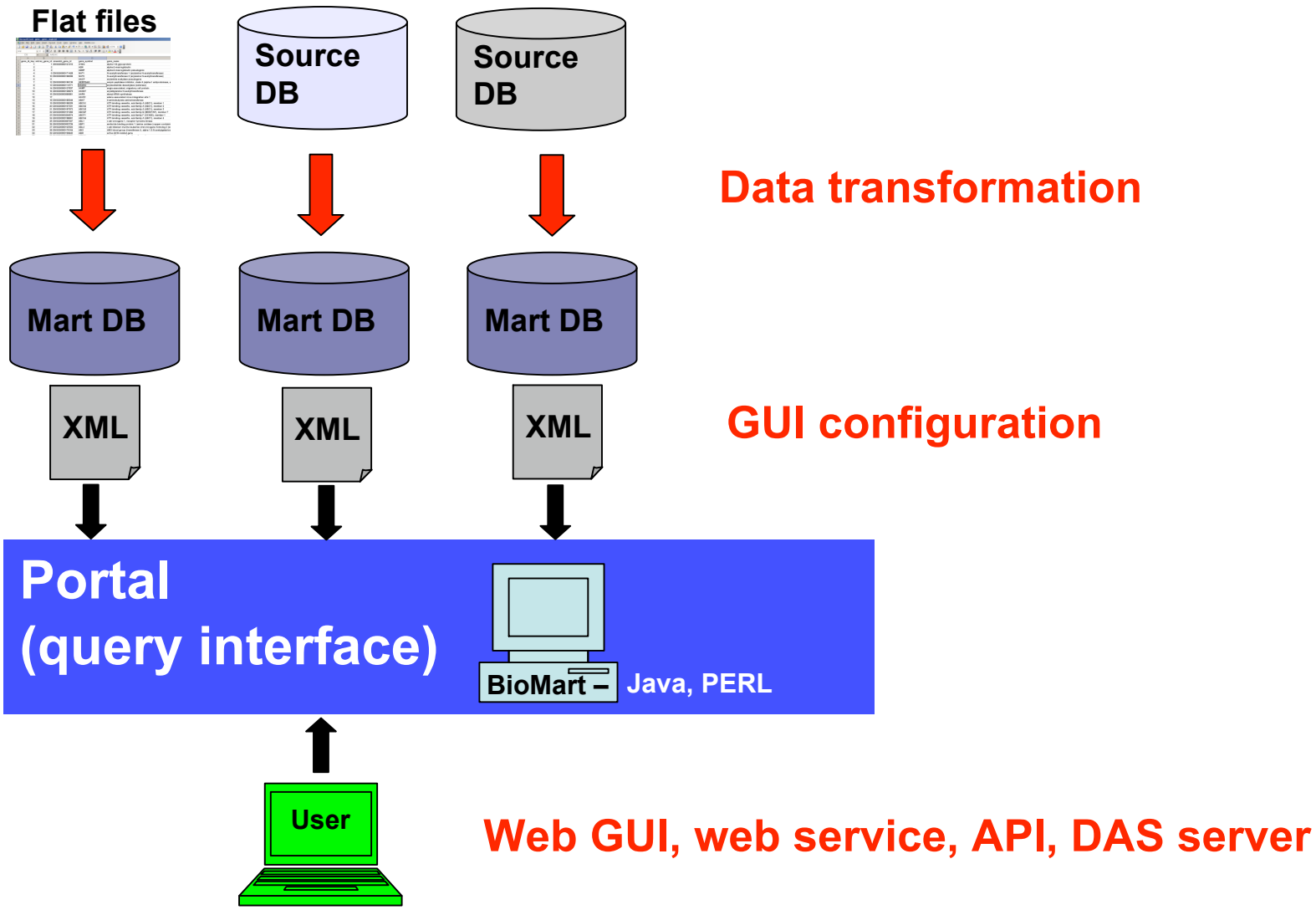
BioMart and GMOD
working towards a closer integration?

Arek Kasprzyk
Ontario Institute for Cancer Research
16th January 2009

BioMart in a nutshell

- open source data management system
  - Large scale datasets
  - Data federation
  - Query optimization
  - supports MySQL, Oracle and Postgres
- www.biomart.org

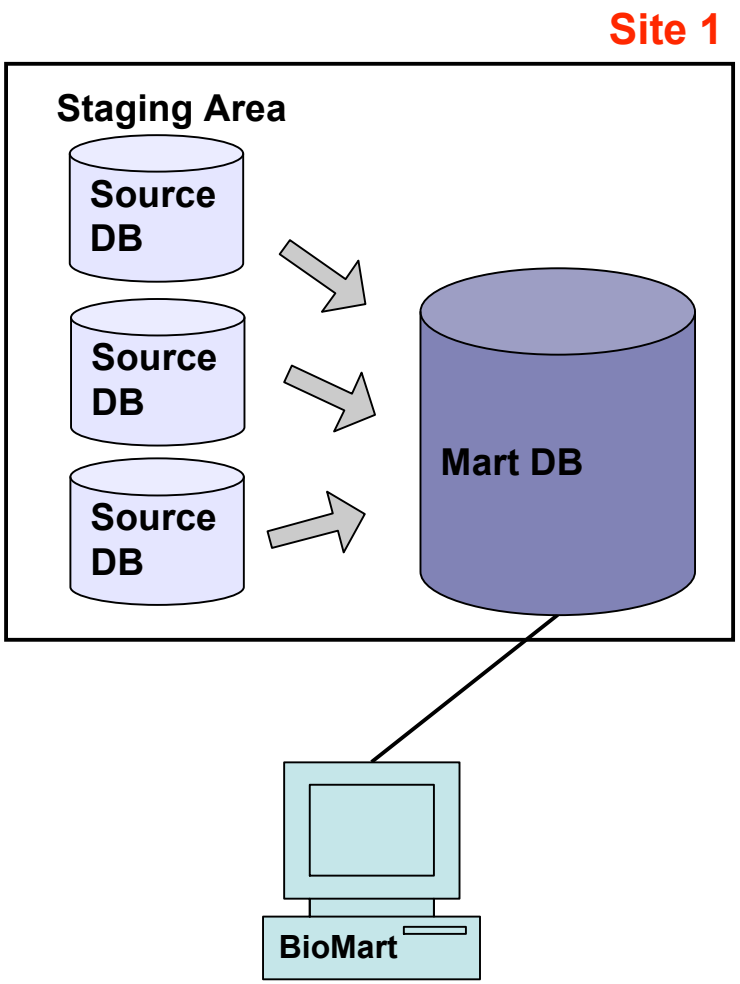
BioMart Work Flow



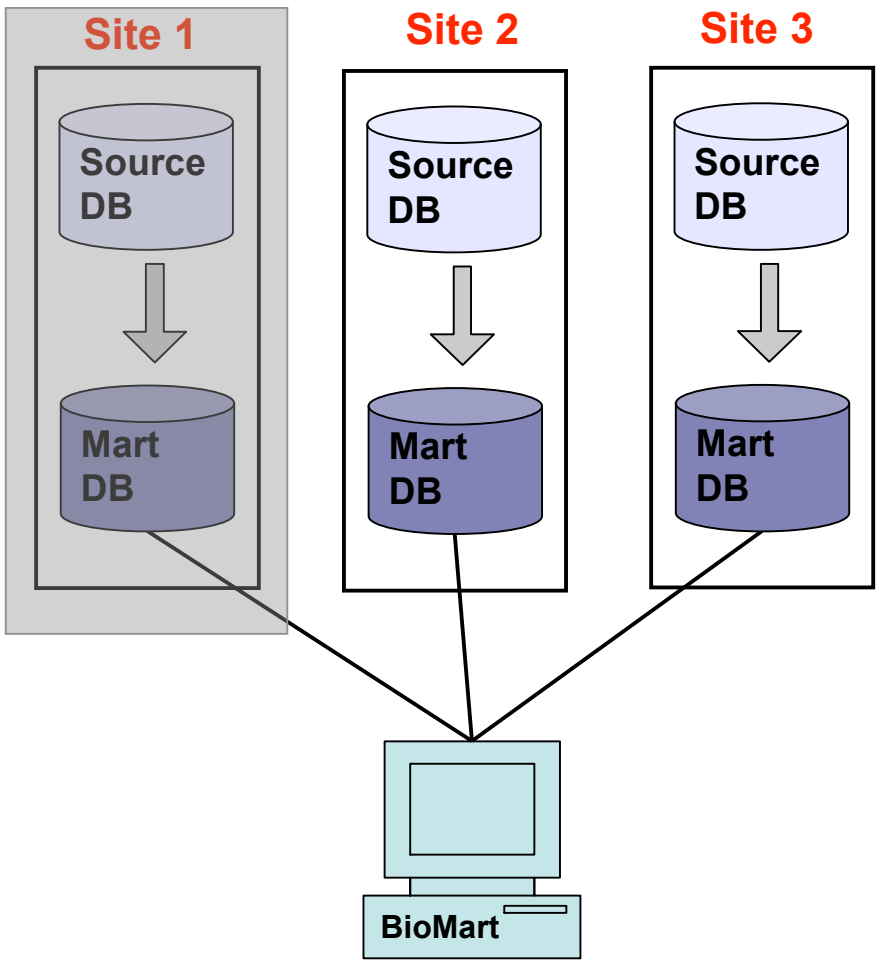
bio mart

Ontario Institute for Cancer Research
science → discoveries → solutions

Centralized Model



Federated Model



Publicly Available Marts

- Ensembl
- HapMap
- High Throughput Gene Targeting Group
- Dictybase
- Wormbase
- Gramene
- Europhenome
- Rat Genome Database
- EU Rat Mart
- ArrayExpress Data Warehouse
- Eurexpress
- DroSpeGe
- GermOnLine
- PRIDE
- PepSeeker
- VectorBase
- Pancreatic Expression Database
- Reactome
- Paramecium DB

www.biomart.org

Software with BioMart Plugin

- Bioclipse
- Bioconductor
- Cystoscape
- Galaxy
- Taverna
- WetLab

Single point of access
Single interface

Coming up …

- Uniprot
- HGNC
- Integr8
- IntAct
- Ensembl genomes
- EMMA
- I-DCC
- CREATE
- MGI
- Cancer Mart
- Mouse Informatics Portal
- International Cancer Genome Consortium (ICGC) Portal

bio mart

Ontario Institute for Cancer Research
science → discoveries → solutions

Ontario Institute for Cancer Research

Themes	Innovation Programs	Innovation Platforms				
Prevention	Ontario Cancer Cohort	Imaging and Interventions	Bio-repositories and Pathology	Genomics and High Throughput Screening	Medicinal Chemistry	Informatics and Bio-computing **Data Coordination Center (DCC)**
Early Diagnosis	One Millimetre Cancer Challenge					
Cancer Targets	Cancer Stem Cells					
	International Cancer Genome Consortium					
New Therapeutics	Selective Agents (Terry Fox Research Institute - Ontario Node)					
	Immuno- and Bio-therapeutics					
Translation Programs	Patents to Products					
	High Impact Clinical Trials					
	Cancer Care and Services (including Health Promotion)					

International Cancer Genome Consortium

Goals

- Catalogue genomic abnormalities in tumors in 50 different cancer types and/or subtypes of clinical and societal importance across the globe
- Generate complementary catalogues of transcriptomic and epigenomic datasets from the same tumors
- Make the data available to the entire research community as rapidly as possible and with minimal restrictions to accelerate research into the causes and control of cancer

50 different tumor types and/or subtypes

500 samples per tumor

25,000 Human Genome Projects!

Data Types

For each specimen

- pathology
- clinical history
- sequence variatns
- structural variants
- copy number variants
- gene expression
- splice variants
- epigenetic variants

Annotations

- gene ontologies
- pathways
- protein-protein interactions
- transcription factors
- other public and licensed annotation

Diverse data types

- images
- clinical notes and tests
- genomic data

bio mart

Ontario Institute for Cancer Research
science → discoveries → solutions

Current Members of ICGC

Country	Funding Organization	Tumor Type
Australia	National Health and Medical Research Council Announcement	Imminent
Canada	Ontario Institute for Cancer Research	Pancreas
China	Chinese Cancer Genome Consortium	Stomach
France	Institut National du Cancer	Liver (alcohol-related) Breast (HER2-positive)
India	Department of Biotechnology, Ministry of Science & Technology	Oral Cavity
Japan	RIKEN, National Cancer Center and National Institute of Biomedical Innovation	Liver (virus-related)
Spain	Spanish Ministry of Science and Innovation	Chroniclymphocytic leukemia
United Kingdom	The Wellcome Trust; Wellcome Trust Sanger Institute	Breast (several subtypes)

ICGC Data Coordination Centre

Mission

- implement project-wide standards for data completeness, quality and protection of confidentiality
- manage the collection and distribution of ICGC data
- manage an ICGC portal that provides researchers with project-wide data search and retrieval services

Challenges

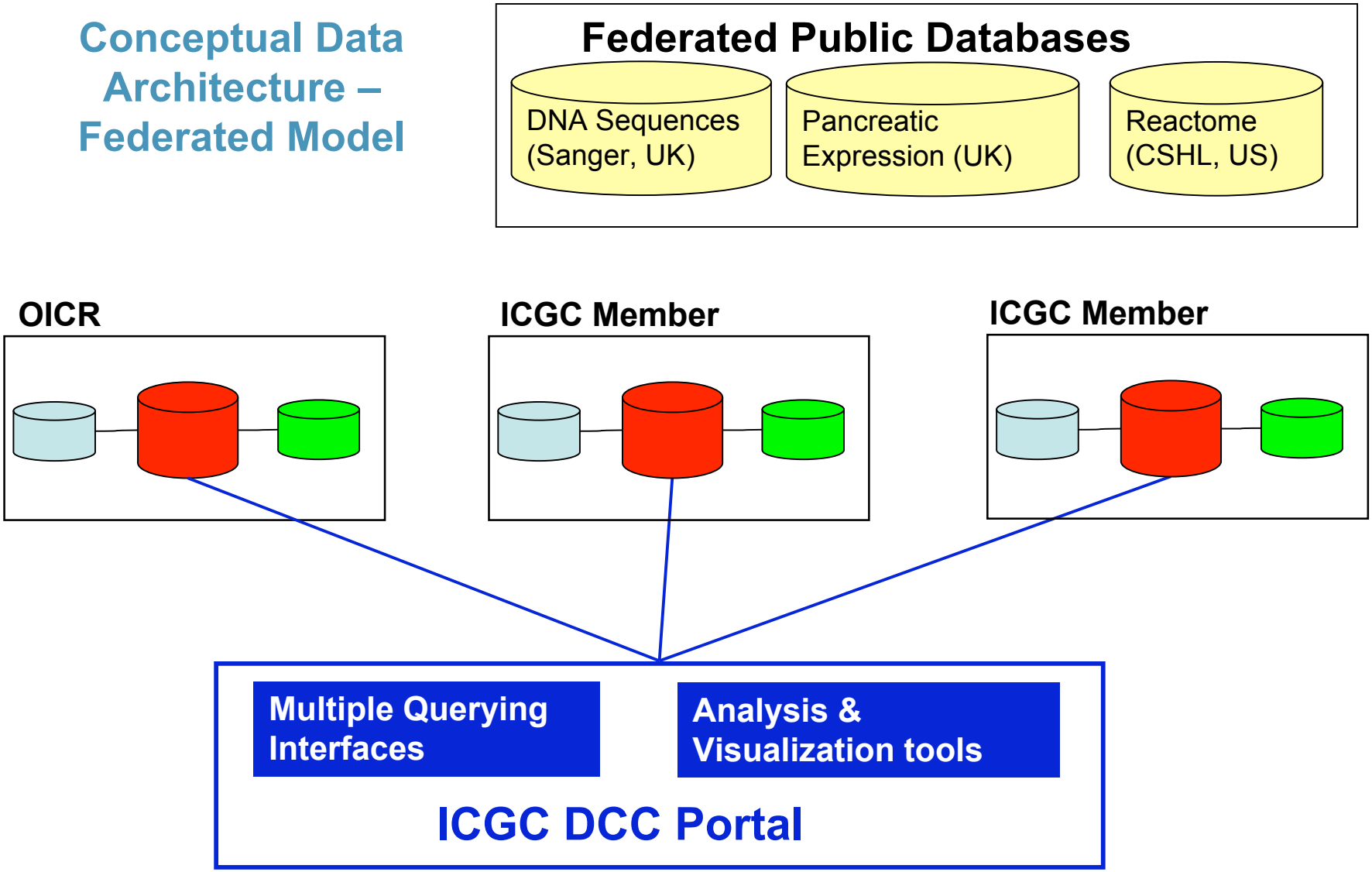
Data

- □ quantity
- □ content
- □ format
- □ location
- □ diversity

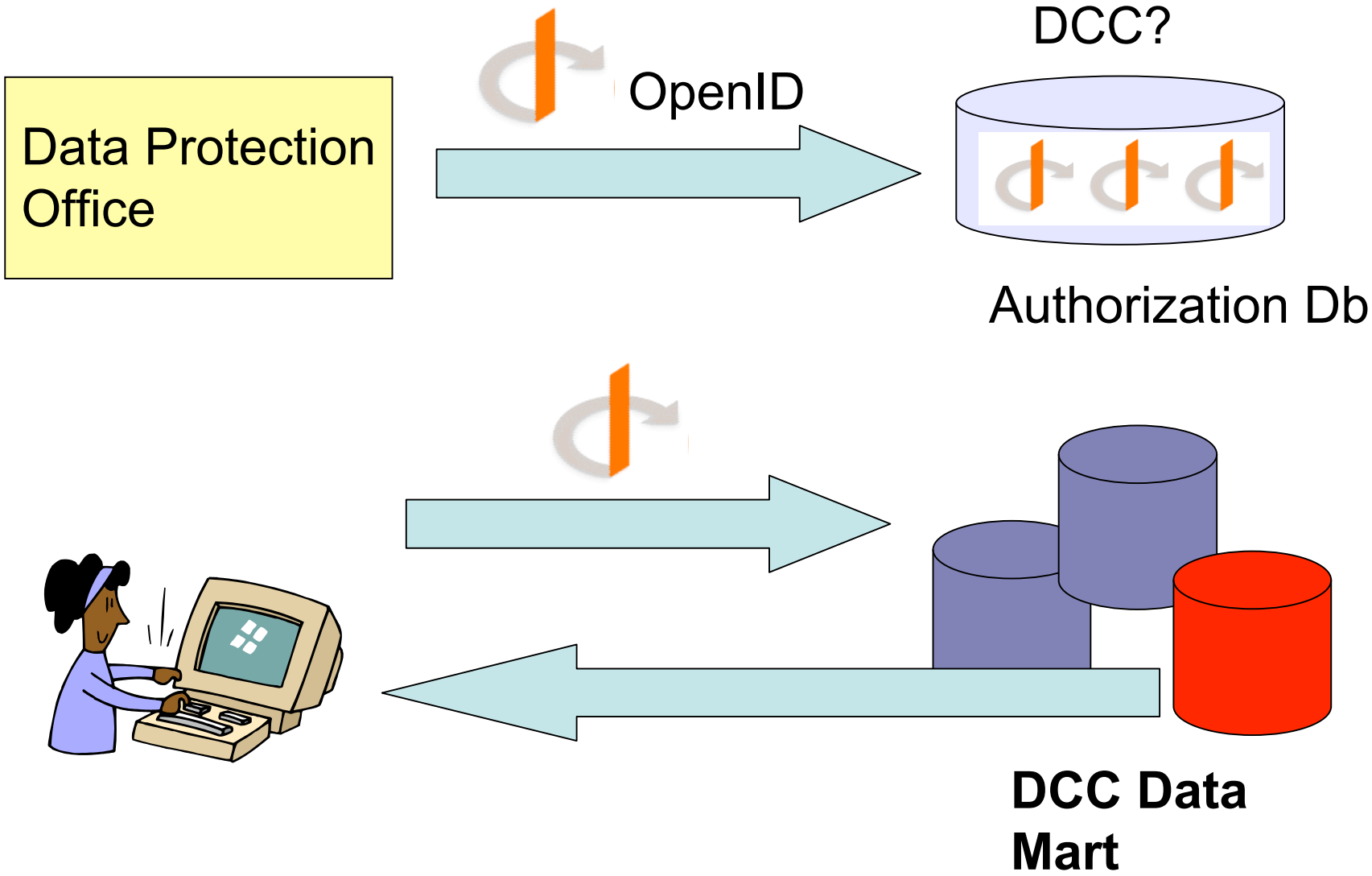
Database & Software

- □ design
- □ implementation
  - □ resources
- □ time scale
- □ synchronization with data

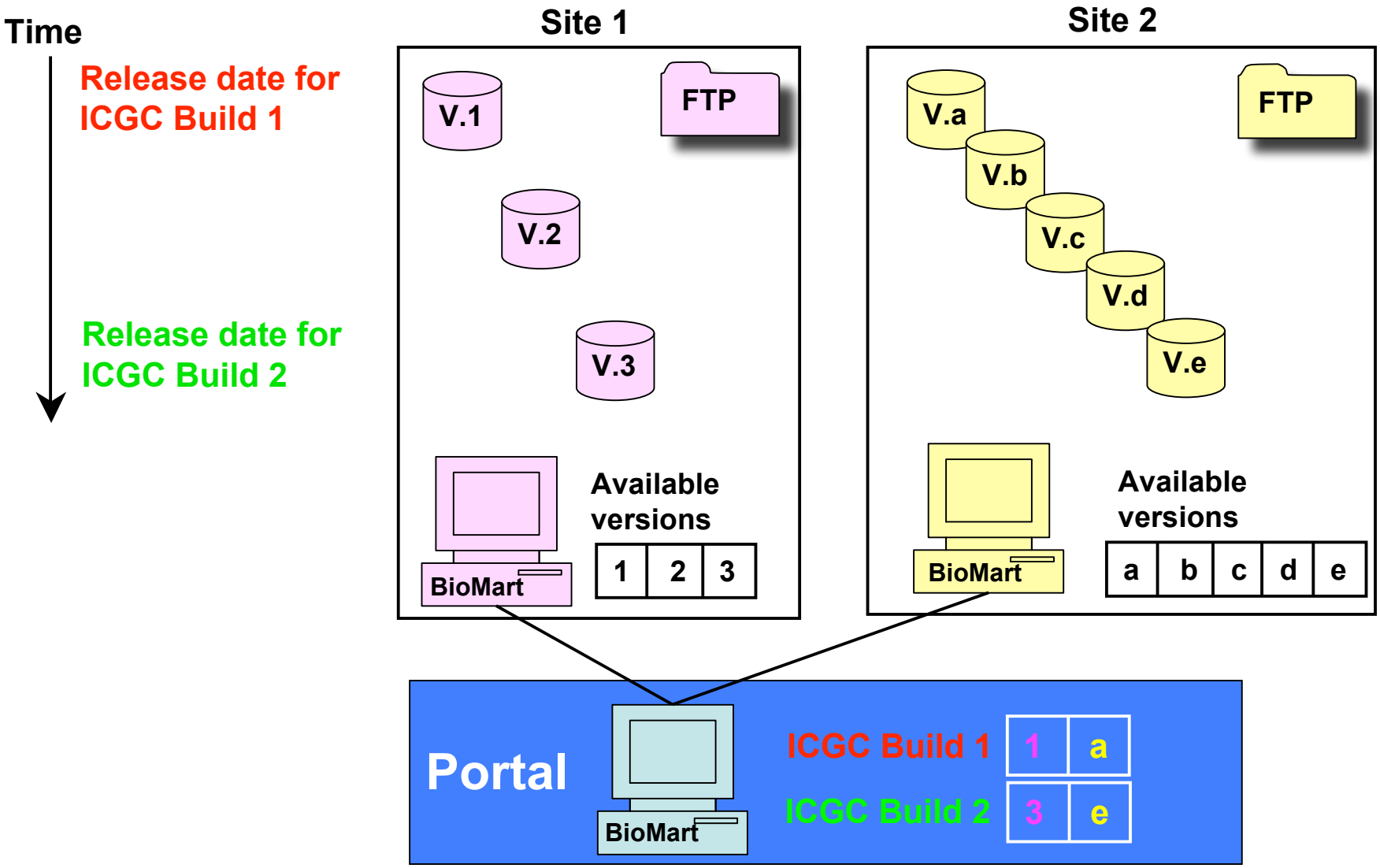
Conceptual Data Architecture – Federated Model



Controlled access data



Versioning of Data



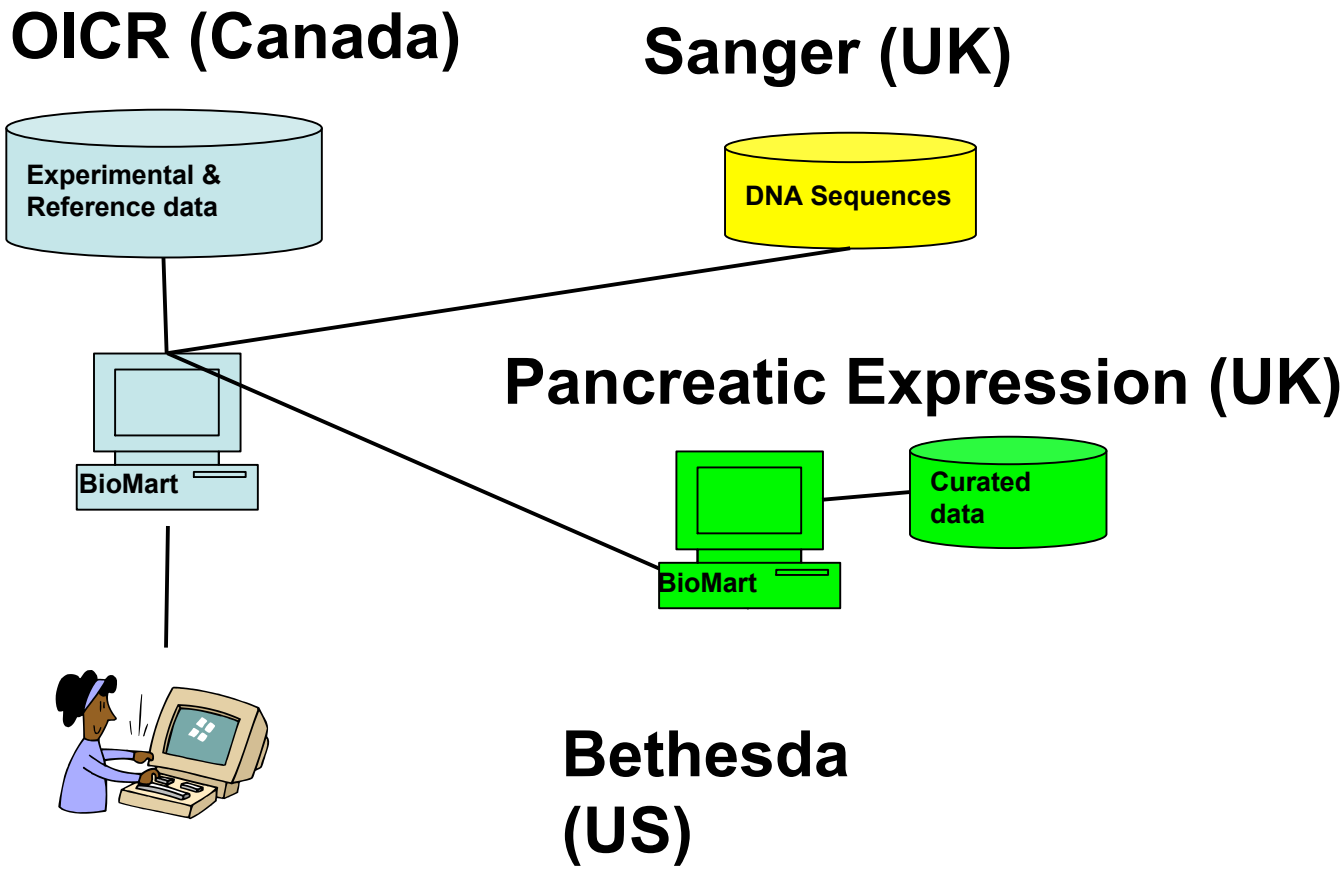
Demo: functionality

Biomarker discovery

In pancreatic tumor, which genes have copy number gain and up-regulated expression?

a. Retrieve DNA sequences 100bp upstream to these genes

b. Compare these fold changes with Pancreatic Expression Database



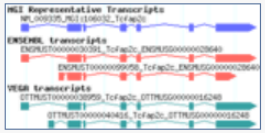
New GUI example

Mouse Informatics Project

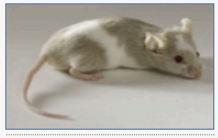
Download ▾ **More Resources** ▾ **Find Mice (EMMA)** **Contact Us**

Explore MIP — All Search Tools

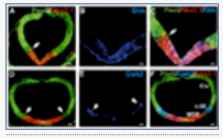
Genes



Phenotypes

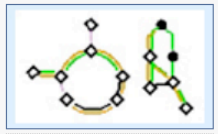


Expression



Function

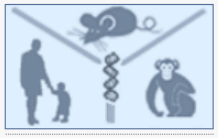
Pathways



Strains / SNPs

Variation Type	DBA/2J	FVB/NJ	129S1/SvImJ	Allele Summary (all strains)
SNP	G	G	A	A/G
SNP	C	C	T	C/T

Orthology



FAQs

How do I...

- .. search for genes? FAQ
- .. find mutations for phenotypes or diseases? FAQ
- .. find expression data? FAQ
- .. view a structural genomic map? FAQ

More FAQs

News 18 September, 2008

- ..... Read more...
- .... Read more...
- ....

More MIP news — MIP Statistics

Contributing Projects:
Mouse Genome Database (MGD), Gene Expression Database (GXD), Mouse Tumor Biology (MTB), Gene Ontology (GO), MouseCyc

Citing These Resources
Funding Information
Warranty Disclaimer & Copyright Notice
Send questions and comments to User Support.

last database update
10/21/2008
MIP_4.12
Web browser compatibility

biomart

Ontario Institute for Cancer Research
science → discoveries → solutions

Mouse Informatics Project

Home | Genes | Phenotypes | Expression | Function | Pathways | Strains / SNPs | Orthology

Search ▾ Download ▾ More Resources ▾ Find Mice (EMMA) Contact Us

Search for genes by name, location etc or modify data to return.

Search Reset

☐ Chromosome 1

☐ Base pair
Gene Start (bp) 1
Gene End (bp) 10000000

☐ ID list limit Ensembl Gene ID(s)

Browse...

Gene data to return

- ☑ Ensembl Gene ID
- ☑ Ensembl Transcript ID
- ☐ Ensembl Protein ID
- ☐ Description
- ☑ Chromosome Name
- ☑ Gene Start (bp)
- ☑ Gene End (bp)

⊞ **Additional gene search fields**

⊞ **Additional pathway search fields**

⊞ **Additional european mouse mutant archive search fields**

⊞ **Additional EUCOMM mouse KO project search fields**

⊞ **Additional Europhenome phenotype search fields**

⊞ **Additional Eurexpress expression search fields**

⊞ **Additional gene data to return**

⊞ **Additional pathway data to return**

⊞ **Additional european mouse mutant archive data to return**

⊞ **Additional EUCOMM mouse KO project data to return**

⊞ **Additional Europhenome phenotype data to return**

⊞ **Additional Eurexpress expression data to return**

Search Reset

Visualization: Gene List Analysis & Clinical Significance

Query

- Cancer types
- Gene annotation
- Chromosomal rearrangement (CR)
- Copy number variation (CNV)
- Single nucleotide polymorphism (SNP)
- Epigenetic modification (EM)
- Alternative splicing (AS)
- Transcript expression
- High throughput screen (HTS)

→ Gene List

Visualization

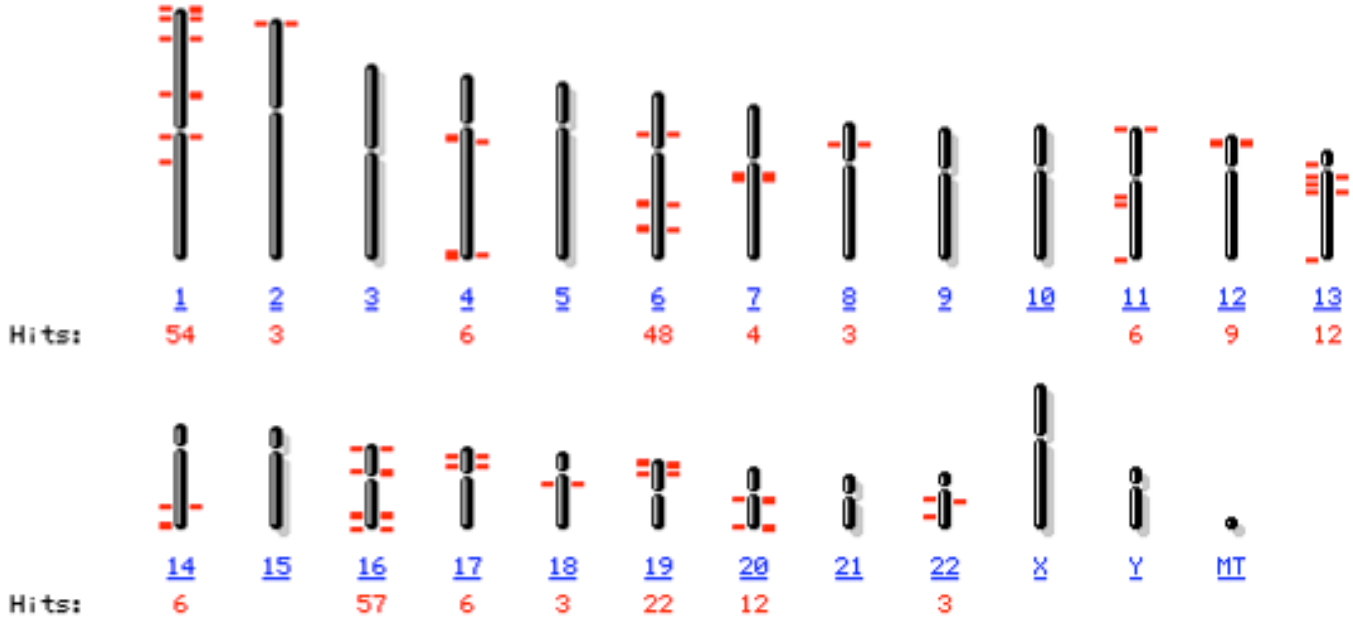
Gene list analysis

- **Map genes onto genome**
- **Map genes onto GO**
- **Map genes onto Pathways**

Clinical Significance

- **Event-free survival**
- **Response therapy**

Map Genes onto Genome



Hits shown: 1 - 100 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y MT

Visualization: Gene List Analysis & Clinical Significance

Query

- Cancer types
- Gene annotation
- Chromosomal rearrangement (CR)
- Copy number variation (CNV)
- Single nucleotide polymorphism (SNP)
- Epigenetic modification (EM)
- Alternative splicing (AS)
- Transcript expression
- High throughput screen (HTS)

→ Gene List

Visualization

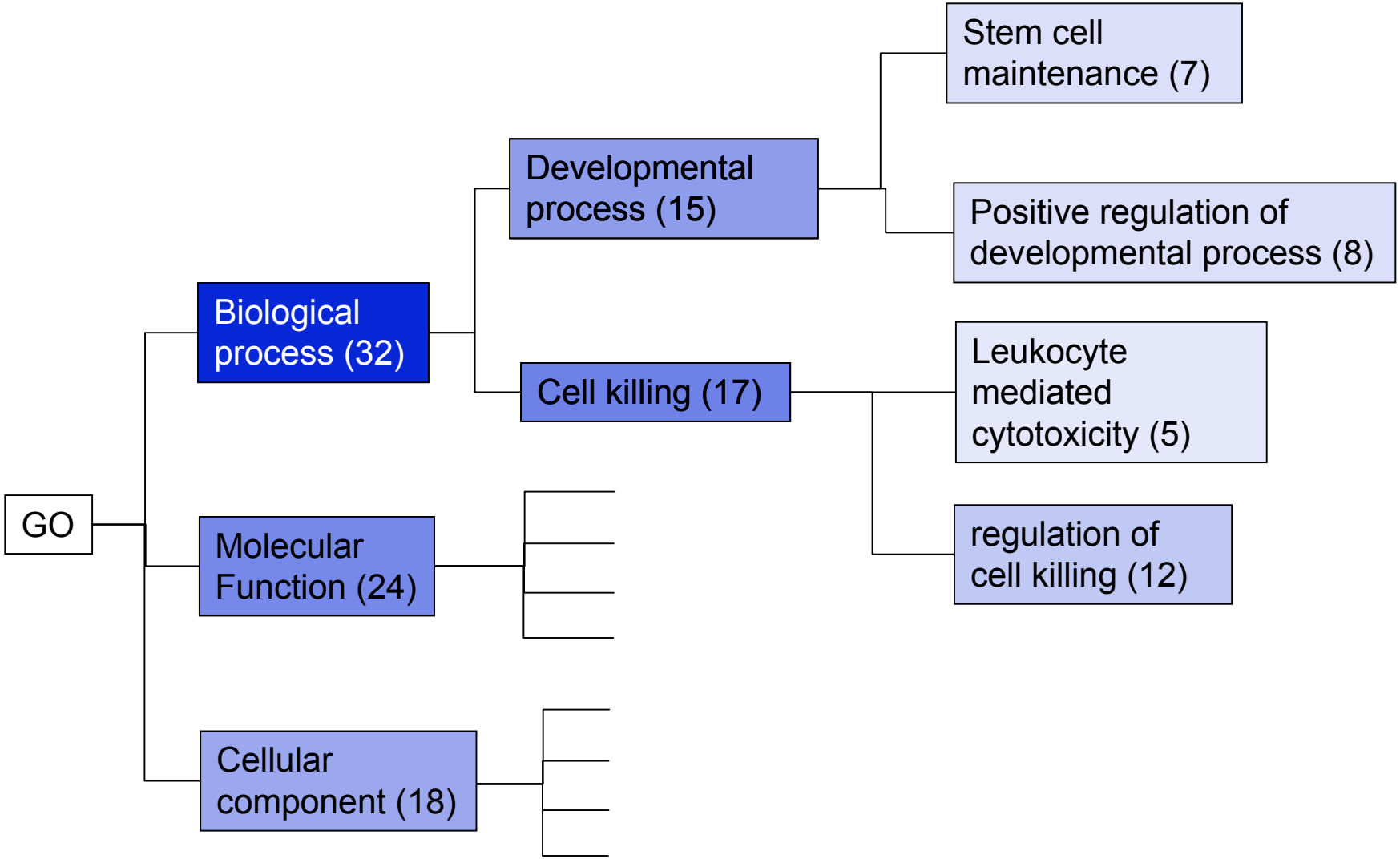
Gene list analysis

- **Map genes onto genome**
- **Map genes onto GO**
- **Map genes onto Pathways**

Clinical Significance

- **Event-free survival**
- **Response therapy**

Map Genes onto GO



Visualization: Gene List Analysis & Clinical Significance

Query

- Cancer types
- Gene annotation
- Chromosomal rearrangement (CR)
- Copy number variation (CNV)
- Single nucleotide polymorphism (SNP)
- Epigenetic modification (EM)
- Alternative splicing (AS)
- Transcript expression
- High throughput screen (HTS)

→ Gene List

Visualization

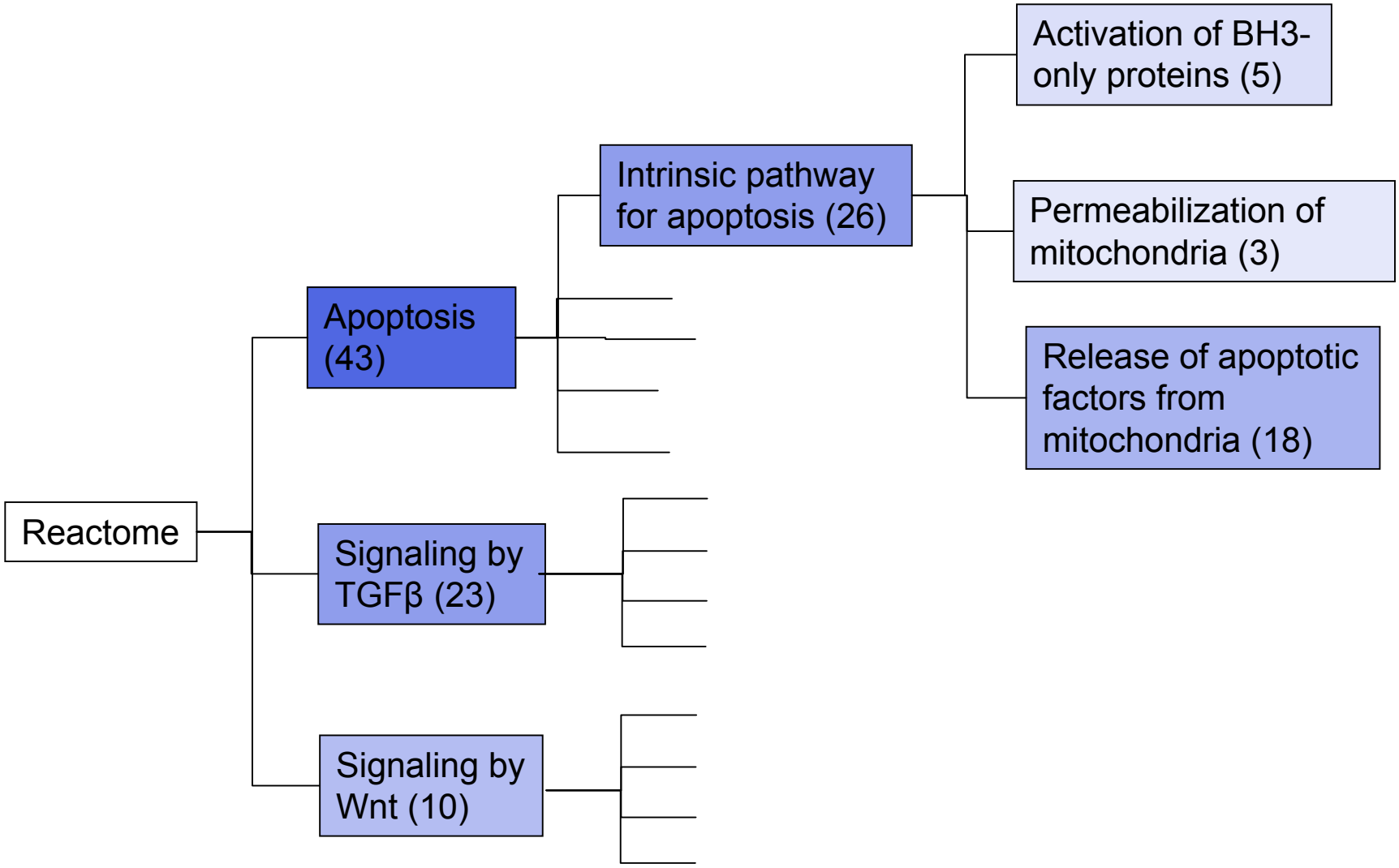
Gene list analysis

- **Map genes onto genome**
- **Map genes onto GO**
- **Map genes onto Pathways**

Clinical Significance

- **Event-free survival**
- **Response therapy**

Map Genes onto Pathways



Visualization: Gene List Analysis & Clinical Significance

Query

- Cancer types
- Gene annotation
- Chromosomal rearrangement (CR)
- Copy number variation (CNV)
- Single nucleotide polymorphism (SNP)
- Epigenetic modification (EM)
- Alternative splicing (AS)
- Transcript expression
- High throughput screen (HTS)

→ Gene List

Visualization

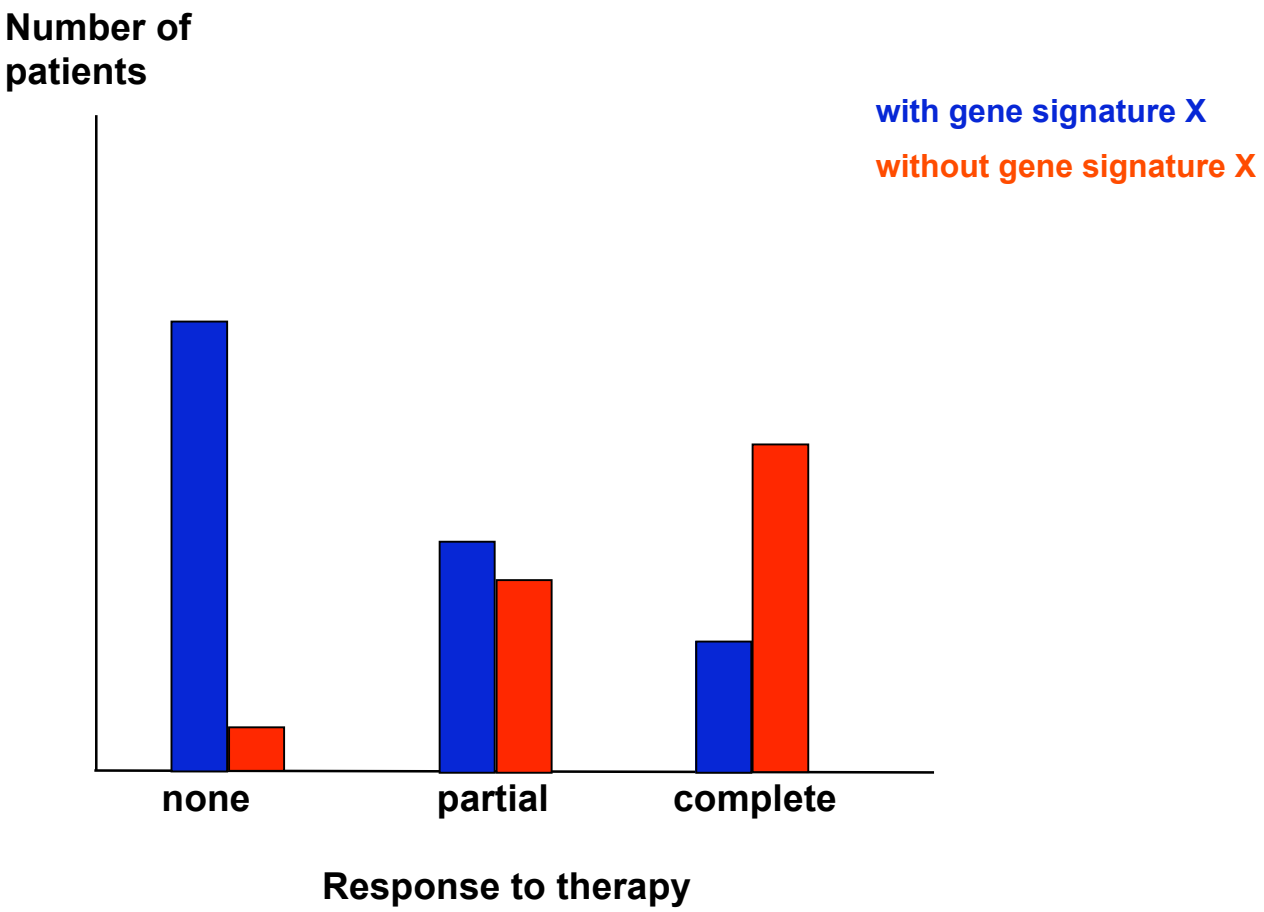
Gene list analysis

- Map genes onto genome
- Map genes onto GO
- Map genes onto Pathways

Clinical Significance

- Response to therapy
- Event – free survival

Stratify Patients' Response to Therapy by Gene Signature



Visualization: Gene List Analysis & Clinical Significance

Query

- Cancer types
- Gene annotation
- Chromosomal rearrangement (CR)
- Copy number variation (CNV)
- Single nucleotide polymorphism (SNP)
- Epigenetic modification (EM)
- Alternative splicing (AS)
- Transcript expression
- High throughput screen (HTS)

→ Gene List

Visualization

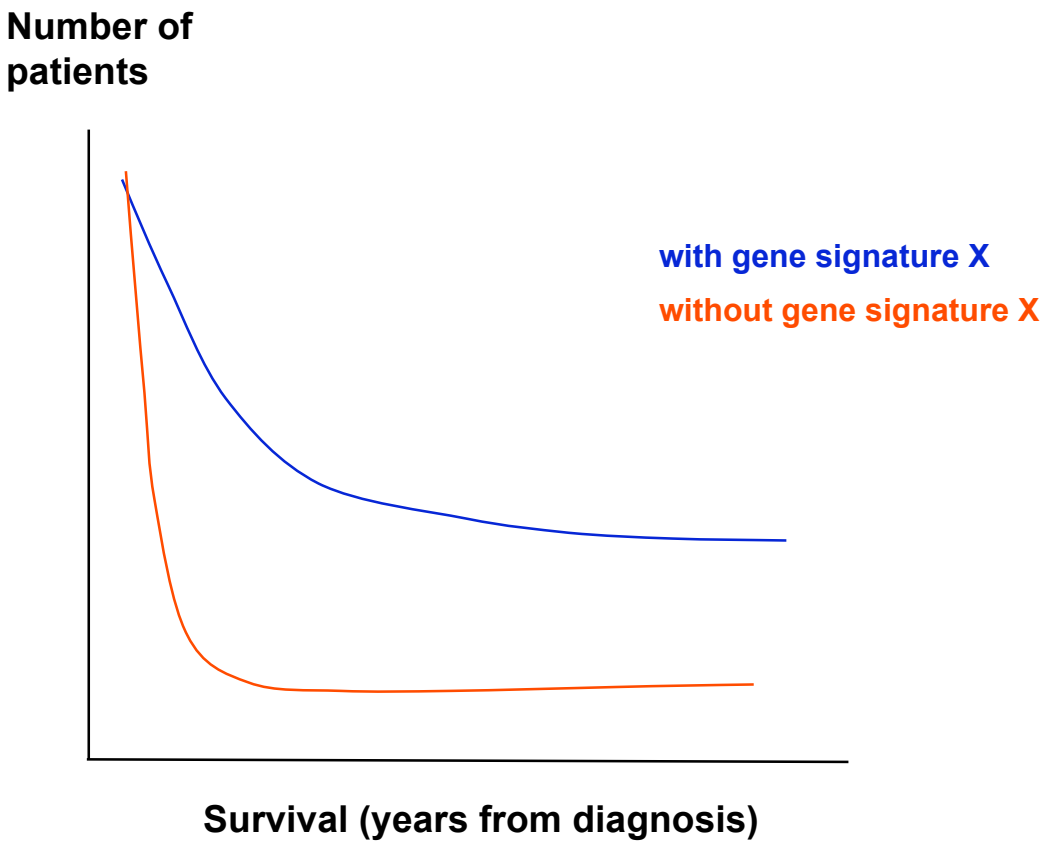
Gene list analysis

- **Map genes onto genome**
- **Map genes onto GO**
- **Map genes onto Pathways**

Clinical Significance

- **Response to therapy**
- **Event – free survival**

Stratify Patients' Survival by Gene Signature



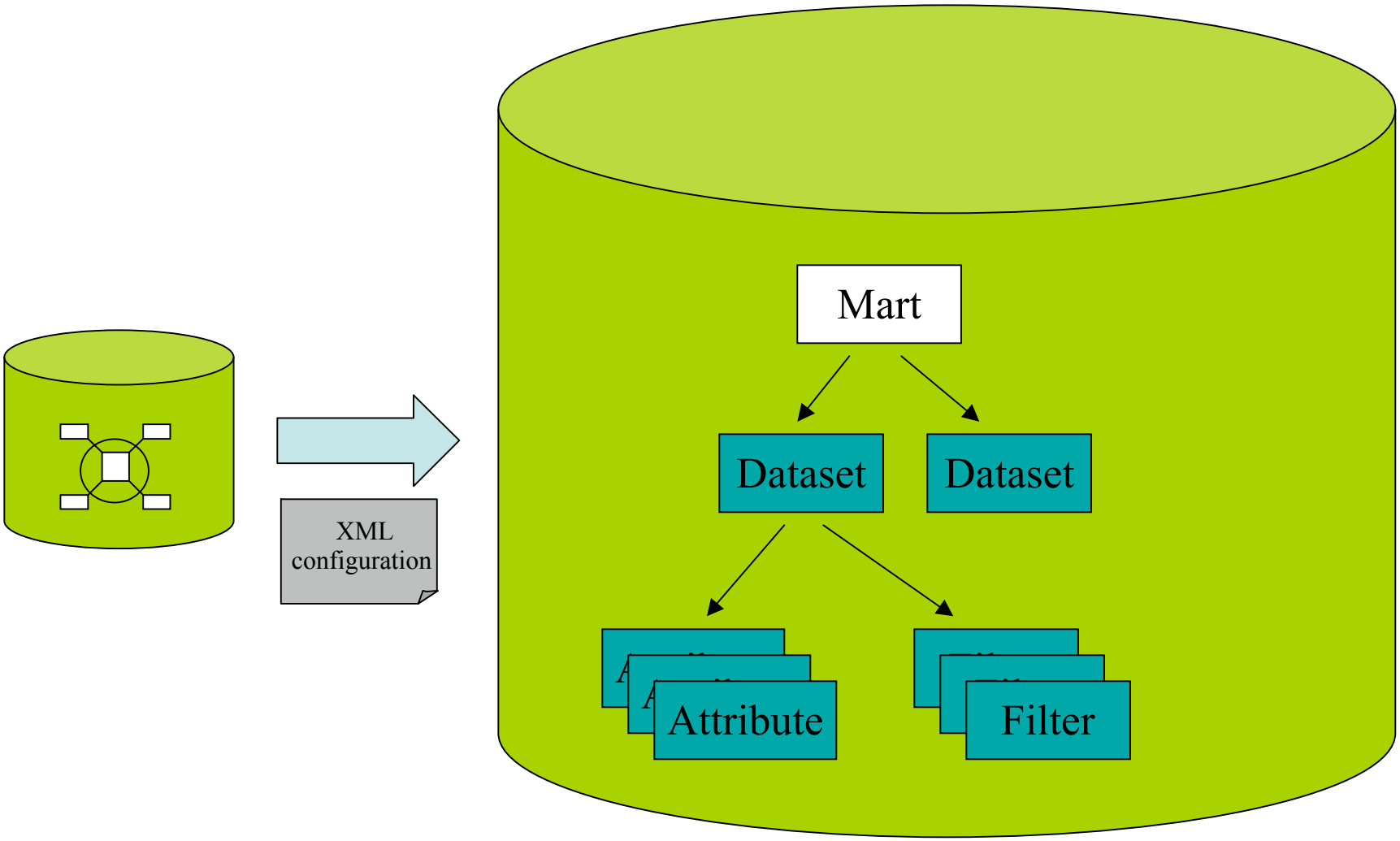
The plan - 0.8

- Full portal support
- New configuration system
  - Multiple GUI framework
  - Analysis and Visualization
  - Multi-tier secure data access
- Better federation support
- Better integration with third party tools

Part II

BioMart GMOD integration

Configuring BioMart Server



biomart

Ontario Institute for Cancer Research
science → discoveries → solutions

MartView (Web GUI)

biomart

HOME | MARTVIEW | MARTSERVICE | DOCS | CONTACT | NEWS | CREDITS

New | Count | Results | URL | XML | Perl | Help

Dataset 741 / 37435 Genes
Homo sapiens genes (NCBI36)

Filters

Chromosome: 1
Band Start : p36.33
Band End : p35.2

Attributes

Ensembl Gene ID
Ensembl Transcript ID
Chromosome Name
Band

Dataset
MSD protein structures

Filters

Experiment type : NMR

Attributes

PDB ID(s)
Release Date
Experiment Type

Export all results to File TSV ☐ Unique results only Go

Email notification to

View 10 rows as HTML ☐ Unique results only

Ensembl Gene ID	Ensembl Transcript ID	Chromosome Name	Band	PDB ID(s)	Release Date	Experiment Type
ENSG00000070831	ENST00000315554	1	p36.12	1aje	1997-11-12	NMR
ENSG00000078900	ENST00000378295	1	p36.32	1cok	1999-08-17	NMR
ENSG00000070831	ENST00000315554	1	p36.12	1ees	2000-04-12	NMR
ENSG00000160049	ENST00000377038	1	p36.22	1koy	2002-09-18	NMR
ENSG00000121769	ENST00000373713	1	p35.2	1g5w	2001-03-07	NMR
ENSG00000070831	ENST00000315554	1	p36.12	1e0a	2000-09-14	NMR
ENSG00000117748	ENST00000373912	1	p35.3	1dpu	2000-11-10	NMR
ENSG00000169598	ENST00000378209	1	p36.32	1ibx	2001-05-30	NMR
ENSG00000160049	ENST00000377038	1	p36.22	1ibx	2001-05-30	NMR
ENSG00000070831	ENST00000315554	1	p36.12	1cee	1999-06-30	NMR

biomart version 0.7

Perl API

```
my $initializer = BioMart::Initializer->new('registryFile'=>$confFile);
my $registry = $initializer->getRegistry;
my $query = BioMart::Query-
>new('registry'=>$registry,'virtualSchemaName'=>'central_server_1');

 $query->setDataset("hsapiens_gene_ensembl");
            $query->addFilter("band_start", ["p36.33"]);
            $query->addFilter("chromosome_name", ["1"]);
            $query->addFilter("band_end", ["p35.2"]);
            $query->addAttribute("ensembl_gene_id");
            $query->addAttribute("ensembl_transcript_id");
            $query->addAttribute("chromosome_name");
            $query->addAttribute("band");

$query->setDataset("msd");
            $query->addFilter("experiment_type", ["NMR"]);
            $query->addAttribute("pdb_id");
            $query->addAttribute("release_date");
            $query->addAttribute("experiment_type");

 my $query_runner = BioMart::QueryRunner->new();
 $query_runner->execute($query);
 $query_runner->printResults();
```

MartService (REST)

```
<Query  virtualSchemaName = "default" formatter = "TSV" datasetConfigVersion = "0.5" >

        <Dataset name = "hsapiens_gene_ensembl" interface = "default" >
                <Filter name = "band_start" value = "p36.33"/>
                <Filter name = "chromosome_name" value = "1"/>
                <Filter name = "band_end" value = "p35.2"/>
                <Attribute name = "ensembl_gene_id" />
                <Attribute name = "ensembl_transcript_id" />
                <Attribute name = "chromosome_name" />
                <Attribute name = "band" />
        </Dataset>

        <Dataset name = "msd" interface = "default" >
                <Filter name = "experiment_type" value = "NMR"/>
                <Attribute name = "pdb_id" />
                <Attribute name = "release_date" />
                <Attribute name = "experiment_type" />
        </Dataset>
</Query>
```

MartService (SOAP)


```
<soapenv:Envelope xmlns:soapenv="http://schemas.xmlsoap.org/soap/envelope/"
xmlns:mar="http://www.biomart.org:80/MartServiceSoap">
  <soapenv:Header/>
  <soapenv:Body>
    <mar:Query>
      <virtualSchemaName>default</virtualSchemaName>
      <header>0</header>
      <count>0</count>
      <uniqueRows>0</uniqueRows>
      <Dataset>
        <name>hsapiens_gene_ensembl</name>
        <Filter><name>chromosome_name</name><value>1</value></Filter>
        <Filter><name>band_start</name><value>p36.33</value></Filter>
        <Filter><name>band_end</name><value>p35.2</value></Filter>
        <Attribute><name>ensembl_gene_id</name></Attribute>
        <Attribute><name>ensembl_transcript_id</name></Attribute>
        <Attribute><name>chromosome_name</name></Attribute>
        <Attribute><name>band</name></Attribute>
      </Dataset>
      <Dataset>
        <name>msd</name>
        <Filter><name>experiment_type</name><value>NMR</value></Filter>
        <Attribute><name>pdb_id</name></Attribute>
        <Attribute><name>release_date</name></Attribute>
        <Attribute><name>experiment_type</name></Attribute>
      </Dataset>
    </mar:Query>
  </soapenv:Body>
</soapenv:Envelope>
```

URL-based access

- **Pre-defined queries (links shown in webpage)**
- **Bookmark / referencing**
- **An example:**
  http://www.biomart.org/biomart/martview?VIRTUALSCHEMANAME=default&ATTRIBUTES=***hsapiens_gene_ensembl***.default.feature_page.ensembl_gene_id|hsapiens_gene_ensembl.default.feature_page.ensembl_transcript_id|hsapiens_gene_ensembl.default.feature_page.chromosome_name|hsapiens_gene_ensembl.default.feature_page.band|***msd***.default.feature_page.pdb_id|msd.default.feature_page.release_date|msd.default.feature_page.experiment_type&FILTERS=hsapiens_gene_ensembl.default.filters.band_start."p36.33"|hsapiens_gene_ensembl.default.filters.chromosome_name."1"|hsapiens_gene_ensembl.default.filters.band_end."p35.2"|msd.default.filters.experiment_type."NMR"&VISIBLEPANEL=resultspanel

Major GMOD components

Community Annotation
- Apollo
- Wiki Table Editor

Comparative Genome Visualization
- CMap
- GBrowse_syn
- SynView
- SynBrowse
- Sybil

Database schema
- Chado

Database tools
- Argos
- BioMart
- Genome grid
- GMODTools
- LuceGene
- XORT
- InterMine

Gene Expression Visualization
- Caryoscope
- GeneXplorer
- Java TreeView

Genome Annotation
- Apollo
- MAKER

Genome Visualization & Editing
- Apollo
- Flash GViewer
- GBrowse
- GMODWeb
- Restriction Graphic Viewer

Literature Tools
- PubSearch
- Textpresso

Molecular Pathway Visualization
- Pathway Tools

Ontology Visualization
- Go Graphic Viewer

Workflow Management
- Ergatis

Middleware
- Modware
- Chado::AutoDBI

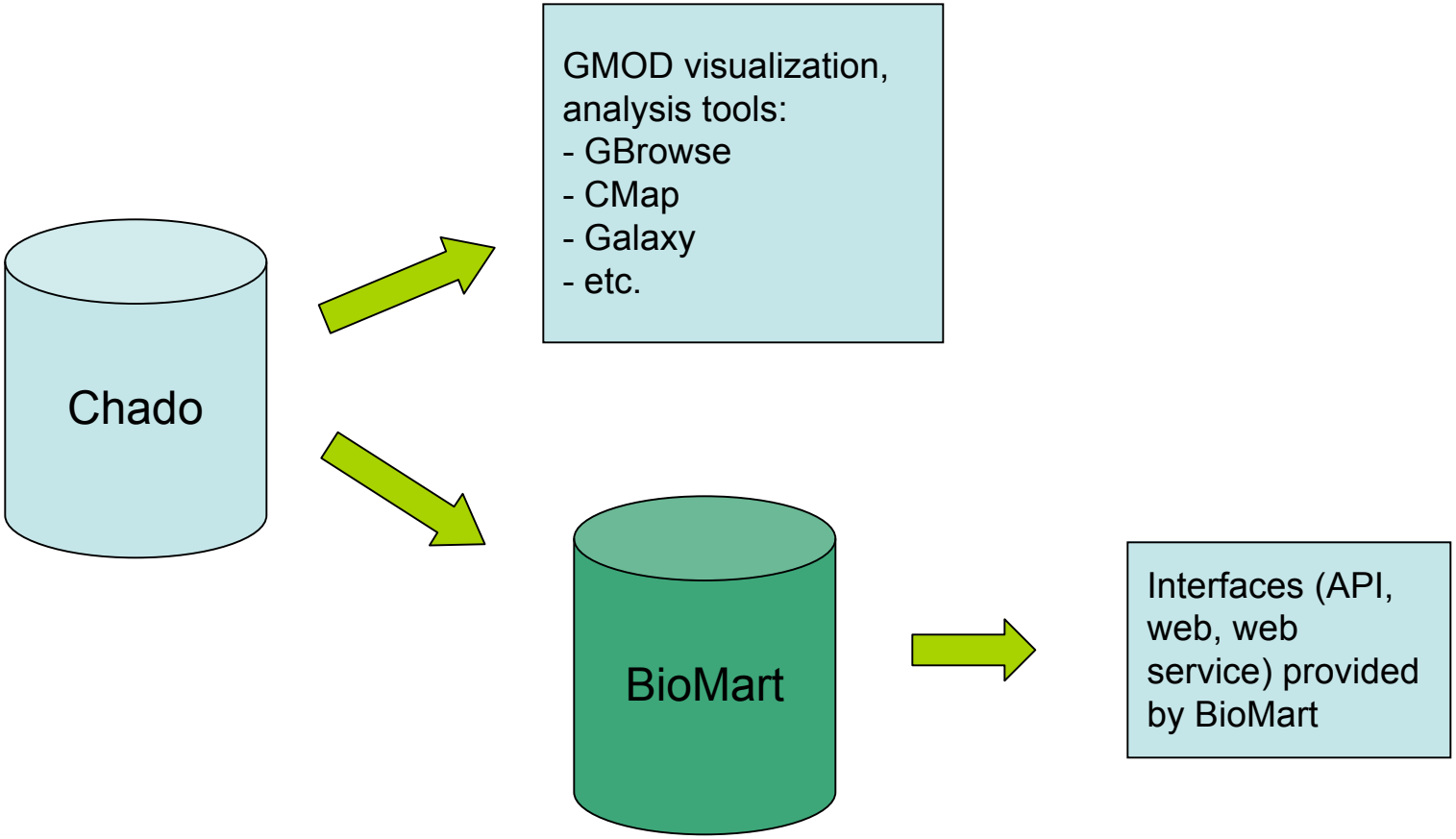
Tool Integration
- Galaxy

Sequence Alignment
- Blast Graphic

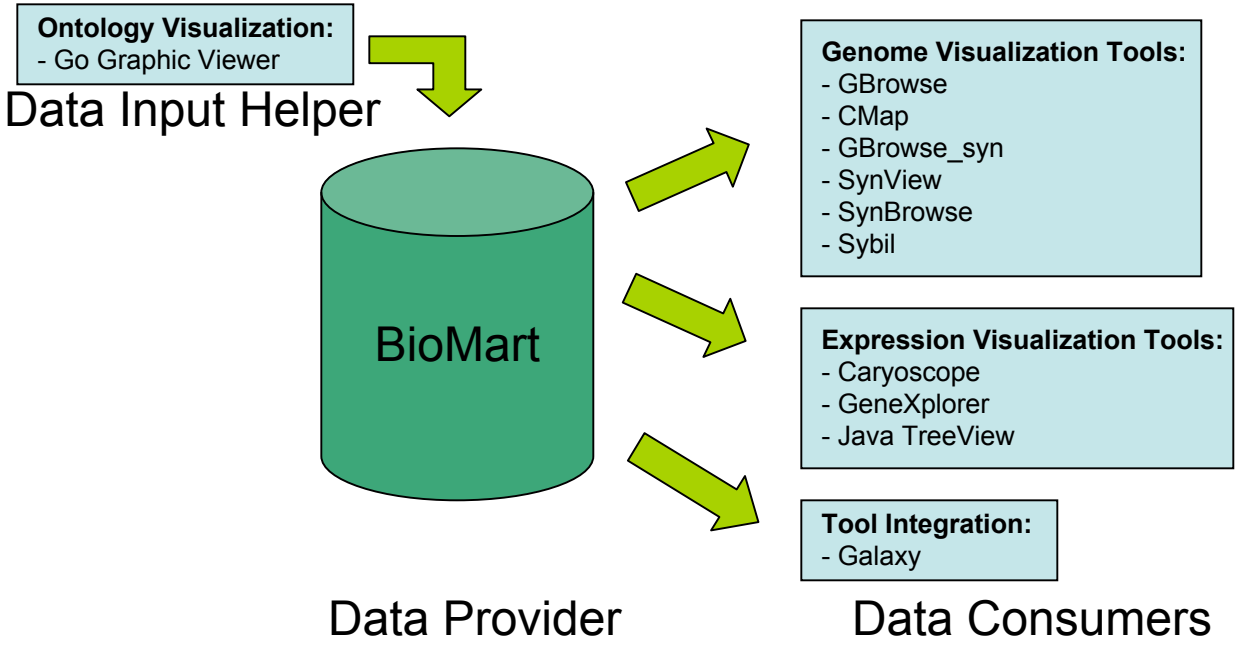
Utilities
- org.bdgp

(gmod.org/wiki/Overview)

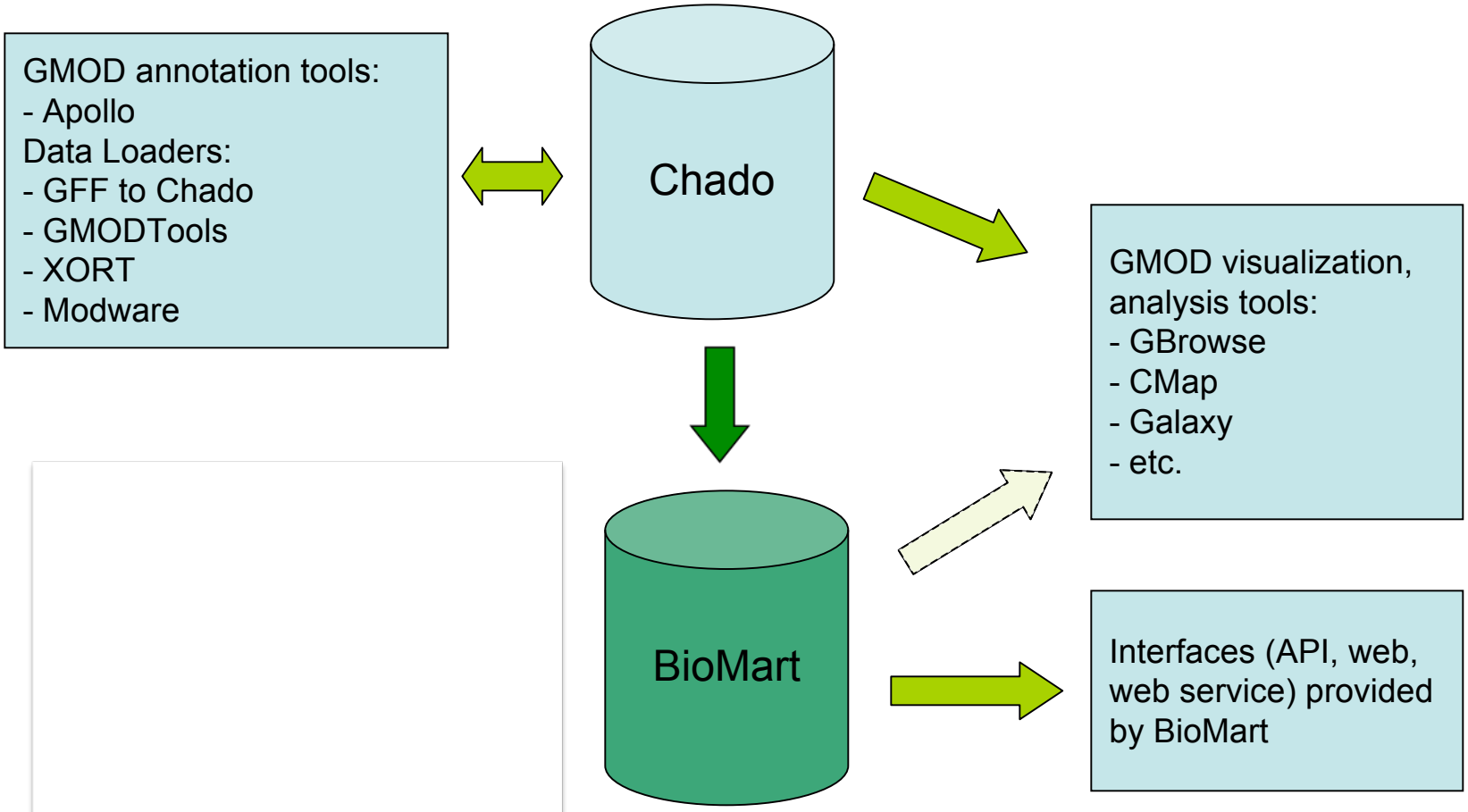
GMOD and BioMart



Possible interactions between BioMart server and other GMOD components



BioMart in addition to Chado as a data source



BioMart Data sources -> GMOD

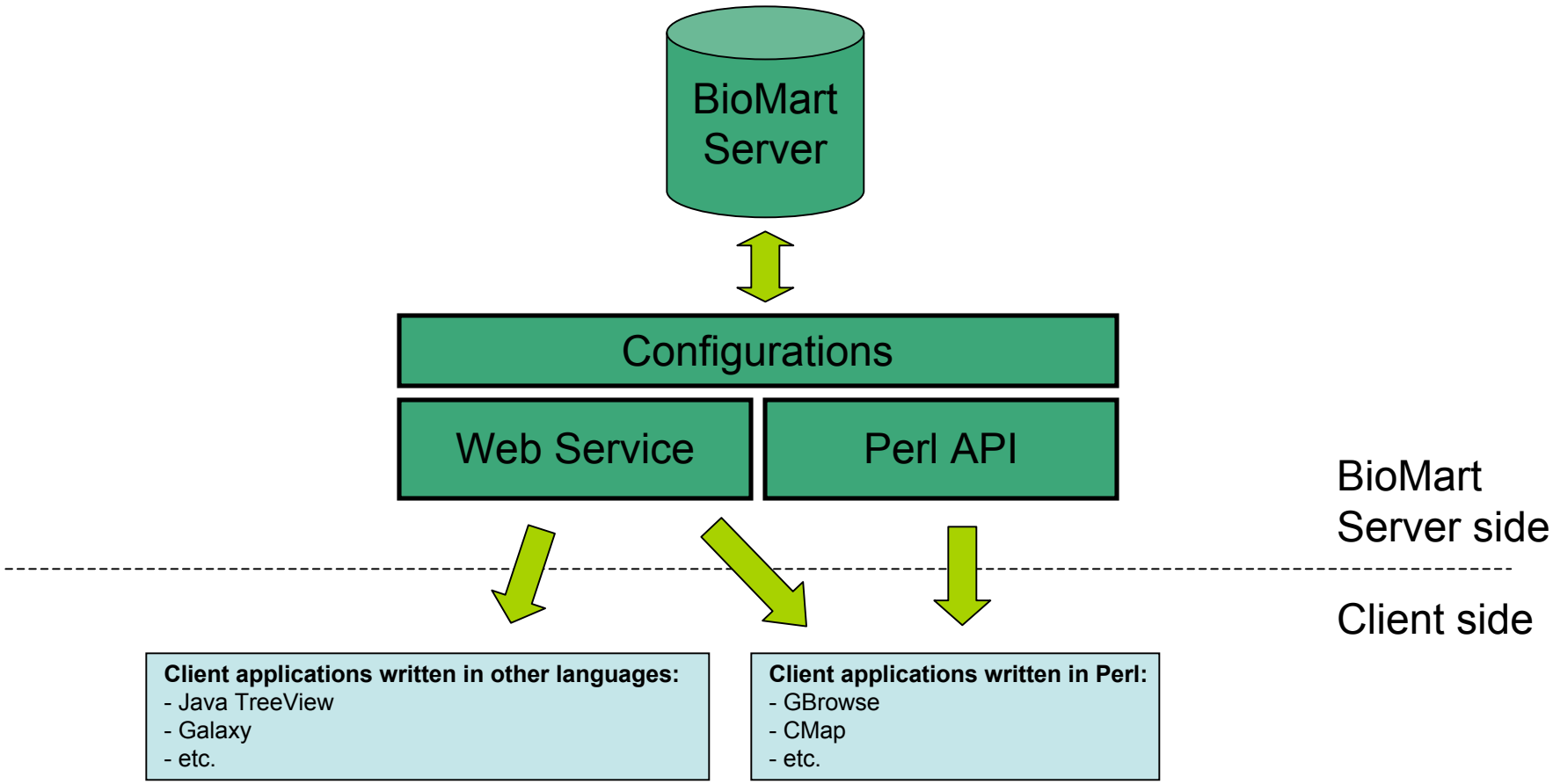
- Ensembl
- HapMap
- High Throughput Gene Targeting Group
- Dictybase
- Wormbase
- Gramene
- Europhenome
- Rat Genome Database
- EU Rat Mart

- ArrayExpress Data Warehouse
- Eurexpress
- DroSpeGe
- GermOnLine
- PRIDE
- PepSeeker
- VectorBase
- Pancreatic Expression Database
- Reactome
- Paramecium DB

- Uniprot
- HGNC
- Integr8
- IntAct
- Ensembl genomes
- EMMA

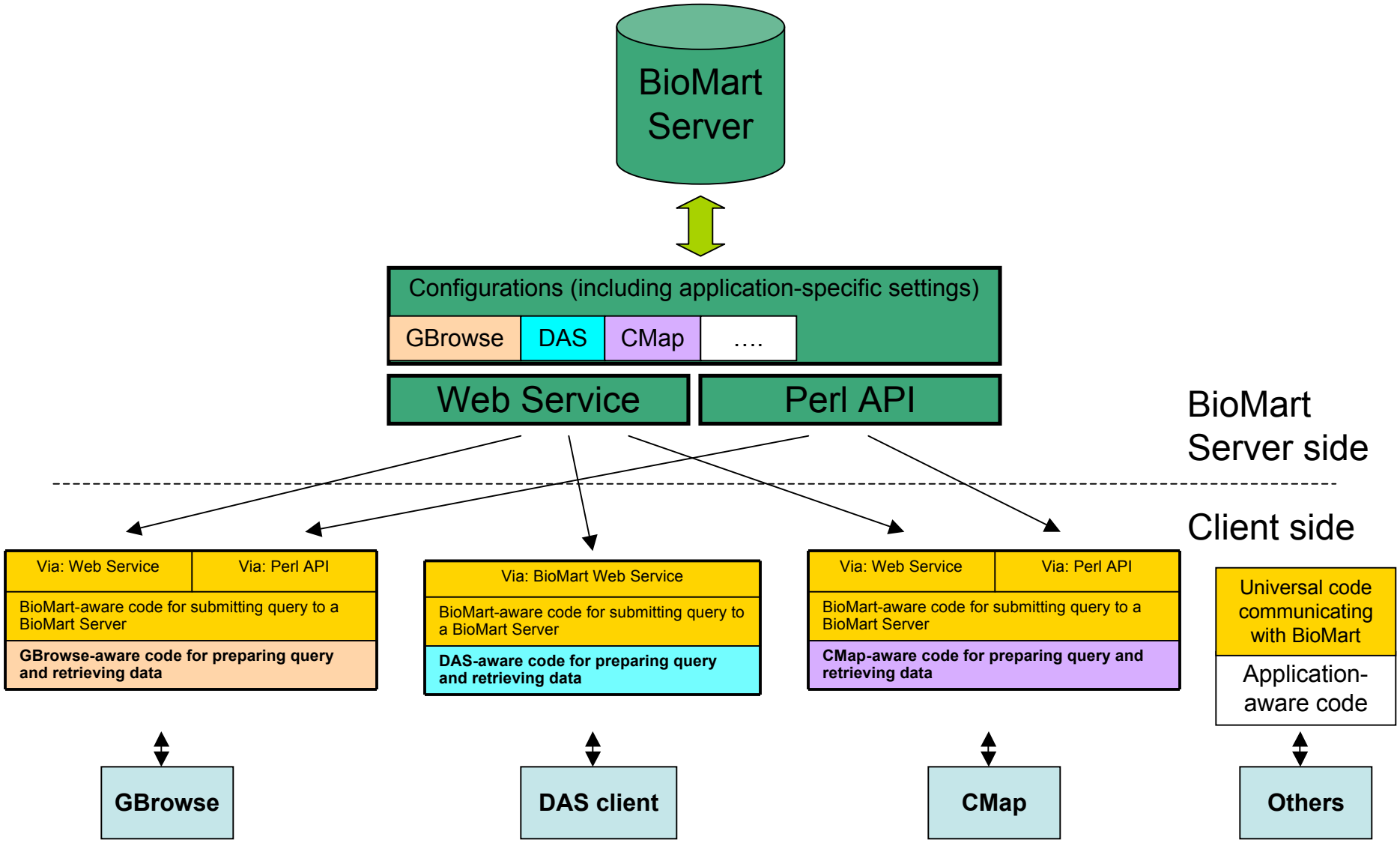
- I-DCC
- CREATE
- MGI
- Cancer Mart
- Mouse Informatics Portal
- ICGC Portal

GMOD/BioMart integration: architecture overview

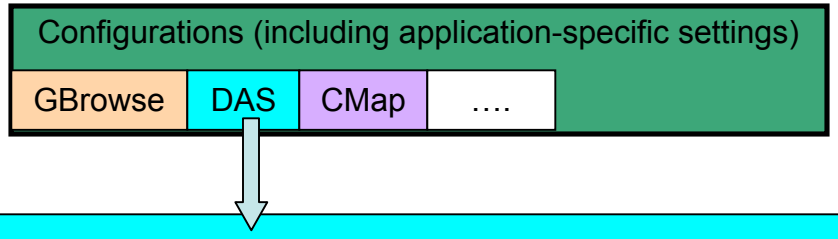


For each client, an application-specific adaptor is needed to communicate with BioMart server

GMOD/BioMart integration: examples



Server side application-specific configuration example

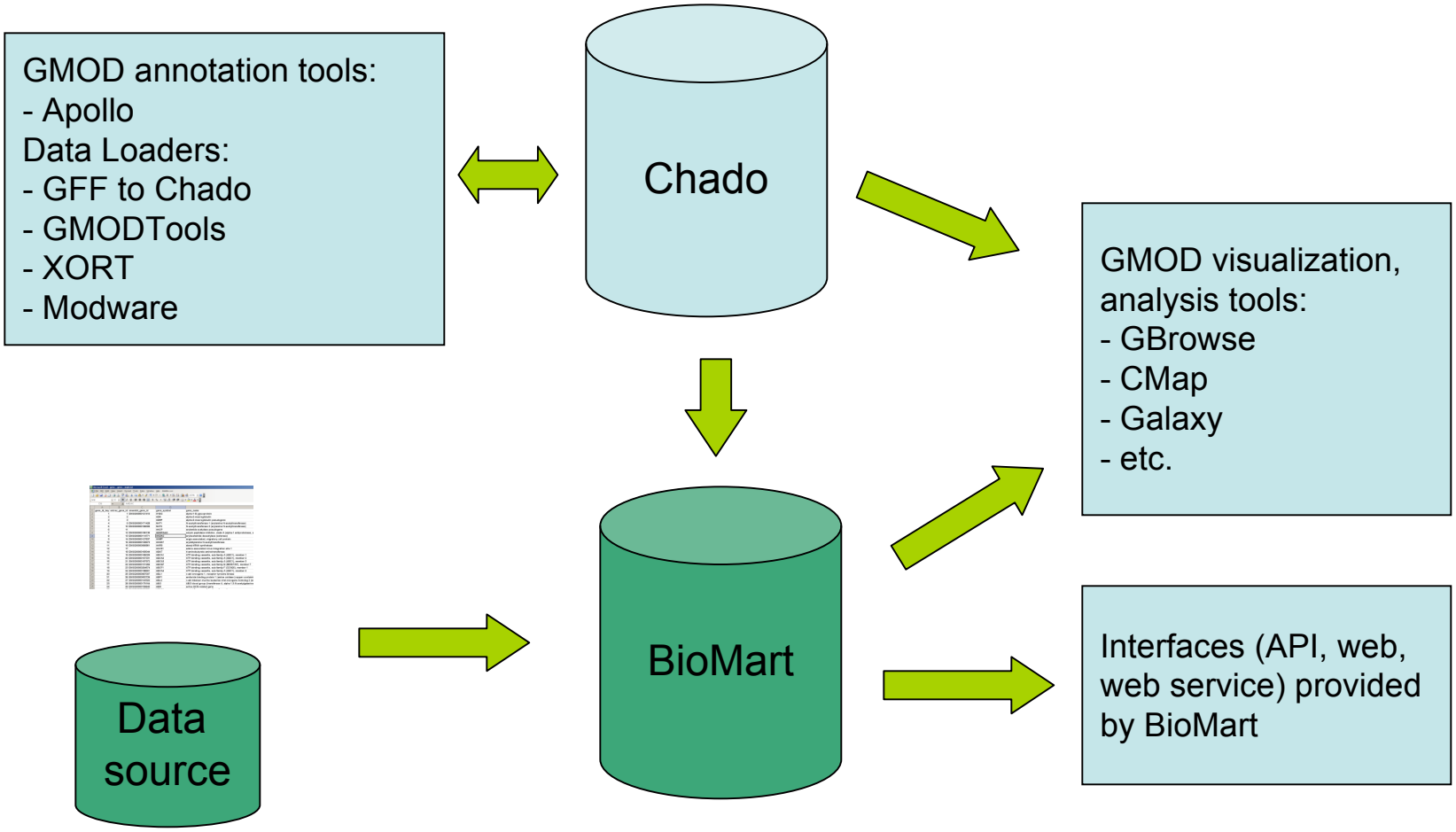



```
<Importable
        filters="chromosome_name,start,end"
        internalName="ensembl_das_chr"
        linkName="ensembl_das_chr"
        name="ensembl_das_chr"
        type="dasChr" />

<Exportable
        attributes="ensembl_gene_id,start_position,end_position,strand,
                ensembl_transcript_id,transcript_start,transcript_end,
                ensembl_exon_id,exon_chrom_start,exon_chrom_end"
        internalName="ensembl_das_chr"
        linkName="ensembl_das_chr"
        name="ensembl_das_chr"
        pointer="true"
        type="dasChr" />
```

Importable defines the filter(s) of the query
Exportable defines the attribute(s) of the query

BioMart in addition to Chado as a data source



BioMart Adaptor

Via: Web Service	Via: Perl API
BioMart-aware code for submitting query to a BioMart Server	
Third party app aware code for preparing query and retrieving data	

Hackathon?

Acknowledgements

- Syed Haider (EBI)
- Arne Stabenau (OICR)
- Junjun Zhang (OICR)

biomart

Ontario Institute for Cancer Research

science → discoveries → solutions