NCRI INFORMATICS INITIATIVE

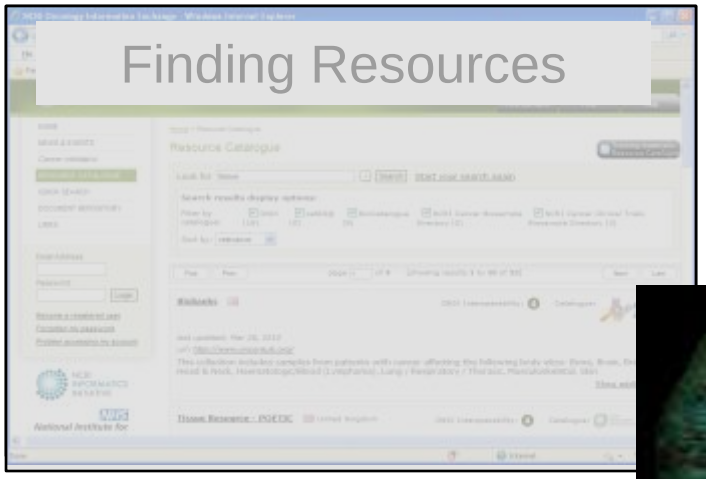
NCRI National Cancer Research Institute

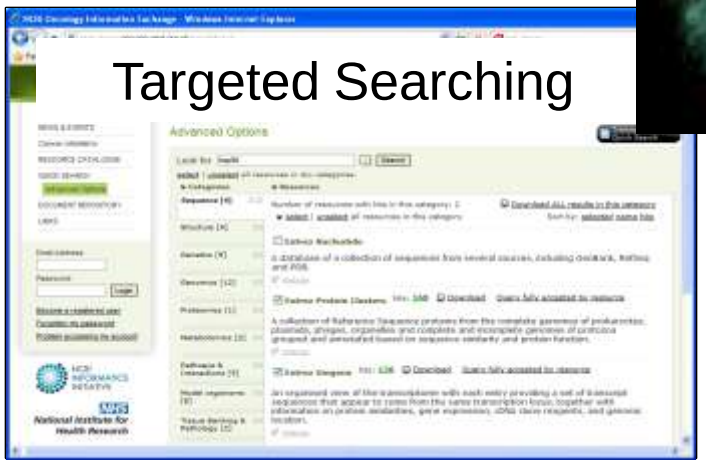
# Information Aggregation and Searching

June 2010 – NCRI-NCI Joint Conference

Dr Alan Hogg

# ONIX Information Portal





NCRI INFORMATICS INITIATIVE

NCRI

# ONIX Quick Search

NCRI Oncology Information Exchange - Windows Internet Explorer

NCRI ONCOLOGY INFORMATION EXCHANGE

HOME
NEWS & EVENTS
Cancer Informatrix
RESOURCE CATALOGUE
QUICK SEARCH
Advanced Options
DOCUMENT REPOSITORY
LINKS

Email Address
Password
Login

Become a registered user
Forgotten my password
Problem accessing my account

NCRI INFORMATICS INITIATIVE

NHS National Institute for Health Research

caBIG

Home > Quick Search > Advanced Options

## Advanced Options

Training materials: Quick Search

Look for mgmt Search

select | unselect all resources in ALL categories

**Categories**

Sequence [4]
Structure [4]
Genetics [9]
Genomics [12]
Proteomics [1]
Metabolomics [2]
Pathways & Interactions [5]
Model organisms [8]
Tissue Banking & Pathology [2]
Imaging [1]
Clinical [4]
Therapeutics [4]
Epidemiology & Pop Studies [5]
Nanotech [0]
Literature [9]

**Resources**

Number of resources with hits in this category: 3

Download ALL results in this category

select | unselect all resources in this category

Sort by: selected name hits

**ENCyclopedia Of DNA Elements portal** hits: 22 Download Query fully accepted by resource

Data generated by members of the ENCODE Consortium is housed in a number of public databases, such as the UCSC Genome Browser and NCBI's Gene Expression Omnibus (GEO). Since issuing queries to these databases is often not intuitive, the ENCODEdb portal was developed to allow biologists to more easily query and retrieve data generated by the Consortium. This portal provides users a single, unified point-of-access to data generated by the ENCODE Consortium, regardless of which public database the primary data is housed in.

Website

**Entrez OMIM** hits: 3 Download Query fully accepted by resource

A comprehensive collection of human genes and genetic phenotypes.

Website

**Entrez SNP** hits: 5841 Download Query fully accepted by resource

A Single Nucleotide Polymorphism (SNP) database that serves as a central repository for both single base nucleotide substitutions and short deletion and insertion polymorphisms.

Website

**Human Variome Project** hits: 0 Query fully accepted by resource

The global initiative to collect and curate all human genetic variation affecting human health.

Website

**IARC TP53 Mutation DB** hits: 0 Query fully accepted by resource

The IARC TP53 Mutation Database compiles all TP53 gene variations identified in human populations and tumor samples. Data are compiled from the peer-reviewed literature and from generalist databases.

Website

NCRI INFORMATICS INITIATIVE

NCRI

# ONIX Quick Search

NCRI Oncology Information Exchange - Windows Internet Explorer

http://www.ncri-onix.org.uk/portal/#5119

File Edit View Favorites Tools Help

Favorites

NCRI ONCOLOGY INFORMATION EXCHANGE

Find out more | FAQ | Help

HOME
NEWS & EVENTS
Cancer InfoMatrix
RESOURCE CATALOGUE
QUICK SEARCH
Advanced Options
DOCUMENT REPOSITORY
LINKS

Email Address:
Password:
Login

Become a registered user
Forgotten my password
Problem accessing my account

NCRI INFORMATICS INITIATIVE

NHS National Institute for Health Research

caBIG

Home > Quick Search > Advanced Options

## Advanced Options

Training materials: Quick Search

Expand this panel

brca1 Search

Sequence [4]

Structure [4]

Genetics [8]

NCBI Probe hits: 1025
Entrez SNP hits: 733
The Pharmacogenetics and Pharmacogenomics Knowledge Base hits: 153
Entrez OMIM hits: 142
ENCyclopedia Of DNA Elements portal hits: 119
The Catalogue of Somatic Mutations (COSMIC) hits: 26
NCBI Cancer Chromosomes hits: 15
Human Variome Project hits: 2

Genomics [12]

Entrez OMI

First Prev page 1 of 8 [showing results 1 to 20 of 142] Next Last

Download selection (0) Entrez OMIM Select fields to display

**BREAST CANCER 1 GENE; BRCA1**
Oid: *113705
AltTitles: PANCREATIC CANCER, SUSCEPTIBILITY TO, 4, INCLUDED; PNCA4, INCLUDED
Locus: 17q21

**BREAST-OVARIAN CANCER, FAMILIAL, SUSCEPTIBILITY TO, 1; BROVCA1**
Oid: #604370
AltTitles: BREAST CANCER, FAMILIAL, SUSCEPTIBILITY TO, 1, INCLUDED
Locus: 17q21, 14q32.3, 6q25.2-q27, 3q26.3

**BRCA1-INTERACTING PROTEIN 1; BRIP1**
Oid: *605882
AltTitles:
Locus: 17q22

**BRCA1-ASSOCIATED RING DOMAIN 1; BARD1**
Oid: *601593
AltTitles:
Locus: 2q34-q35
See Entrez OMIM entry for BRCA1-ASSOCIATED RING DOMAIN 1; BARD1

**CHROMOSOME 19 OPEN READING FRAME 62; C19ORF62**
Oid: *612766
AltTitles:
Locus: 19p13.11

**ZINC FINGER PROTEIN 350; ZNF350**
Oid: *605422
AltTitles:
Locus:

**NEIGHBOR OF BRCA1 GENE 1; NBR1**
Oid: *166945
AltTitles:
Locus: 17q21.1

http://www.ncbi.nlm.nih.gov/sites/entrez?Db=omim&Cmd=ShowDetailView&TermToSearch=601593&ordinalpos=18&tool=EntrezSystem2.PEntre

Internet 100%

# ONIX Quick Search

- A simple text matching search of currently 67 target resources
- ONIX sends the search string via an ‘ONIX adapter’ to the target
- Target resource undertakes the search
- ONIX collates the returned results which can be saved ,or used to ‘jump’ to the full details within the target resource

# Findings (1) –'Can we talk?'

- Larger institutions (NCBI and KEGG) very consistent computer and human interfaces
- Some configure interfaces for human interaction and not for computer based access
- 75 targets reviewed –
  - Little data
  - No API
  - Data duplicates other resources
  - User Interface not suitable, no operators etc
  - 11 particularly challenging – 8 not deployed

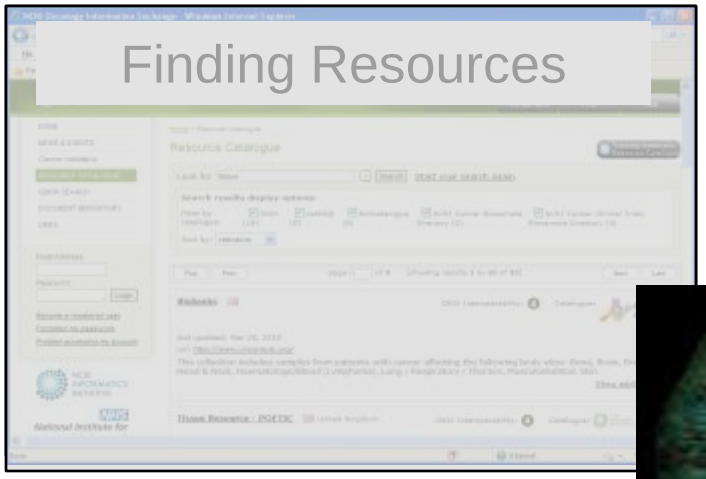
# Findings (2) – an AND or an OR

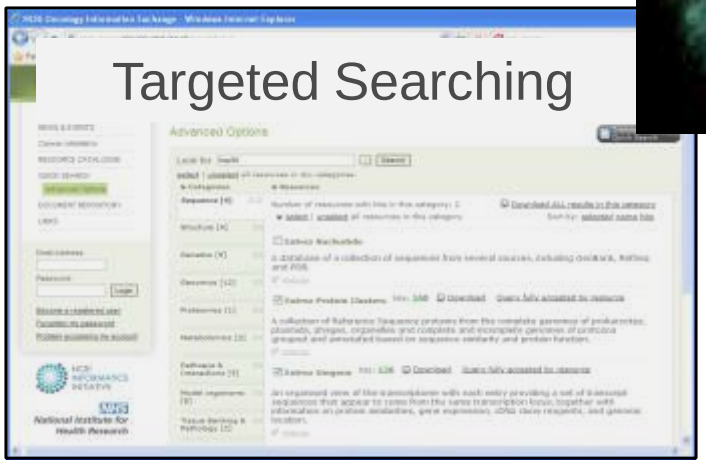
- Majority use a more specific, targeted approach
  - Utilises blank = AND
  - Number of operators available
  - Brings back only perfect matches
- A number utilise Google technology ,or their approach
  - Utilises blank = OR
  - Brings back any possible result and tries to prioritise
- Nine targets only allow 1 search term

# Findings (3) Contents

- Mapping identical items between targets ……
- Inclusion of standards on searching/search operators in development frameworks ……..

# Semantics






'Biologically Aware' Search, Linkage and Data Extraction

# NCRI-UCL Collaboration

- Aims
  - Exploit the rich caGrid semantic infrastructure to support
    - Queries across the information models driven by the existing semantic annotations
    - Effective data integration, minimising the effort for users/clients of the caGrid infrastructure
- Methods
  - Extend the model-driven approach with semantic web/linked data technologies, creating a semantic layer on top of the caGrid infrastructure

# Proof-of-concept projects

- High level queries over caGrid data services
  - Using concepts from NCI thesaurus and associations
    - Paper “Domain Concept-Based Queries for Cancer Research Data Sources” by A González-Beltrán, A Finkelstein, JM Wilkinson, J Kramer – CBMS 2009
- Data integration and data-level queries over caGrid data services
  - Enables analysis of data coming from different resources
    - Paper “Semantic web data warehousing for caGrid”by JP McCusker, JA Phillips, A González-Beltrán, A Finkelstein and M Krauthammer - BMC Bioinformatics 2009
- Linked-Data interface for caGrid data services
  - Enables linking with the datasets available in the Linked Open Data
    - Paper “Exposing caGrid Data Services as Linked Data” by JA Phillips, A González-Beltrán, A Finkelstein, J Pathak - AMIA CRI 2010

## Intelligent Search

Single Nucleotide Polymorphisms
associated with the Gene brca%

GO

**caBIG Query Results: 70**

**objectsalleleA alleleB bigid chrXPseudoAutosomalRegion DBSNPID flank validationStatus id aminoAcidChange codingStatus**

A G hdl://2500.1.PMEUQUCCL5/5R6P5BE3HZ 0 rs8176126 attattgttgaaaatc[A/G]cttgatcacagatgta YES 14433267

A G hdl://2500.1.PMEUQUCCL5/EMCENSYLWJ 0 rs1012130 tatgctgtatgggaactaaataatg[A/G]atacaactcccacccctaatttaga YES 12161581

C T hdl://2500.1.PMEUQUCCL5/LYLATUJCIP 0 rs11571594 tccttccttggttata[A/G]atcaattctagcaggt YES 12422249

**Thanks to – Dr Alejandra González Beltrán**

# Ongoing and Future Work

- Improvement to the services produced by the proof-of-concept projects
- Coordinate the architectural design of the services with the evolving caBIG semantic Infrastructure
  - Knowledge Repository Project
  - caBIG Semantic Web /W3C Projects

# Team at UCL

- Alejandra González-Beltrán
- Ben Tagger
- Anthony Finkelstein

Thanks also to:

- Joshua Phillips

# Thank You

- NCRI Informatics Initiative:
- http://www.cancerinformatics.org.uk

- ONIX Portal:
- http://www.ncri-onix.org.uk

# Interoperability Summary

- Fundamentally computer interfaces are a necessity
- Consistency in approach to search strategy, language and commands could be better
- Need increased consistency in describing the same thing
  - Effort to link to each target is very high at present
- Only if we address these points can we truly start to link data easily and discover previously unknown linkages