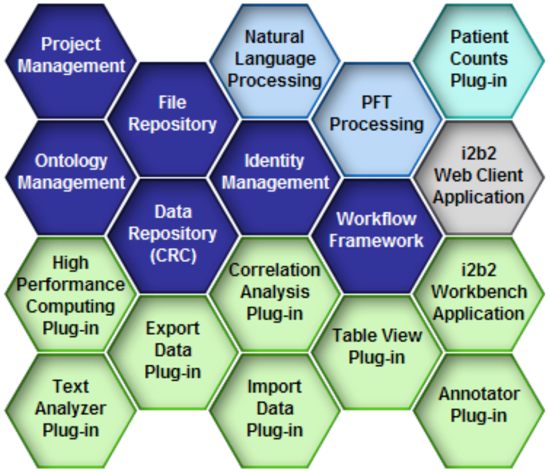
GARLIC - Genomic Analysis Result Library Interface Cell

Harvard Medical School
2011 Open Source Developer Retreat

- Brief Overview Of i2b2
- Assumptions – Requirements – Challenges
- Learning From Others
- Proposed Strategy
- Data Model
- Ontologies & Data Dictionaries
- Core Technology / Mongo / BioMart / Galaxy
- External System & Application Integration

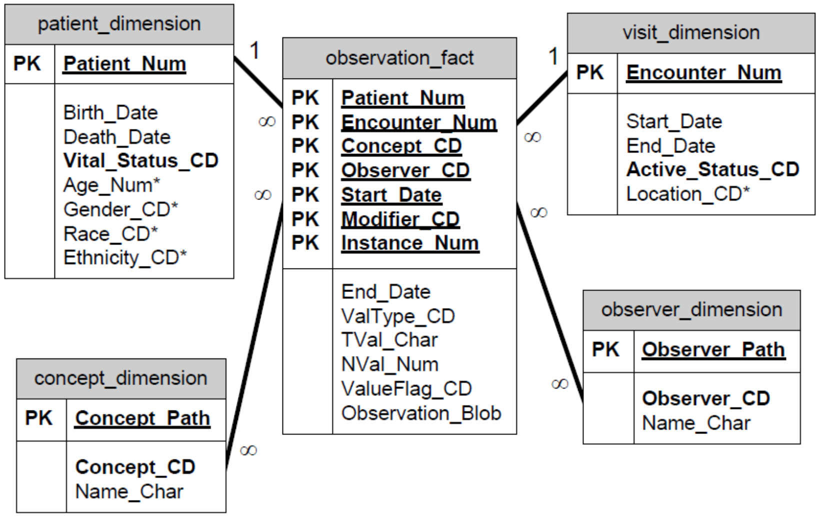
Brief Overview Of i2b2

> *….a scalable informatics framework that will enable clinical researchers to use existing clinical data for discovery research and, when combined with IRB-approved genomic data, facilitate the design of targeted therapies for individual patients with diseases having genetic origins.*



Project Management
File Repository
Natural Language Processing
PFT Processing
Patient Counts Plug-in
Ontology Management
Identity Management
i2b2 Web Client Application
Data Repository (CRC)
Workflow Framework
High Performance Computing Plug-in
Correlation Analysis Plug-in
i2b2 Workbench Application
Export Data Plug-in
Table View Plug-in
Text Analyzer Plug-in
Import Data Plug-in
Annotator Plug-in

- Modular – "Cell"
- HTTP – XML Restful (& SOAP)
- Java 1.5
  - JBOSS 4 & AXIS
- Rich Client
  - Eclipse Frame Work
- Web Client
  - Yahoo UI Javascript
- Data Warehouse Star Schema
  - SQLServer & Oracle



patient_dimension
PK **Patient_Num**
Birth_Date
Death_Date
Vital_Status_CD
Age_Num*
Gender_CD*
Race_CD*
Ethnicity_CD*

observation_fact
PK **Patient_Num**
PK **Encounter_Num**
PK **Concept_CD**
PK **Observer_CD**
PK **Start_Date**
PK **Modifier_CD**
PK **Instance_Num**
End_Date
ValType_CD
TVal_Char
NVal_Num
ValueFlag_CD
Observation_Blob

visit_dimension
PK **Encounter_Num**
Start_Date
End_Date
Active_Status_CD
Location_CD*

concept_dimension
PK **Concept_Path**
Concept_CD
Name_Char

observer_dimension
PK **Observer_Path**
Observer_CD
Name_Char

https://www.i2b2.org/events/slides/i2b2_AMIA_Tutorial_20100310.pdf

Assumptions – Requirements - Challenges

- Initial Focus On Variant Data
  Sequence & Chip based

- Balancing the Needs Of Multiple Consumer / Producer Types
  Clinical Researcher, Bioinformatician, Biostatistician, Programmer…
  Variant Data From Research
  Variant Data From Clinical Practice

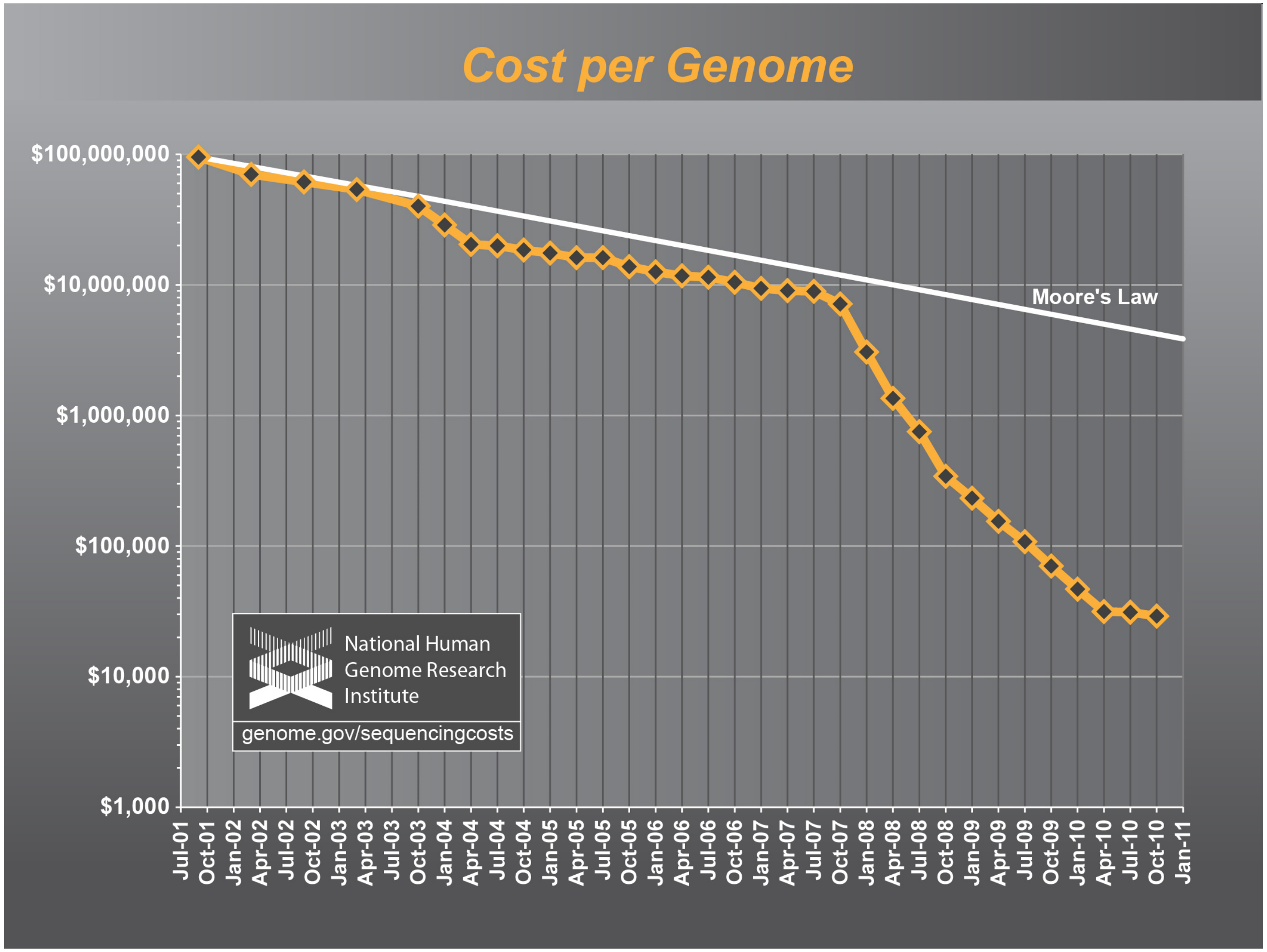
- Extensibility & Scalability
  - Store 2M Variants per Patient
  - Store 11 Facts Per Variant
  - Store 5000 Patients
  - Vertical Vs Horizontal scaling
- Minimize Impact On Exist i2b2 Core Software Components

Currently in the demo i2b2 project, the OBSERVATION_FACT table requires around 0.18KB per fact (row)

2M Variants x 11 Facts x 5000 patients => ~ 19TB

Utilizing OBSERVATION_BLOB field could offset this overhead but reduces data accessibility

Predicted Demand



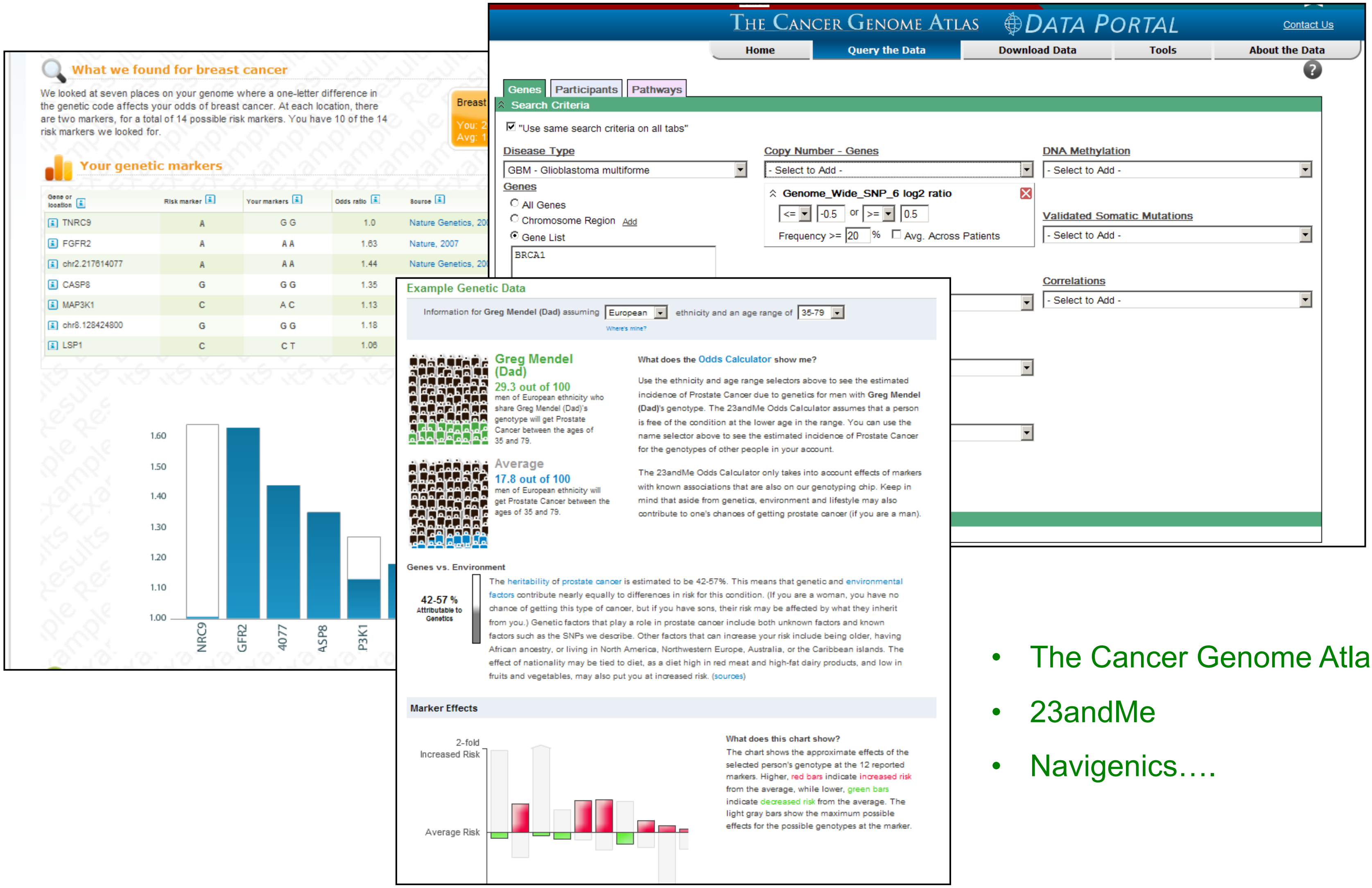
http://www.genome.gov/sequencingcosts/

The challenge……

Distill and present the information derived from an NGS variant analysis dataset within i2b2 to enable researches to exploit the genomic knowledge it contains…

Learning From Others

Clinical Researcher Perspective



What we found for breast cancer

We looked at seven places on your genome where a one-letter difference in the genetic code affects your odds of breast cancer. At each location, there are two markers, for a total of 14 possible risk markers. You have 10 of the 14 risk markers we looked for.

Breast
You: 2
Avg: 1

Your genetic markers

Gene or location	Risk marker	Your markers	Odds ratio	Source
TNRC9	A	G G	1.0	Nature Genetics, 20
FGFR2	A	A A	1.63	Nature, 2007
chr2.217614077	A	A A	1.44	Nature Genetics, 20
CASP8	G	G G	1.35	
MAP3K1	C	A C	1.13	
chr8.128424800	G	G G	1.18	
LSP1	C	C T	1.06	

The Cancer Genome Atlas — Data Portal

Search Criteria

☑ "Use same search criteria on all tabs"

Disease Type: GBM - Glioblastoma multiforme

Genes:
○ All Genes
○ Chromosome Region Add
● Gene List

BRCA1

Copy Number - Genes: - Select to Add -

Genome_Wide_SNP_6 log2 ratio
<= -0.5 or >= 0.5
Frequency >= 20 % □ Avg. Across Patients

DNA Methylation: - Select to Add -

Validated Somatic Mutations: - Select to Add -

Correlations: - Select to Add -

Example Genetic Data

Information for **Greg Mendel (Dad)** assuming European ethnicity and an age range of 35-79

Where's mine?

Greg Mendel (Dad)
29.3 out of 100
men of European ethnicity who share Greg Mendel (Dad)'s genotype will get Prostate Cancer between the ages of 35 and 79.

Average
17.8 out of 100
men of European ethnicity will get Prostate Cancer between the ages of 35 and 79.

What does the Odds Calculator show me?

Use the ethnicity and age range selectors above to see the estimated incidence of Prostate Cancer due to genetics for men with **Greg Mendel (Dad)**'s genotype. The 23andMe Odds Calculator assumes that a person is free of the condition at the lower age in the range. You can use the name selector above to see the estimated incidence of Prostate Cancer for the genotypes of other people in your account.

The 23andMe Odds Calculator only takes into account effects of markers with known associations that are also on our genotyping chip. Keep in mind that aside from genetics, environment and lifestyle may also contribute to one's chances of getting prostate cancer (if you are a man).

Genes vs. Environment

42-57 %
Attributable to Genetics

The heritability of prostate cancer is estimated to be 42-57%. This means that genetic and environmental factors contribute nearly equally to differences in risk for this condition. (If you are a woman, you have no chance of getting this type of cancer, but if you have sons, their risk may be affected by what they inherit from you.) Genetic factors that play a role in prostate cancer include both unknown factors and known factors such as the SNPs we describe. Other factors that can increase your risk include being older, having African ancestry, or living in North America, Northwestern Europe, Australia, or the Caribbean islands. The effect of nationality may be tied to diet, as a diet high in red meat and high-fat dairy products, and low in fruits and vegetables, may also put you at increased risk. (sources)

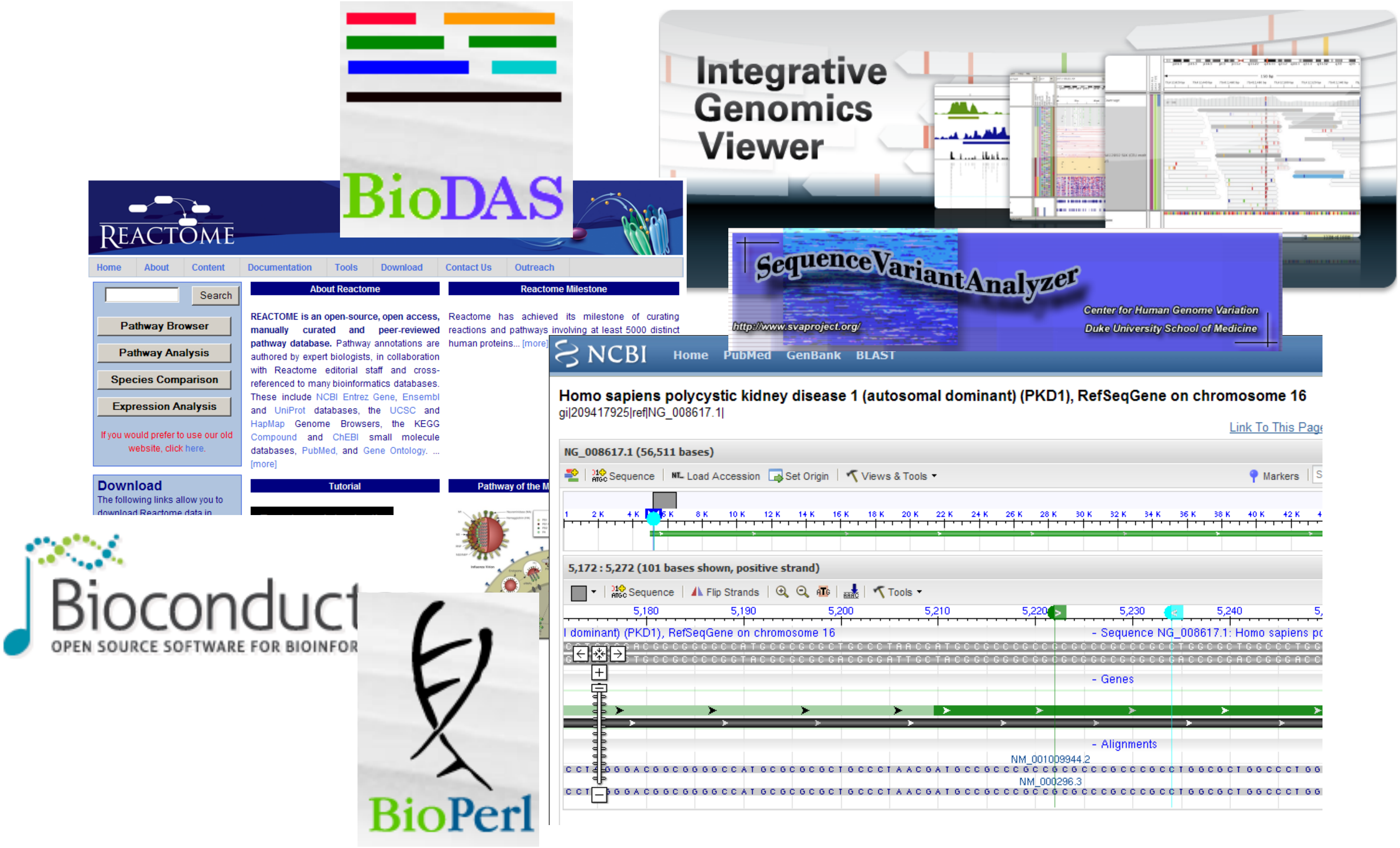
Marker Effects

What does this chart show?
The chart shows the approximate effects of the selected person's genotype at the 12 reported markers. Higher, red bars indicate increased risk from the average, while lower, green bars indicate decreased risk from the average. The light gray bars show the maximum possible effects for the possible genotypes at the marker.

- The Cancer Genome Atlas
- 23andMe
- Navigenics….

Learning From Others

Domain Expert Perspective



If you would prefer to use our old website, click here.

Download

The following links allow you to download Reactome data in

About Reactome

REACTOME is an open-source, open access, manually curated and peer-reviewed pathway database. Pathway annotations are authored by expert biologists, in collaboration with Reactome editorial staff and cross-referenced to many bioinformatics databases. These include NCBI Entrez Gene, Ensembl and UniProt databases, the UCSC and HapMap Genome Browsers, the KEGG Compound and ChEBI small molecule databases, PubMed, and Gene Ontology. ... [more]

Reactome Milestone

Reactome has achieved its milestone of curating reactions and pathways involving at least 5000 distinct human proteins... [more]

Homo sapiens polycystic kidney disease 1 (autosomal dominant) (PKD1), RefSeqGene on chromosome 16

gi|209417925|ref|NG_008617.1|

NG_008617.1 (56,511 bases)

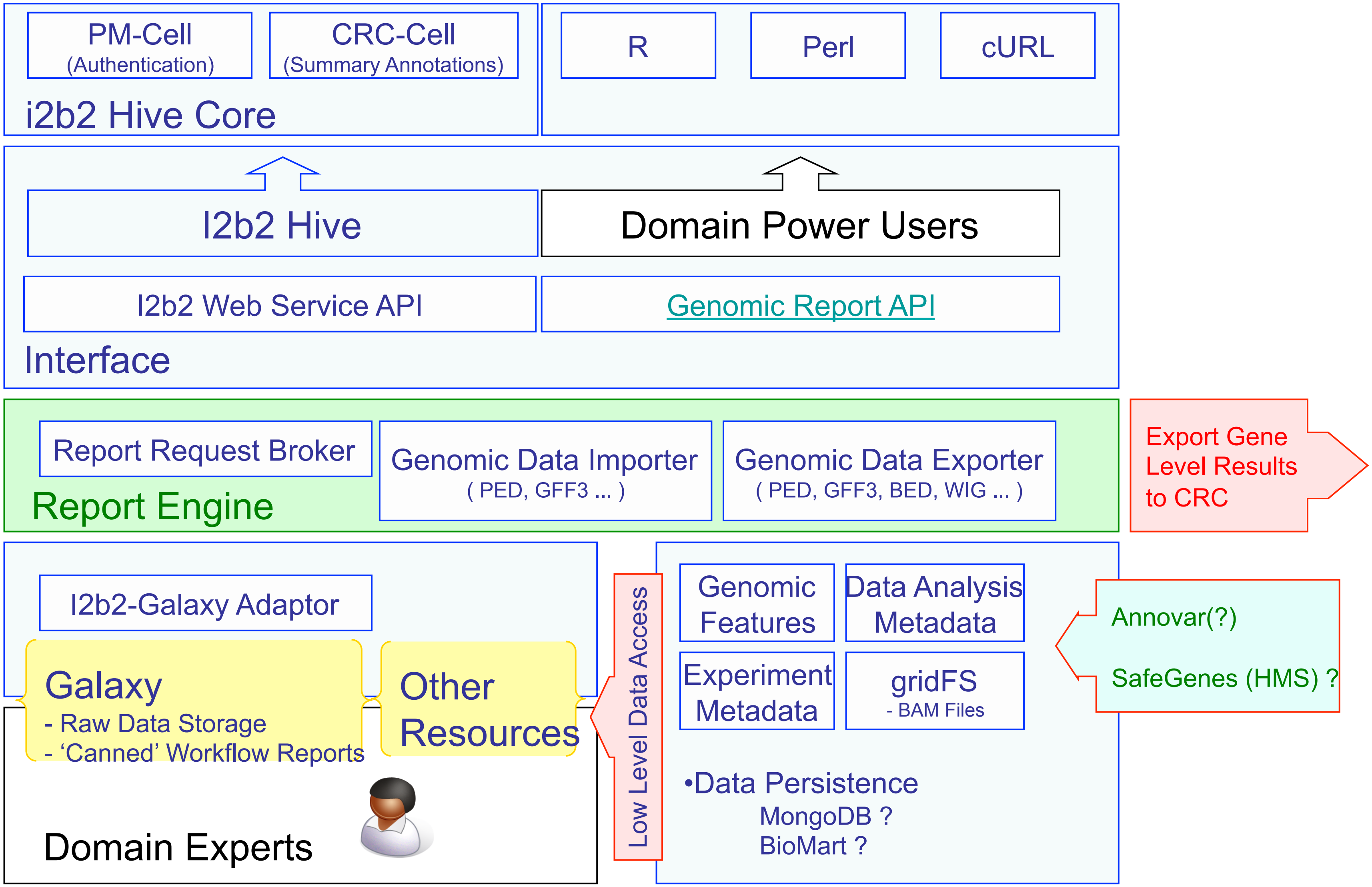
5,172 : 5,272 (101 bases shown, positive strand)

l dominant) (PKD1), RefSeqGene on chromosome 16 - Sequence NG_008617.1: Homo sapiens pc

Proposed Strategy

1. Store Summarized Genomic Annotation Information Within the current OBSERVATION_FACT table
    - Genomic Landmark-Centric e.g. Gene
2. Separate Storage Detailed Genomic Annotation & Result Information
    - Object Based (MongoDB?)
    - BioMart (http://www.biomart.org/) ?
3. Store Genomic Datasets (BAM, PED etc…) Within A Secure File System – Indexed within i2b2 Data Mart
4. Analytics "Workflow Engine"
    - Galaxy (http://galaxy.psu.edu)

Component Diagram



PM-Cell
(Authentication)

CRC-Cell
(Summary Annotations)

i2b2 Hive Core

R

Perl

cURL

I2b2 Hive

Domain Power Users

I2b2 Web Service API

Genomic Report API

Interface

Report Request Broker

Report Engine

Genomic Data Importer
(PED, GFF3 ...)

Genomic Data Exporter
(PED, GFF3, BED, WIG ...)

Export Gene Level Results to CRC

I2b2-Galaxy Adaptor

Galaxy
- Raw Data Storage
- 'Canned' Workflow Reports

Other Resources

Domain Experts

Low Level Data Access

Genomic Features

Data Analysis Metadata

Experiment Metadata

gridFS
- BAM Files

- Data Persistence
  MongoDB ?
  BioMart ?

Annovar(?)

SafeGenes (HMS) ?

Data Model

- i2b2 Encounter = ‘genomic analysis’
  - Single Patient
  - Single Sample
  - Single Assay / Platform

- Analysis Results – Annotation (GFF3/GVF)
  - Utilize The GFF3/GVF File Format For Describing Annotations
  - Summary stored within OBSERVATION_FACT Table
  - Leverage the MODIFIER_CD concept (> i2b2 1.6 RC2)
  - Detailed Annotations Stored Within A Genomic Centric DB (MongoDB, Cassandra, CouchDB, BioMART)

- Analysis Results – Data Sets (PED, BAM, FASTA, CEL)
  - Stored Within A File System
  - Indexed Within i2b2 Observation Fact table
  - Ultimately, create a workflow that can create the GFF annotations based on the uploaded Data Sets (Nice To Have)

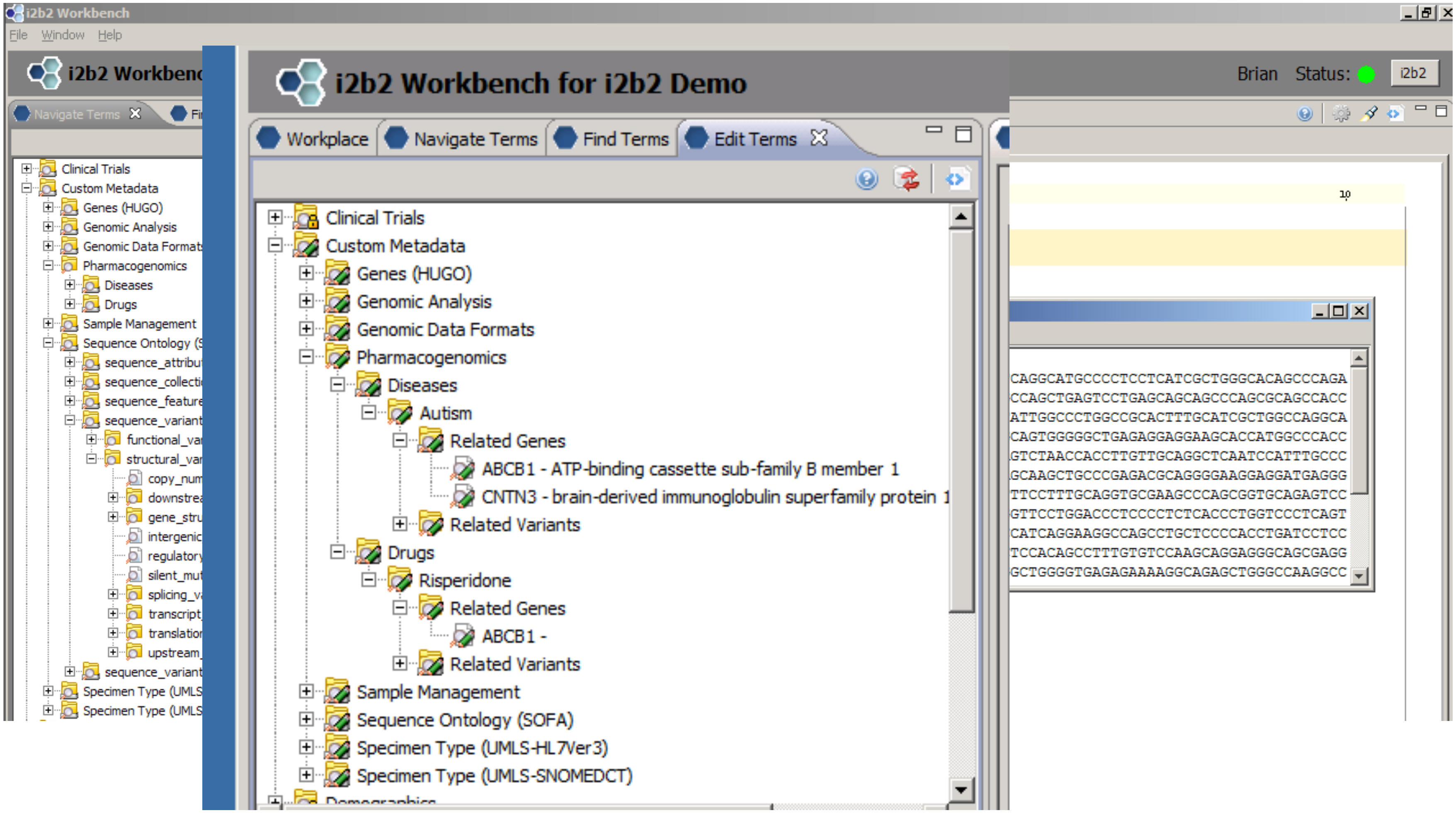
Ontologies & Data Dictionaries

- Sequence Ontology Feature Annotation (SOFA)
  Describe the genomic features detected
  - http://www.sequenceontology.org/
- Experimental Factor Ontology (EFO)
  Broadly describe the type of genomic assay performed
- Software Ontology (SWO)
  Track data file formats
- UMLS (SNOMEDCT / HL7 Ver3)
  Sample Type & Pathology
- Human Genome Organisation (HUGO)
  Gene Symbols
- Annotation Pipelines (Annovar, Safegenes, PharmGKB)
  - Disease Gene relationship
  - Drug Gene relationship

Core Element

- Java 1.6
- JBOSS 5
- JERSEY Restful Web Service Layer
- Oracle & SQL Server
- jQuery (YUI plug-in development)

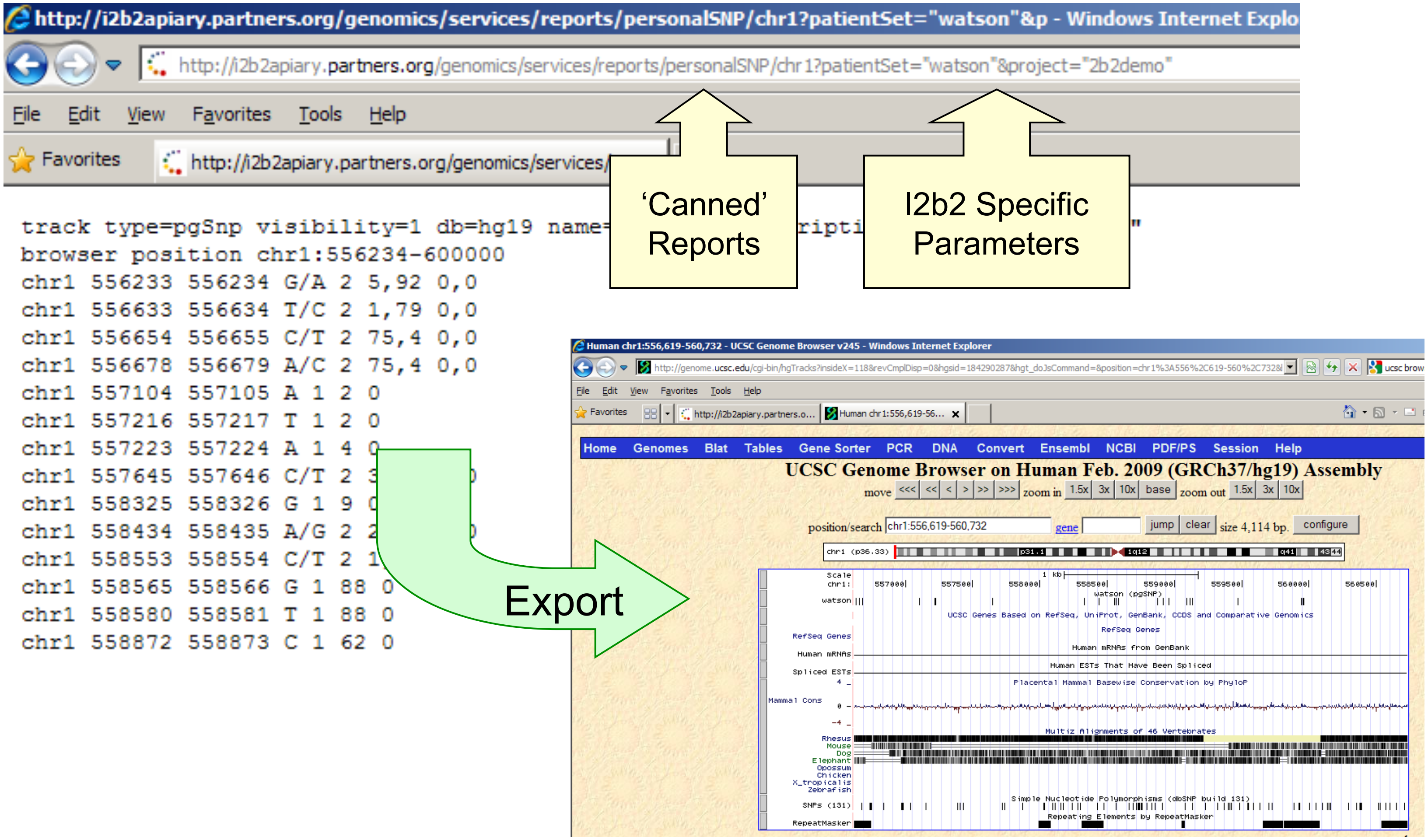
Visual Example



```
CAGGCATGCCCCTCCTCATCGCTGGGCACAGCCCAGA
CCAGCTGAGTCCTGAGCAGCAGCCCAGCGCAGCCACC
ATTGGCCCTGGCCGCACTTTGCATCGCTGGCCAGGCA
CAGTGGGGGCTGAGAGGAGGAAGCACCATGGCCCACC
GTCTAACCACCTTGTTGCAGGCTCAATCCATTTGCCC
GCAAGCTGCCCGAGACGCAGGGGAAGGAGGATGAGGG
TTCCTTTGCAGGTGCGAAGCCCAGCGGTGCAGAGTCC
GTTCCTGGACCCTCCCCTCTCACCCTGGTCCCTCAGT
CATCAGGAAGGCCAGCCTGCTCCCCACCTGATCCTCC
TCCACAGCCTTTGTGTCCAAGCAGGAGGGCAGCGAGG
GCTGGGGTGAGAGAAAAGGCAGAGCTGGGCCAAGGCC
```

Accessing Detailed Results

- Simple ‘Entrez’ like API for Bioinformaticians



http://i2b2apiary.partners.org/genomics/services/reports/personalSNP/chr1?patientSet="watson"&p - Windows Internet Explo

http://i2b2apiary.partners.org/genomics/services/reports/personalSNP/chr1?patientSet="watson"&project="i2b2demo"

```
track type=pgSnp visibility=1 db=hg19 name=          ripti                 "
browser position chr1:556234-600000
chr1 556233 556234 G/A 2 5,92 0,0
chr1 556633 556634 T/C 2 1,79 0,0
chr1 556654 556655 C/T 2 75,4 0,0
chr1 556678 556679 A/C 2 75,4 0,0
chr1 557104 557105 A 1 2 0
chr1 557216 557217 T 1 2 0
chr1 557223 557224 A 1 4 0
chr1 557645 557646 C/T 2 3       0
chr1 558325 558326 G 1 9 0
chr1 558434 558435 A/G 2 2       0
chr1 558553 558554 C/T 2 1
chr1 558565 558566 G 1 88 0
chr1 558580 558581 T 1 88 0
chr1 558872 558873 C 1 62 0
```

‘Canned’ Reports

I2b2 Specific Parameters

Export

Human chr1:556,619-560,732 - UCSC Genome Browser v245 - Windows Internet Explorer

http://genome.ucsc.edu/cgi-bin/hgTracks?insideX=118&revCmplDisp=0&hgsid=184290287&hgt_doJsCommand=&position=chr1%3A556%2C619-560%2C732&

UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

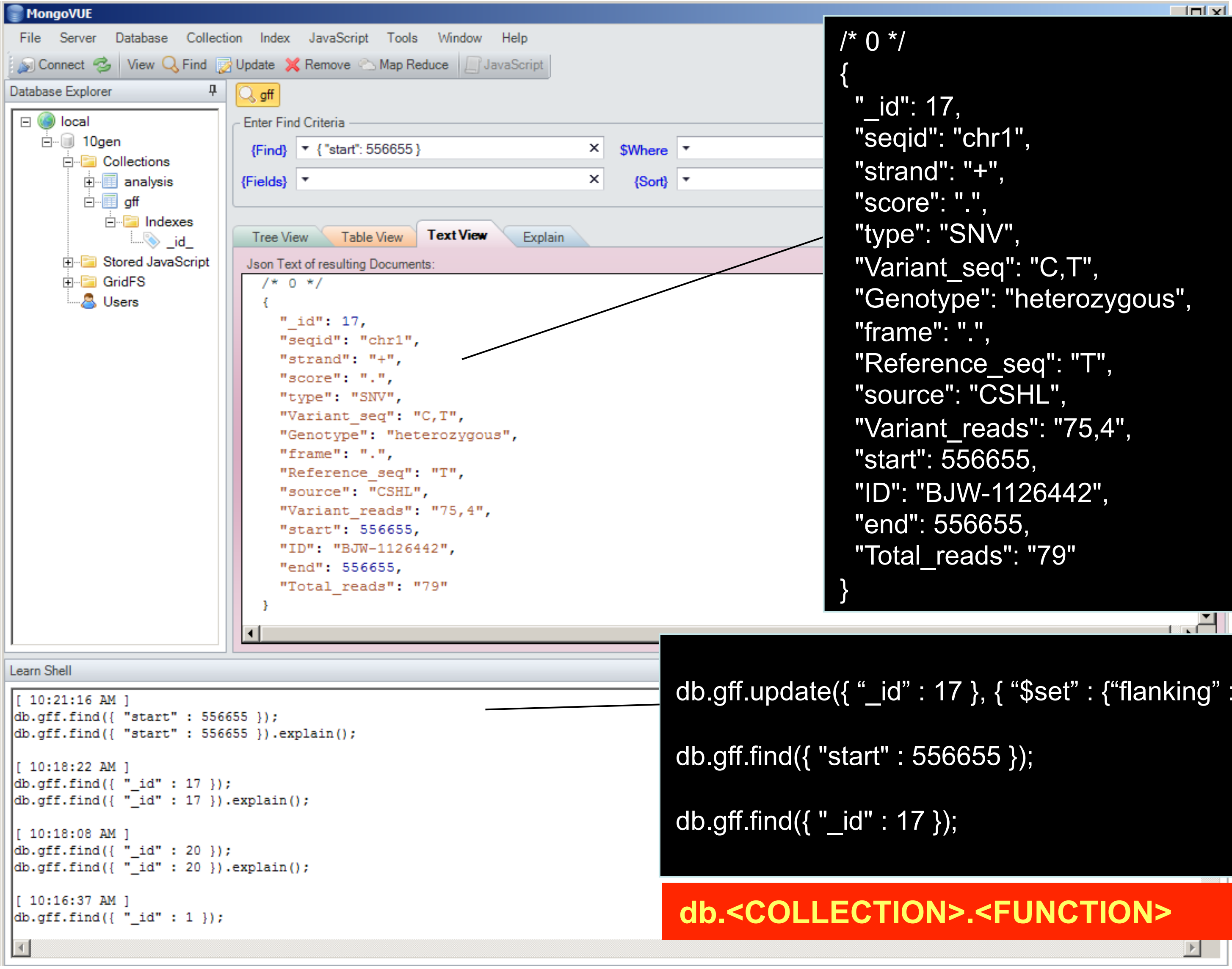
position/search chr1:556,619-560,732 size 4,114 bp.

MongoDB

- Document Oriented database written in C++
- JSON formatted Documents
- Uses Javascript
- Supports sharding
- Use MapReduce
- Language specific drivers (C, C++, Java, Perl, Python, Ruby, Scala…)

- Document ~ Row in RDBMS
- Collection ~ Table in RDBMS

MongoVUE



{Find} { "start": 556655 }

```
/* 0 */
{
  "_id": 17,
  "seqid": "chr1",
  "strand": "+",
  "score": ".",
  "type": "SNV",
  "Variant_seq": "C,T",
  "Genotype": "heterozygous",
  "frame": ".",
  "Reference_seq": "T",
  "source": "CSHL",
  "Variant_reads": "75,4",
  "start": 556655,
  "ID": "BJW-1126442",
  "end": 556655,
  "Total_reads": "79"
}
```

```
[ 10:21:16 AM ]
db.gff.find({ "start" : 556655 });
db.gff.find({ "start" : 556655 }).explain();

[ 10:18:22 AM ]
db.gff.find({ "_id" : 17 });
db.gff.find({ "_id" : 17 }).explain();

[ 10:18:08 AM ]
db.gff.find({ "_id" : 20 });
db.gff.find({ "_id" : 20 }).explain();

[ 10:16:37 AM ]
db.gff.find({ "_id" : 1 });
```

```
/* 0 */
{
  "_id": 17,
  "seqid": "chr1",
  "strand": "+",
  "score": ".",
  "type": "SNV",
  "Variant_seq": "C,T",
  "Genotype": "heterozygous",
  "frame": ".",
  "Reference_seq": "T",
  "source": "CSHL",
  "Variant_reads": "75,4",
  "start": 556655,
  "ID": "BJW-1126442",
  "end": 556655,
  "Total_reads": "79"
}
```

```
db.gff.update({ “_id” : 17 }, { “$set” : {“flanking” : “AGGGGTCGT” }});

db.gff.find({ "start" : 556655 });

db.gff.find({ "_id" : 17 });
```

db.<COLLECTION>.<FUNCTION>

BioMart

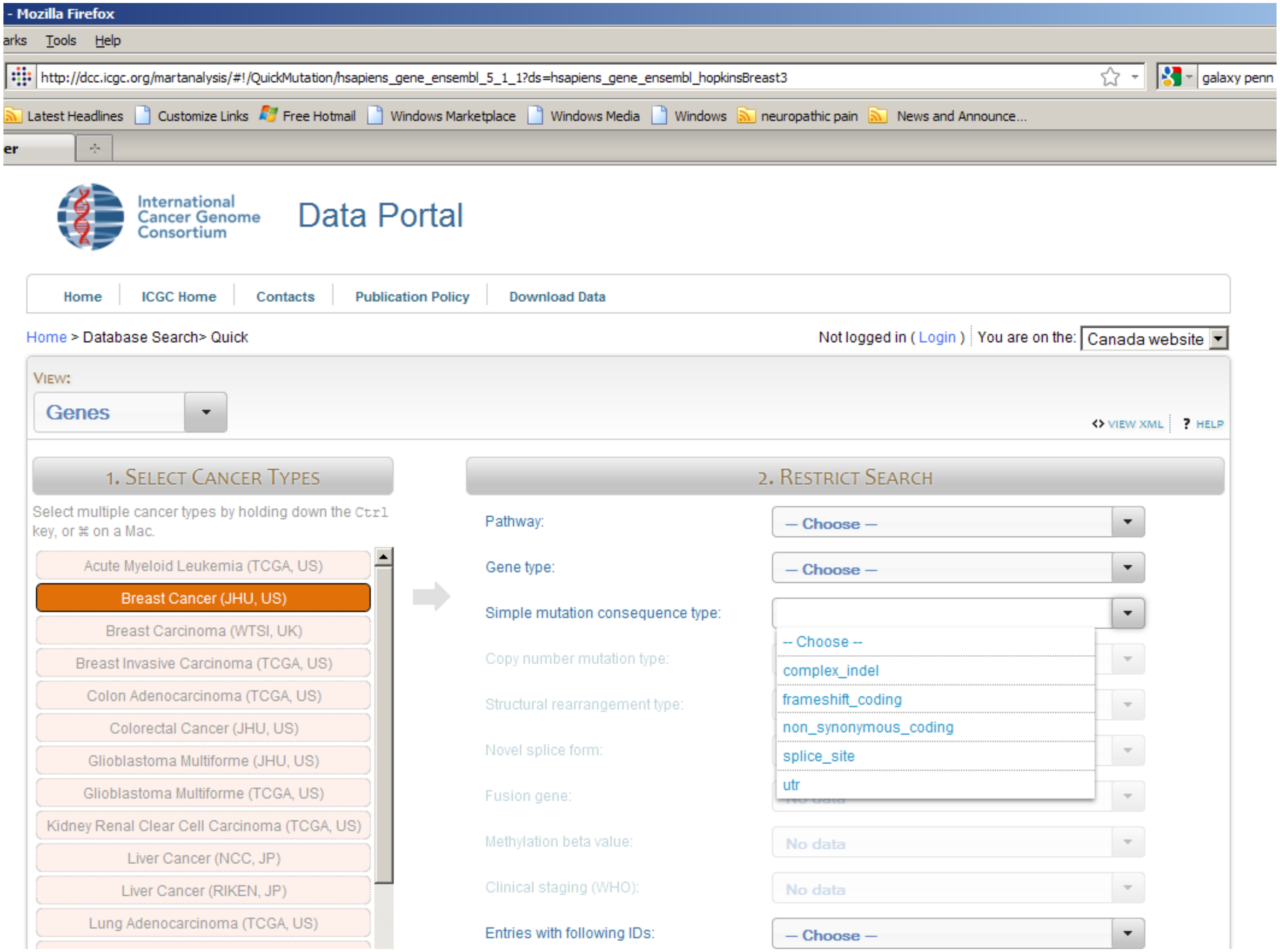
- Developed By
  - European Bioinformatics Institute (http://www.ebi.ac.uk)
  - Cold Spring Harbor Laboratory (http://www.cshl.edu)
  - Ontario Cancer Research Institute (http://www.oicr.on.ca)

- 3 Tier architecture
  - One ore more RDBMS (Oracle, MySQL, Postgres) w/ BioMart compliant schemas & multiple datasets per schema
  - APIs (Perl & Java)
  - Query interfaces
    - MartView, a web browser interface, based on the Perl API.
    - MartService, a web services interface, based on the Perl API.
    - MartURLAccess, a URL based access to MartView, based on Perl API.
    - MartExplorer, a standalone GUI tool, based on the Java API.
    - MartShell, a command-line tool, also based on the Java API.

- Distributed Annotation System (http://www.ensembl.org/info/docs/das/index.html)

  DAS is a specification of a protocol for requesting and returning annotation data for genomic regions. DAS allows sequence annotation to be stored in a decentralised manner, by multiple third-party annotators, and integrated on an as-needed basis by client-side software.

BioMart – 0.8 RC5


- Mozilla Firefox
arks Tools Help
galaxy penn
Latest Headlines Customize Links Free Hotmail Windows Marketplace Windows Media Windows neuropathic pain News and Announce...
er

http://dcc.icgc.org/martanalysis/#!/QuickMutation/hsapiens_gene_ensembl_5_1_1?ds=hsapiens_gene_ensembl_hopkinsBreast3

International Cancer Genome Consortium
Data Portal
Home | ICGC Home | Contacts | Publication Policy | Download Data

Home > Database Search> Quick

Not logged in (Login) | You are on the: Canada website

View: Genes

1. Select Cancer Types

Select multiple cancer types by holding down the Ctrl key, or ⌘ on a Mac.

- Acute Myeloid Leukemia (TCGA, US)
- Breast Cancer (JHU, US)
- Breast Carcinoma (WTSI, UK)
- Breast Invasive Carcinoma (TCGA, US)
- Colon Adenocarcinoma (TCGA, US)
- Colorectal Cancer (JHU, US)
- Glioblastoma Multiforme (JHU, US)
- Glioblastoma Multiforme (TCGA, US)
- Kidney Renal Clear Cell Carcinoma (TCGA, US)
- Liver Cancer (NCC, JP)
- Liver Cancer (RIKEN, JP)
- Lung Adenocarcinoma (TCGA, US)

2. Restrict Search

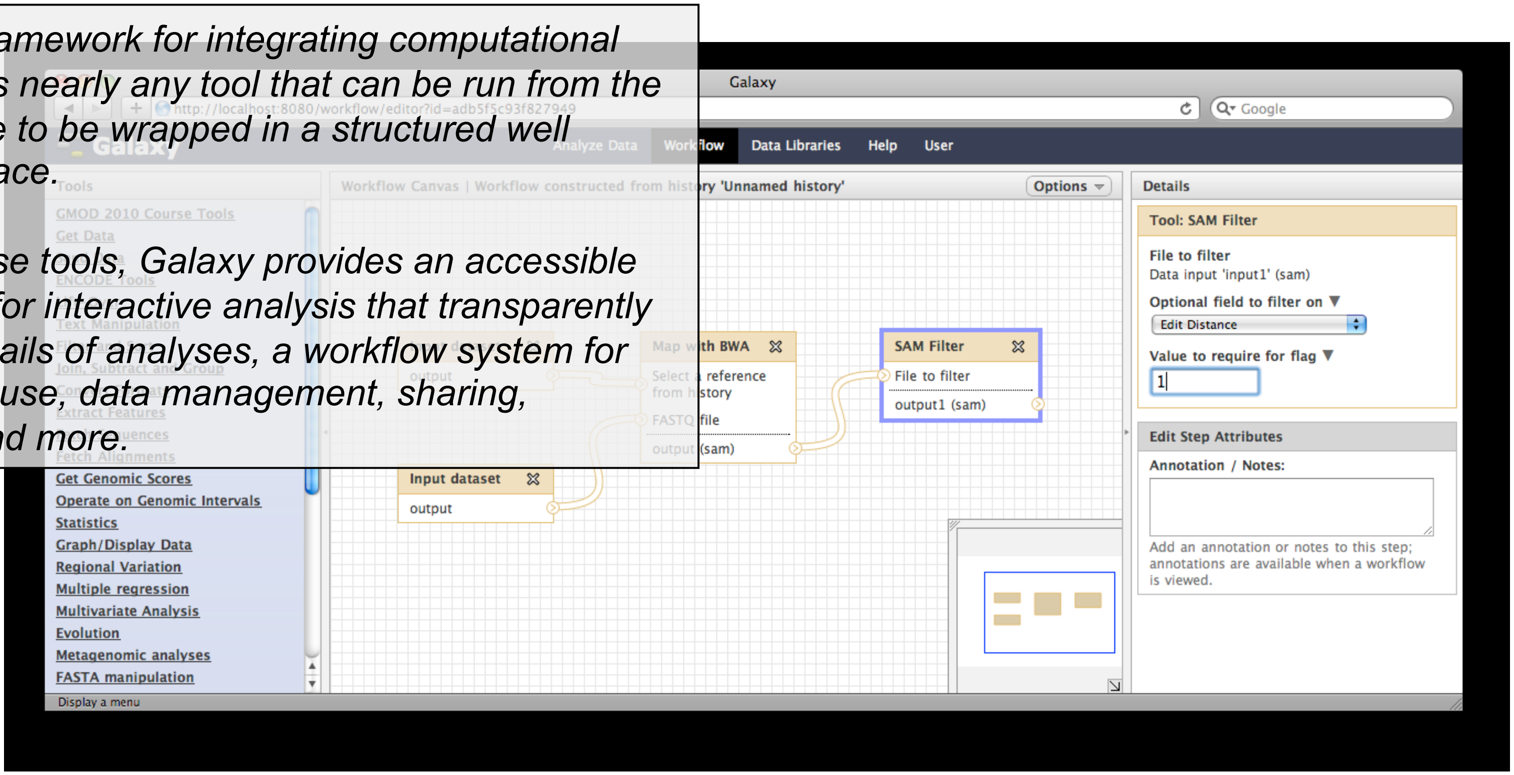
Pathway:	-- Choose --
Gene type:	-- Choose --
Simple mutation consequence type:	-- Choose -- complex_indel frameshift_coding non_synonymous_coding splice_site utr
Copy number mutation type:	
Structural rearrangement type:	
Novel splice form:	
Fusion gene:	No data
Methylation beta value:	No data
Clinical staging (WHO):	No data
Entries with following IDs:	-- Choose --

Galaxy

- Developed By
  - Penn State Uni. (http://galaxy.psu.edu)
  - Python based analysis bioinformatic environment

Galaxy is a framework for integrating computational tools. It allows nearly any tool that can be run from the command line to be wrapped in a structured well defined interface.

On top of these tools, Galaxy provides an accessible environment for interactive analysis that transparently tracks the details of analyses, a workflow system for convenient reuse, data management, sharing, publishing, and more.



Tool: SAM Filter

File to filter
Data input 'input1' (sam)

Optional field to filter on: Edit Distance

Value to require for flag: 1

https://bitbucket.org/galaxy/galaxy-central/wiki/ISMB2010_GalaxyTutorial_3_RunningYourOwn

Integration

- Genome Browsers
  - UCSC
  - Ensembl
  - NCBI

- Applications
  - **GALAXY** (http://galaxy.psu.edu/)
    Penn State University
  - **WGAViewer** (http://people.genome.duke.edu/~dg48/WGAViewer/whatis.php)
    Duke University School Of Medicine
  - **SequenceVariantAnalyzer** (http://www.svaproject.org/)
    Duke University School Of Medicine
  - **Cytoscape**
    w/ Reactome Plugin
  - **JalView**
    Sequence alignment visualization