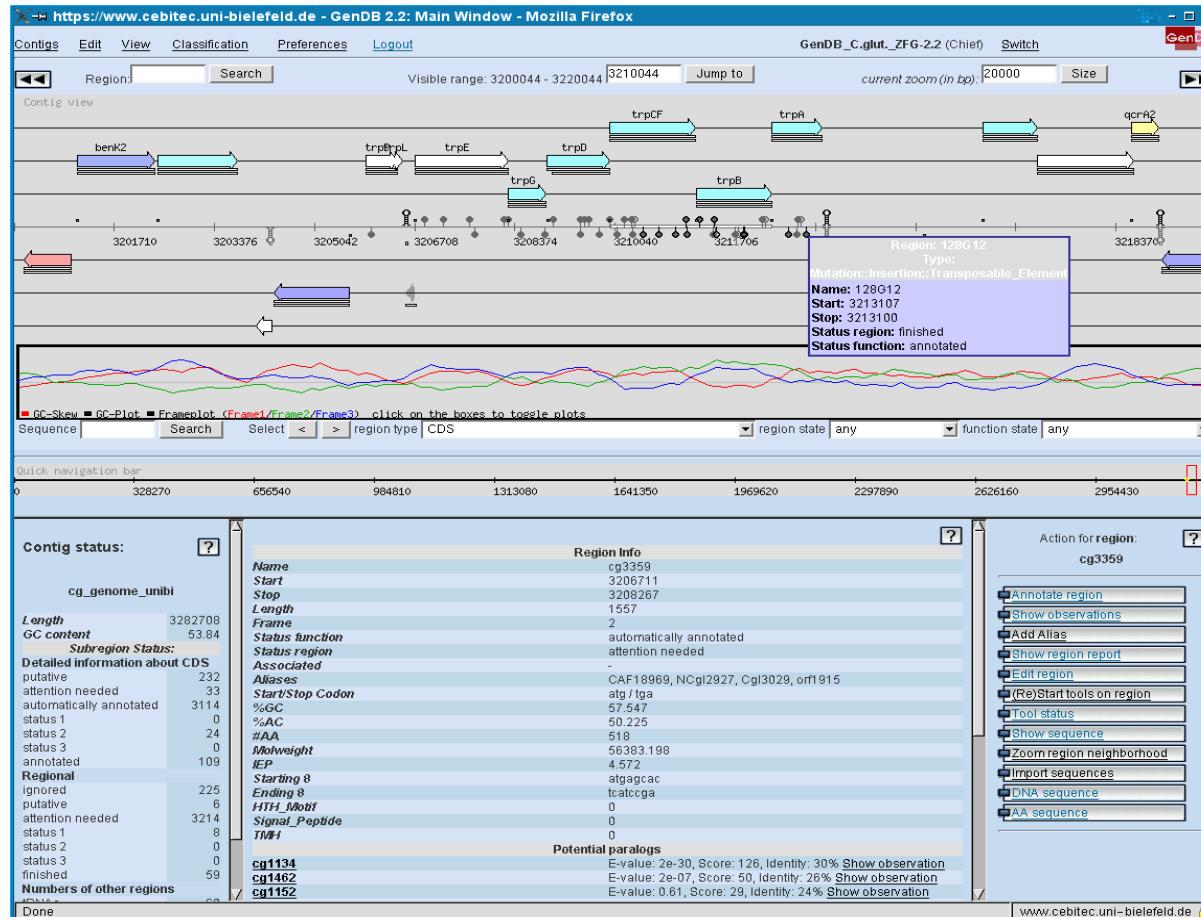
### Ultrafast Genome Sequencing and Annotation

- 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

bio  
informatics  
CeBiTec

A large, faint watermark or background text on the right side of the slide reads "bio", "informatics", "CeBiTec" vertically from bottom to top. The letters are in a light gray sans-serif font.

# GenDB – Distributed genome annotation



- Automatic & manual annotation

- CDS, tRNA, rRNA, IS elements, oligos, mutations, operons, terminators, ...

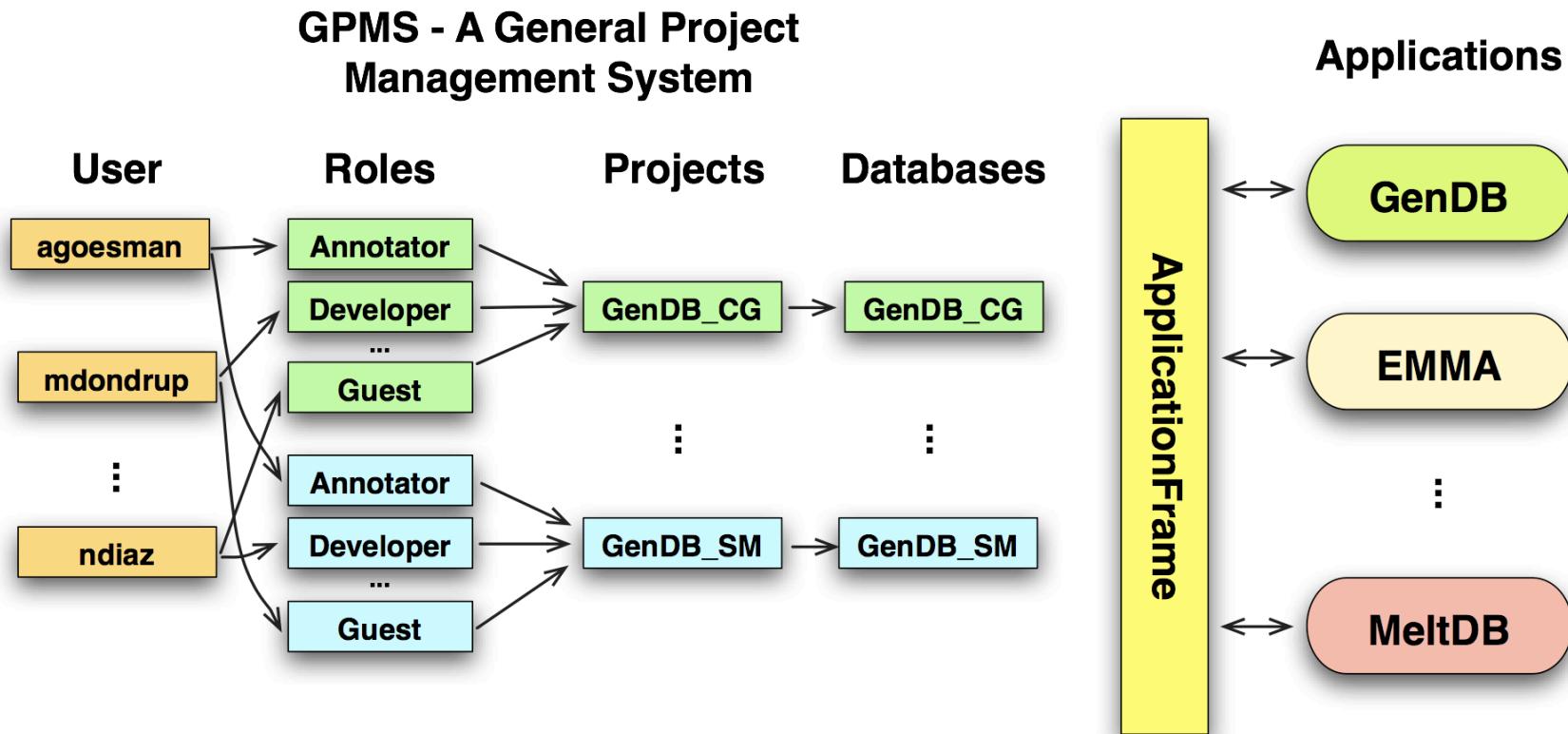
- KEGG, COG, GO

- genome maintenance and re-annotation

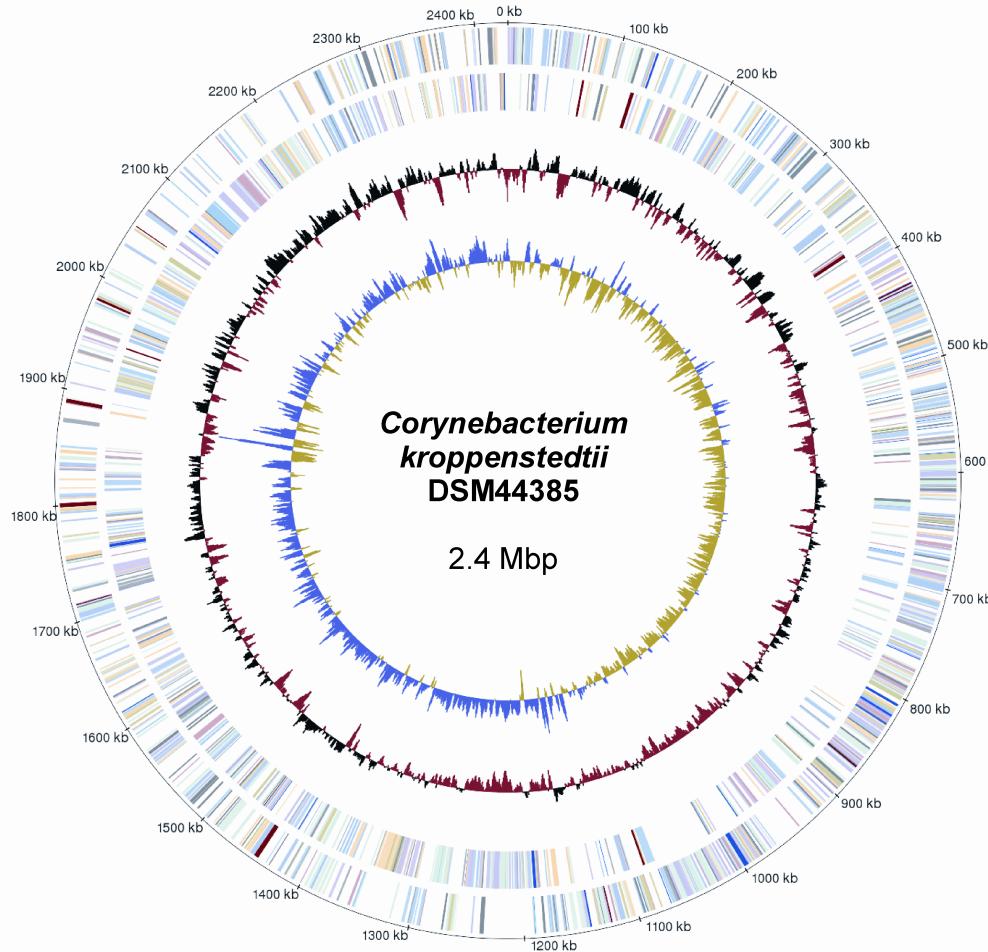
Manual genome annotation by distributed teams via web interface.

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
<b>Assembled contigs</b>	<b>6</b>
(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

## One run – Done!

- training
  - end of September, 2007



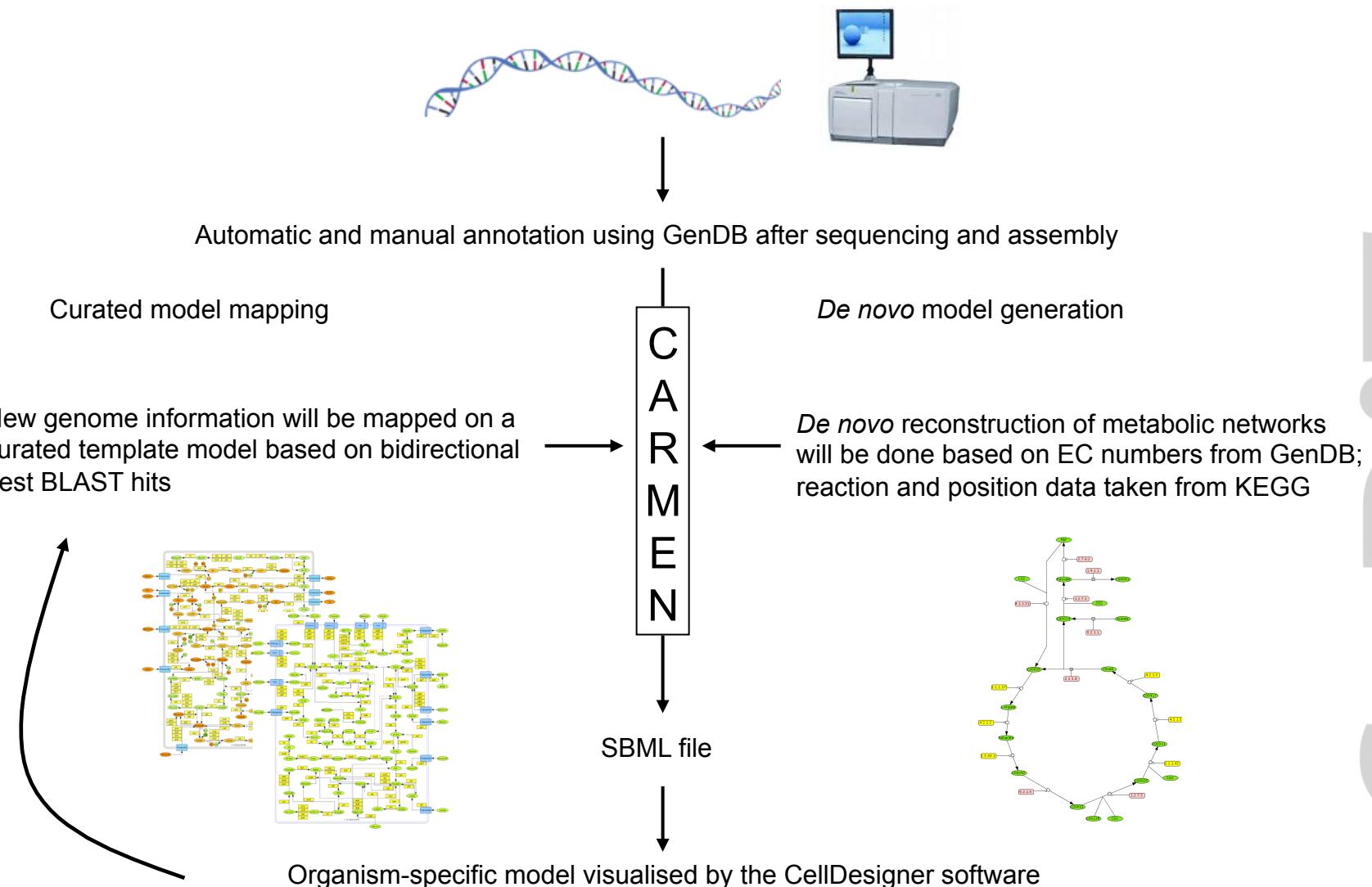
- paper submitted
  - December, 2007
- paper accepted
  - March, 2008

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

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

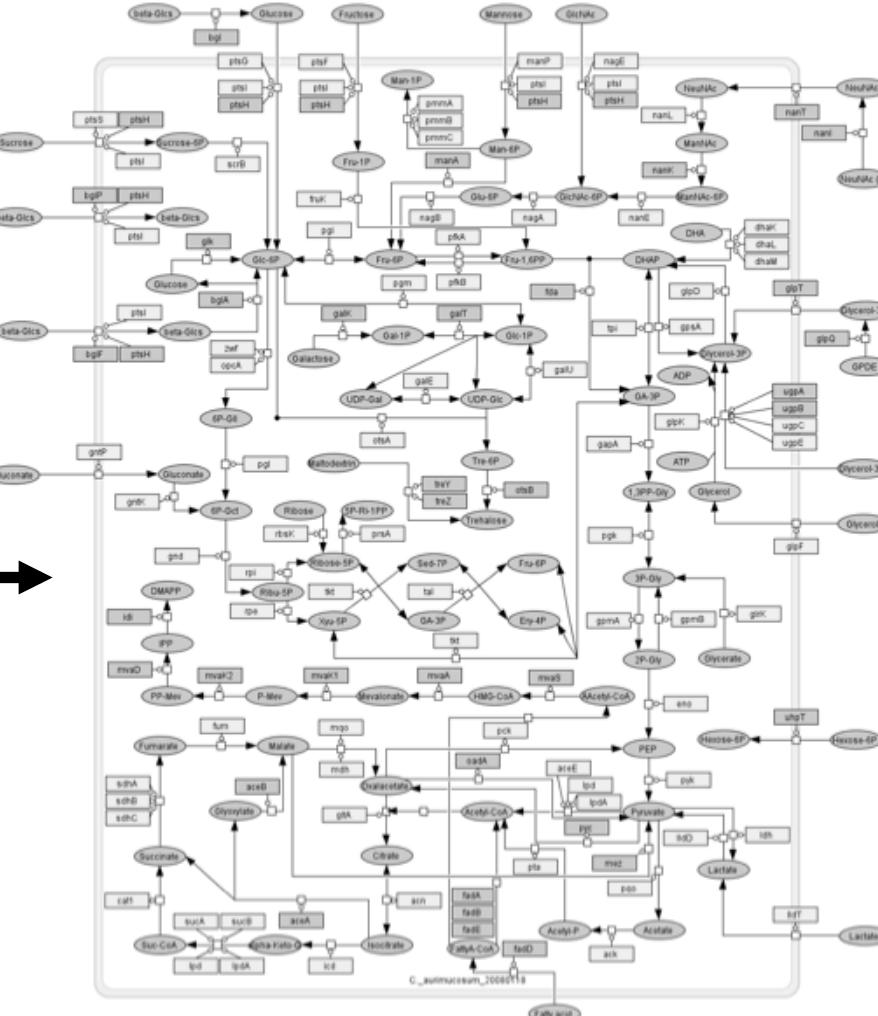
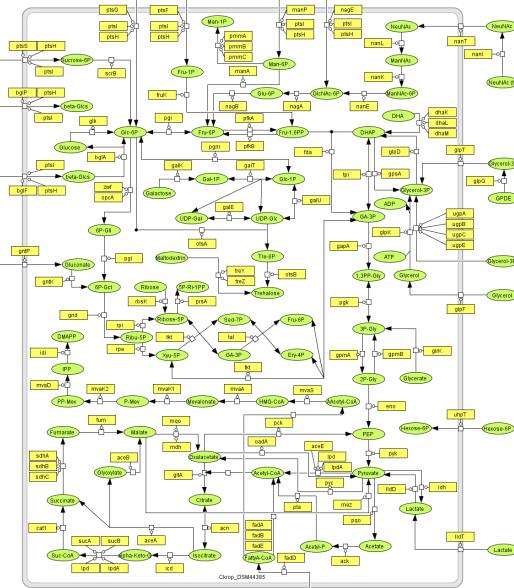
one training → one run → one genome → one publication

## In silico reconstruction of metabolic pathways



# *In silico* reconstruction by curated model mapping

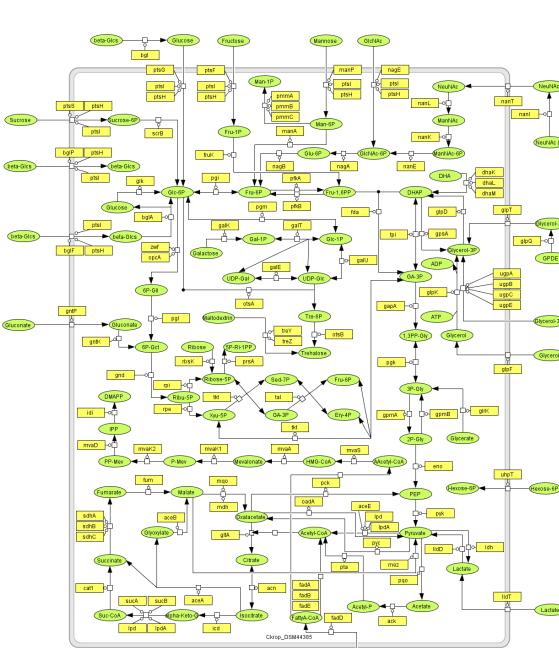
*C. aurimucosum*  
genome sequence



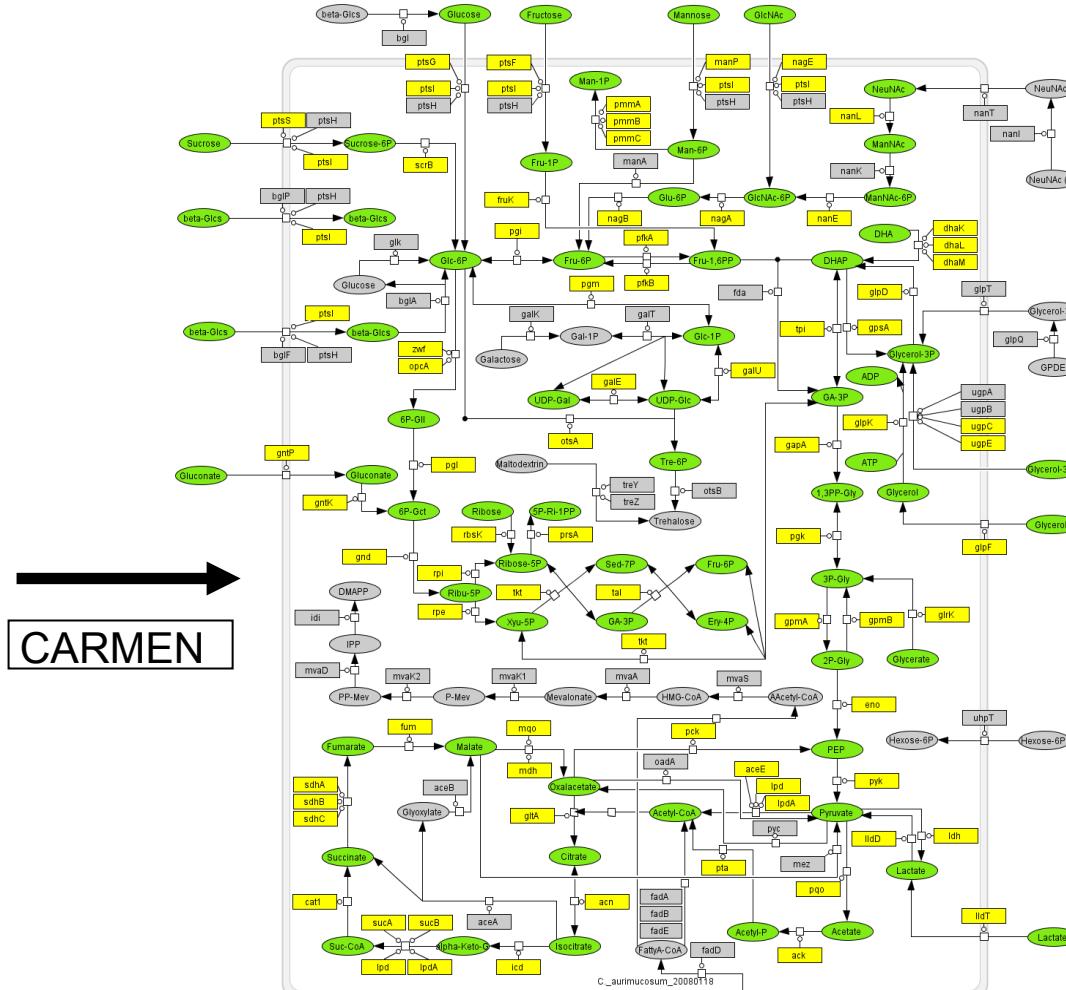
Curated model of *C. kroppenstedtii*

# In silico reconstruction by curated model mapping

*C. aurimucosum*  
genome sequence



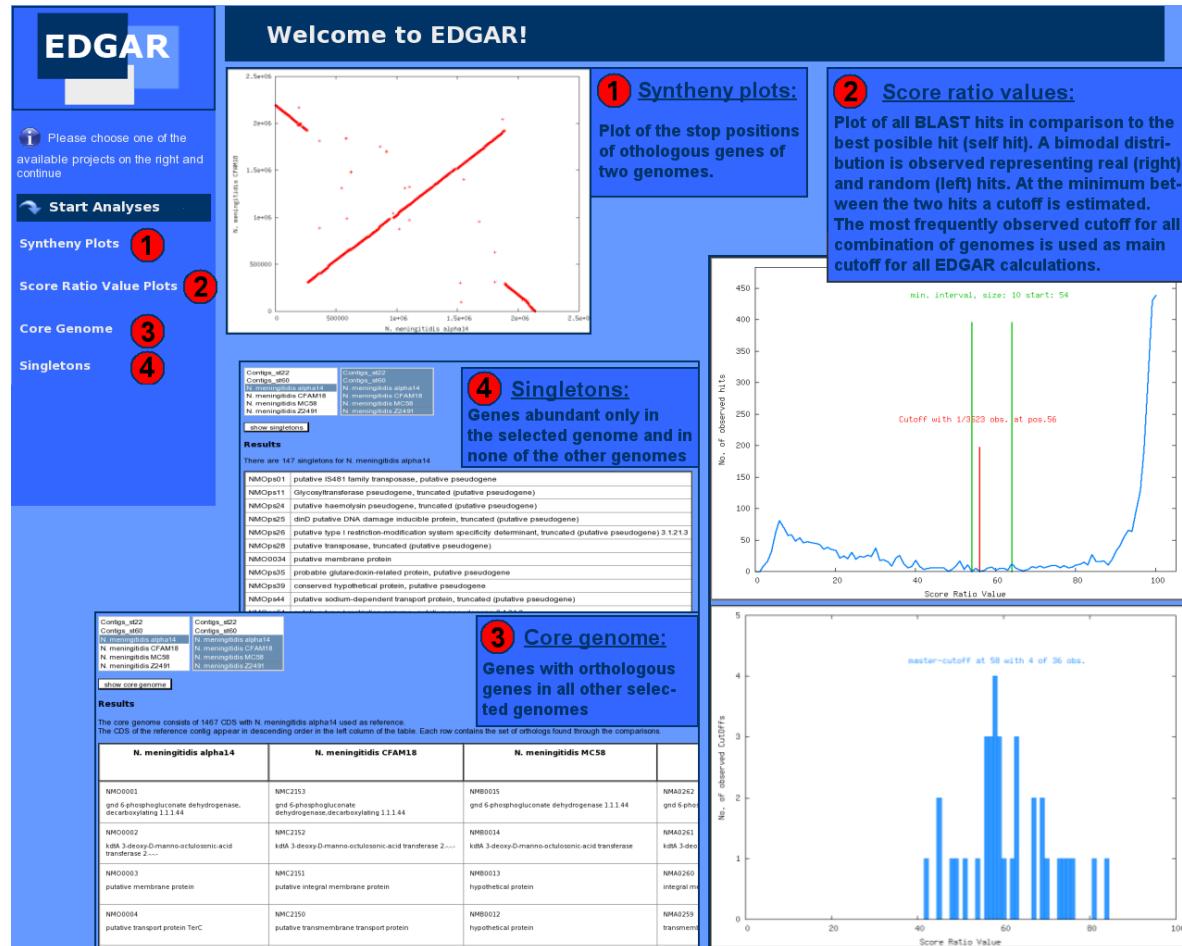
CARMEN



Curated model of *C. kroppenstedtii*

Carbohydrate metabolism of *C. aurimucosum*

# Core genome and pan genome analysis with EDGAR



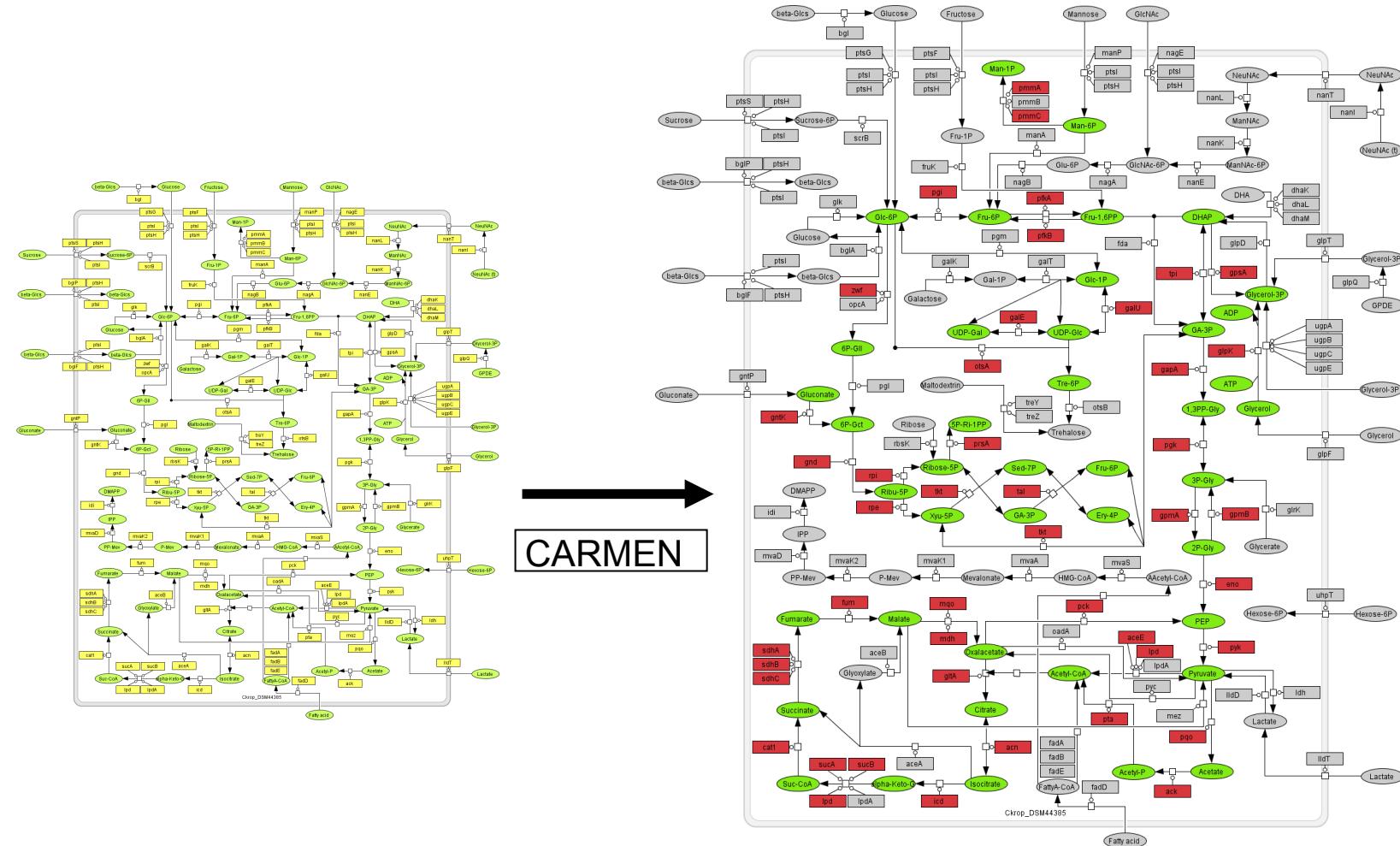
<http://edgar.cebitec.uni-bielefeld.de>

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.

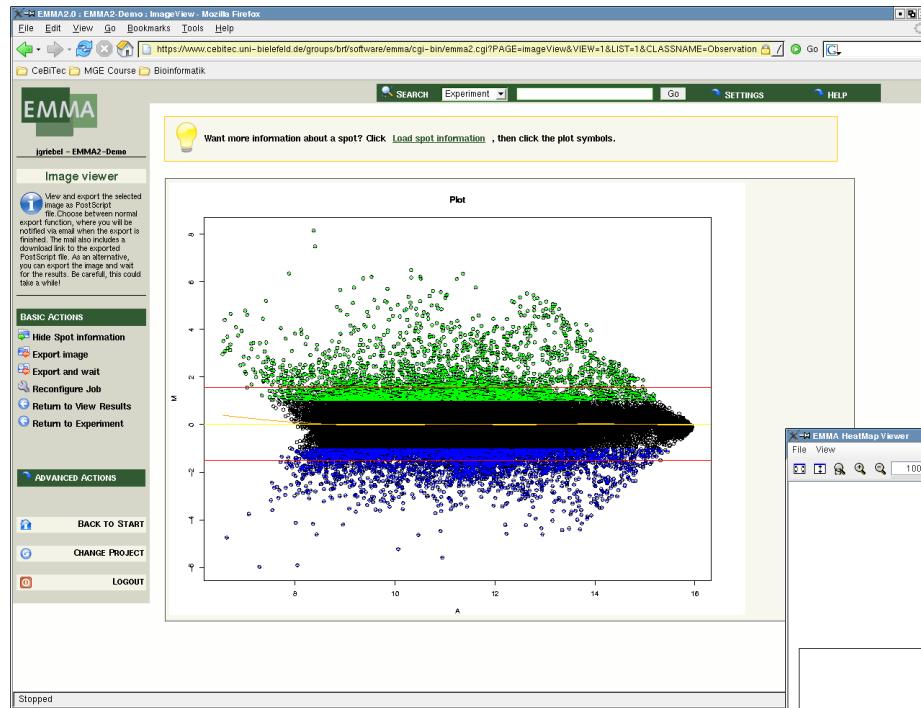
- based on Blast score ratios
- precomputed data for 582 genomes across 75 genus groups
- Venn diagrams
- Synteny plots
- Core genome
- Singletons
- Pan genome
- Comparative Viewer
- Phylogenetic Tree

# CARMEN and EDGAR – Core genome mapping

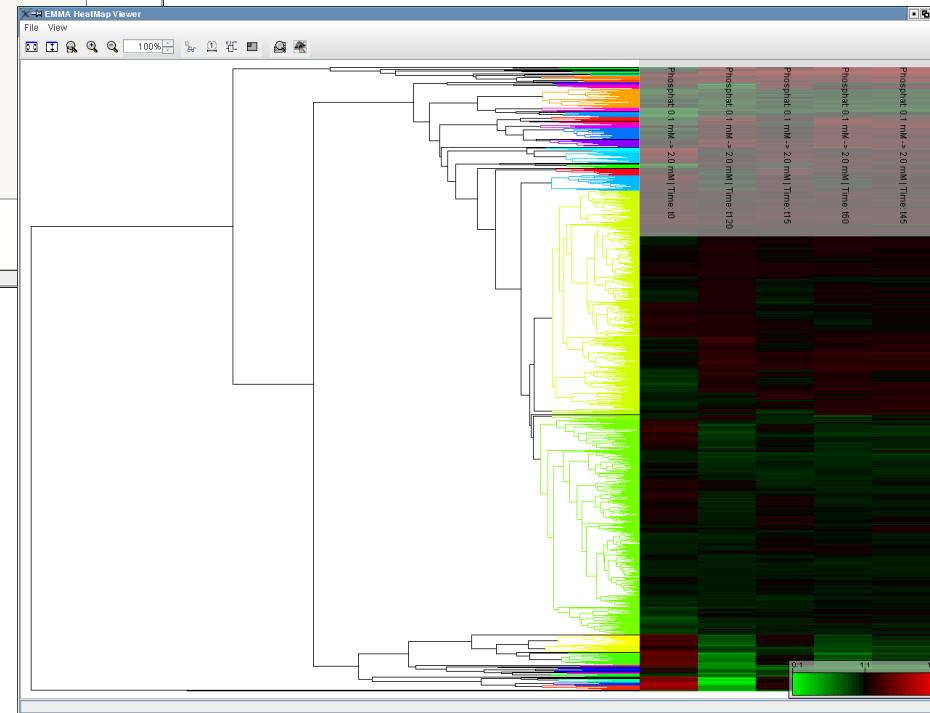


Core genome of *C. kroppenstedtii*, *C. diphtheriae*, *C. aurimucosum*, *C. jeikeium*, *C. urealyticum* and *C. glutamicum* based on EDGAR

# EMMA 2 – High-throughput Transcriptomics Software



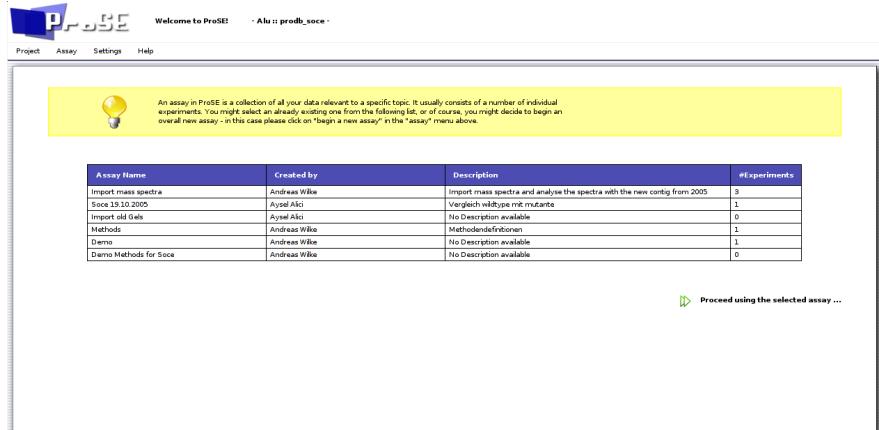
- MAGE compatible
- Separate LIMS (ArrayLIMS)
- Completely web-based



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

Dondrup *et al.*, BMC Bioinformatics, 2009

# QuPE

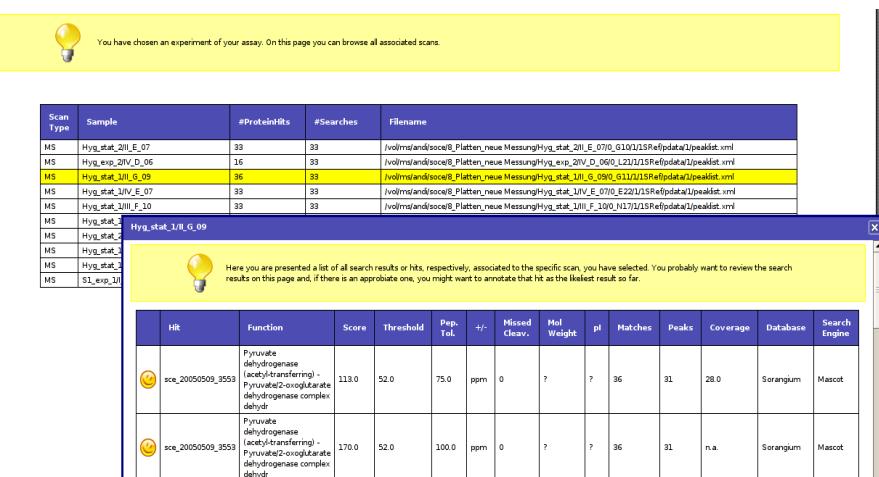


Welcome to ProSE - Alu :: prodB\_socE ::

An assay in ProSE is a collection of all your data relevant to a specific topic. It usually consists of a number of individual experiments. You might select an already existing one from the following list, or, of course, you might decide to begin an overall new assay - in this case please click on "begin a new assay" in the "assay" menu above.

Assay Name	Created by	Description	#Experiments
Import mass spectra	Andreas Wike	Import mass spectra and analyse the spectra with the new config from 2005	3
Soco 19.10.2005	Ayseg Akg	Vergleich wildtyp vs mutanten	1
Import old Gels	Ayseg Akg	No Description available	0
Methods	Andreas Wike	Methodenfraktionen	1
Demo	Andreas Wike	No Description available	1
Demo Methods for Soco	Andreas Wike	No Description available	0

Proceed using the selected assay ...



You have chosen an experiment of your assay. On this page you can browse all associated scans.

Scan Type	Sample	#ProteinHRs	#Searches	Filename
MS	Hyg_stat_20II_E_07	39	33	/vol/mranda/soco@Platten_neue_Messung/Hyg_stat_20II_E_070_G10/2/15Refidata/1/peaklist.xml
MS	Hyg_exp_20V_D_06	16	33	/vol/mranda/soco@Platten_neue_Messung/Hyg_exp_20V_D_060_G060_L21/2/15Refidata/1/peaklist.xml
MS	Hyg_stat_20II_G_09	36	33	/vol/mranda/soco@Platten_neue_Messung/Hyg_stat_20II_G_090_G11/2/15Refidata/1/peaklist.xml
MS	Hyg_stat_20II_E_07	39	33	/vol/mranda/soco@Platten_neue_Messung/Hyg_stat_20II_E_070_G22/2/15Refidata/1/peaklist.xml
MS	Hyg_stat_20II_F_10	39	33	/vol/mranda/soco@Platten_neue_Messung/Hyg_stat_20II_F_100_G17/2/15Refidata/1/peaklist.xml
MS	Hyg_stat_20II_H_09			
MS	Hyg_stat_20II_I_09			
MS	Hyg_stat_20II_J_09			
MS	SL_exp_20I			

Here you are presented a list of all search results or hits, respectively, associated to the specific scan, you have selected. You probably want to review the search results on this page and, if there is an appropriate one, you might want to annotate that hit as the likeliest result so far.

Hit	Function	Score	Threshold	Pep. Tol.	+/-	Missed Cleav.	Mol Weight	pI	Matches	Peaks	Coverage	Database	Search Engine
sce_20050509_3553	Pyruvate dehydrogenase (lipoate)-transfering - Pyruvate2-oxoglutarate dehydrogenase complex dehydr	113.0	52.0	75.0	ppm	0	?	?	36	31	28.0	Sorangium	Mascot
sce_20050509_3553	Pyruvate dehydrogenase (lipoate)-transfering - Pyruvate2-oxoglutarate dehydrogenase complex dehydr	170.0	52.0	100.0	ppm	0	?	?	36	31	n.a.	Sorangium	Mascot

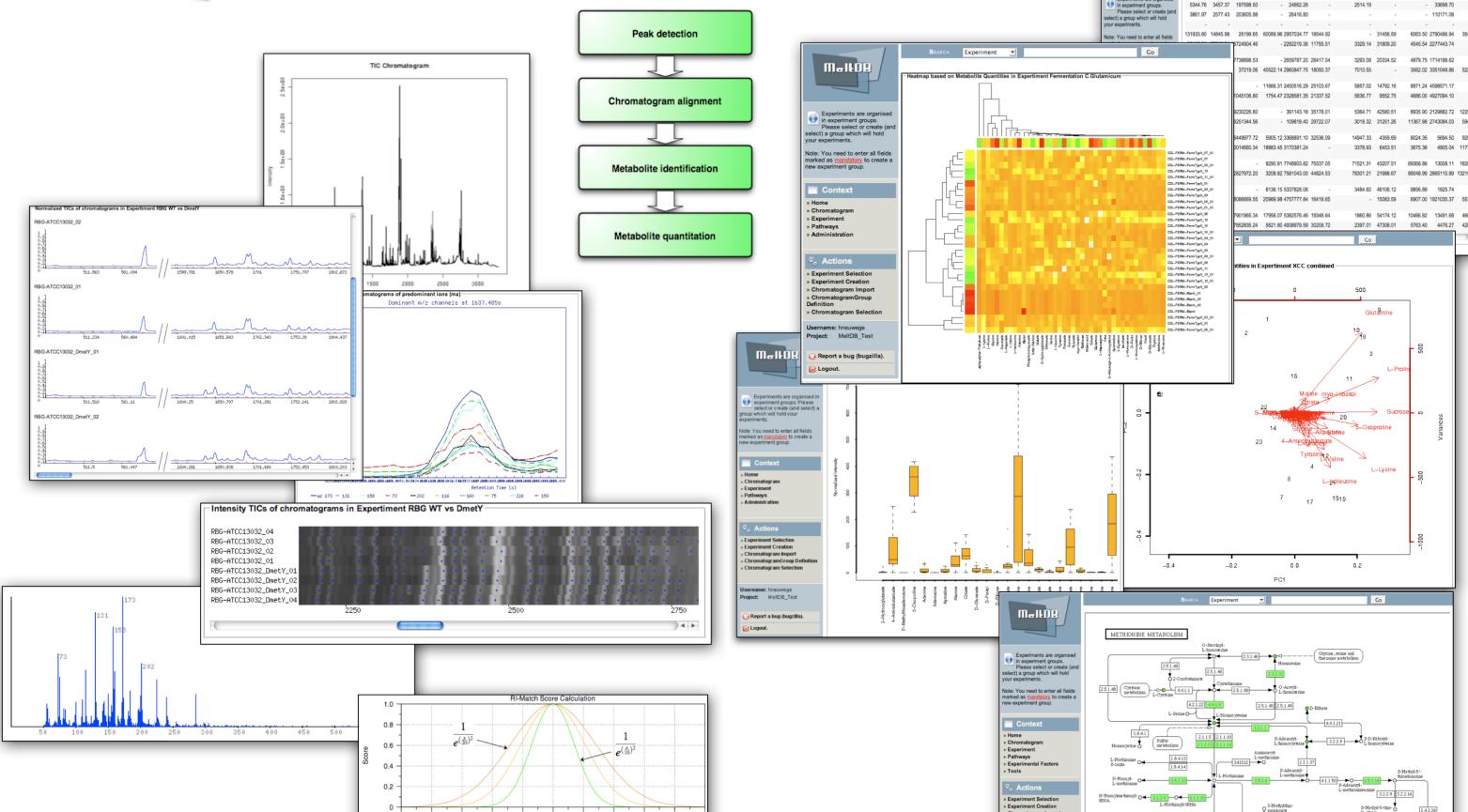
Albaum *et al.*, Bioinformatics, 2009

- 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

# MeltDB

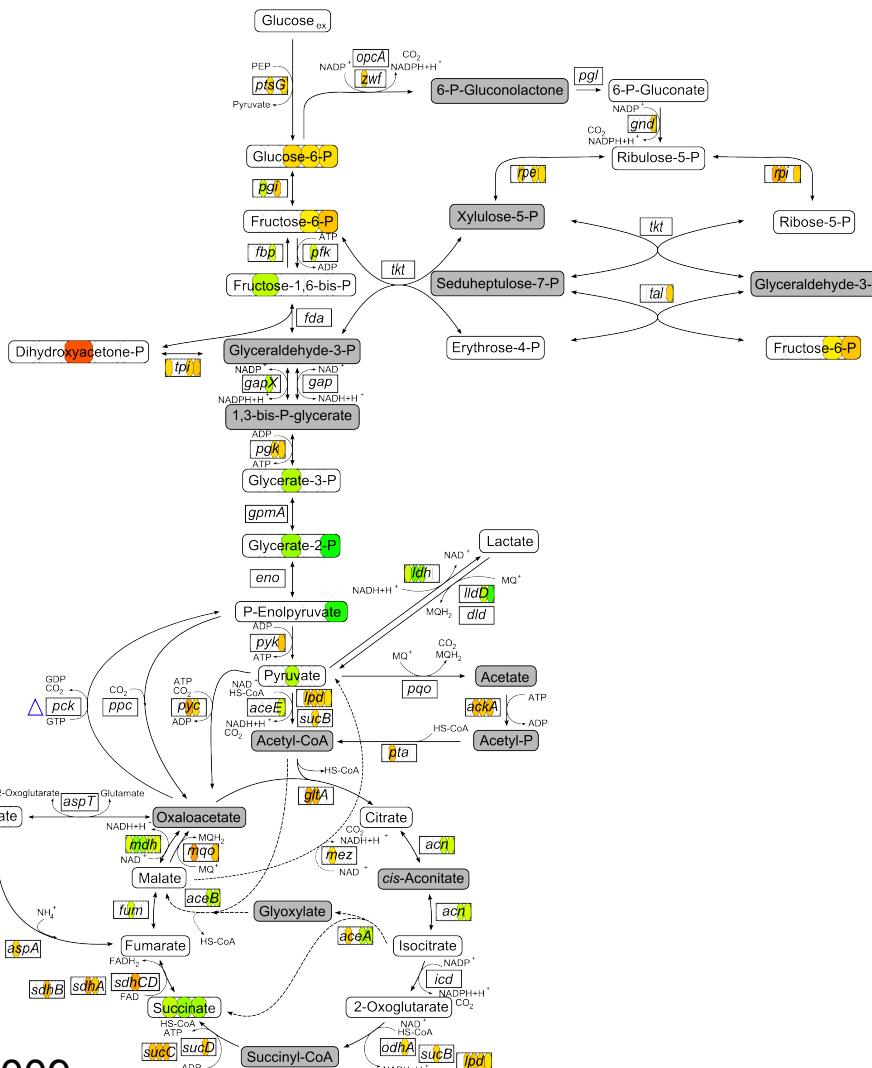
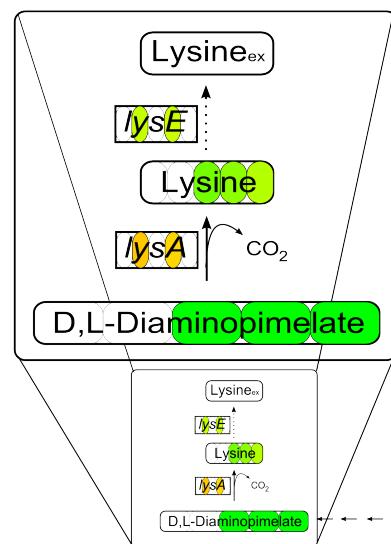
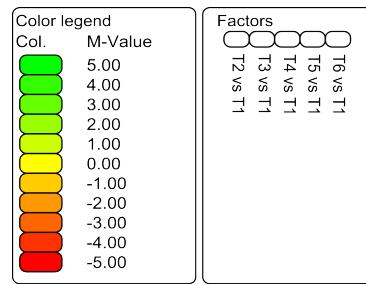


## A Framework for the Analysis and Integration of Metabolomic Experiment Data



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

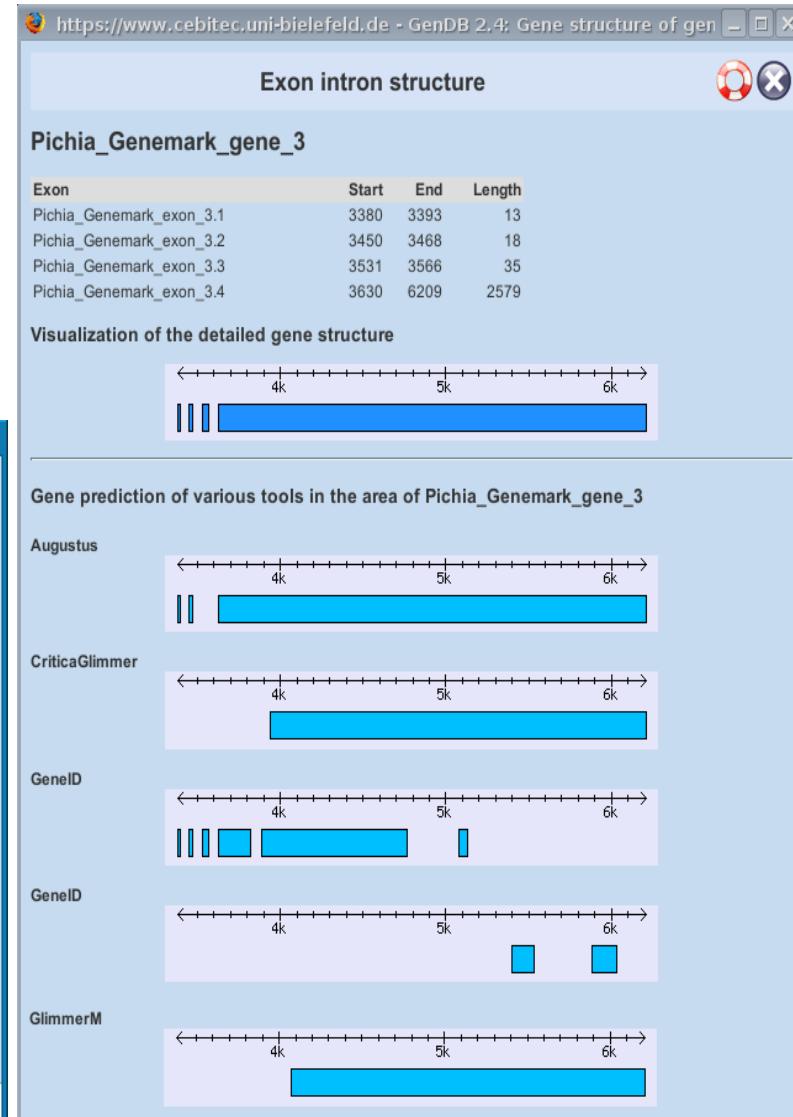
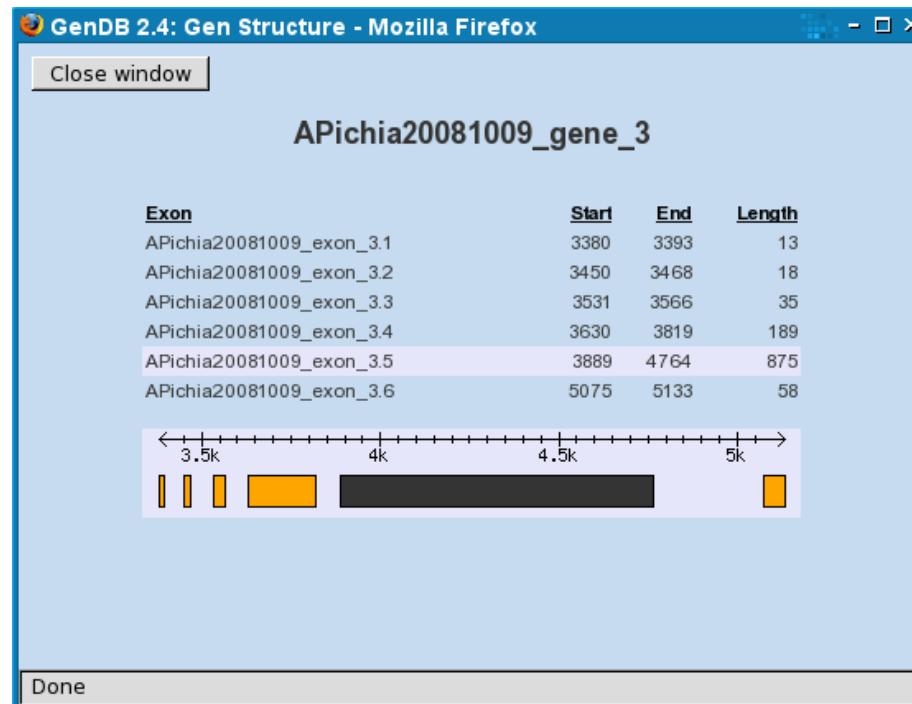
# ProMeTra: Visualization of –omics data sets



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 **CUDA** and **N- Exact algorithm, no heuristic
- Allows for deletions, insertions and substitutions
- Sample runtimes:**

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

## 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



bio  
data  
mining  
and  
analysis  
of  
metagenomes

A large, semi-transparent watermark on the right side of the slide contains the text "bio data mining and analysis of metagenomes" in a light gray, sans-serif font. The words are stacked vertically, with "bio" at the top, followed by "data", "mining", "and", "analysis", "of", "metagenomes".

## 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](mailto:agoesman@CeBiTec.Uni-Bielefeld.DE)
- Homepage: <http://www.cebitec.uni-bielefeld.de/brf>
- Software: <http://www.cebitec.uni-bielefeld.de/brf/software/brfsoftware.html>

C  
e  
B  
iT  
e  
c  
G

A large, semi-transparent watermark or background text is present on the right side of the slide, reading "CeBiTec" vertically, followed by "G".