NEXTBIO

Making Discoveries Using the NextBio Search Engine

User Guide

January 2009

www.nextbio.com

NextBio User Guide

Contents

Basic Concepts of Overall Search Process

The NextBio search engine is a completely novel web-based platform designed to search and correlate global collections of heterogeneous, large-scale experimental data across diverse platforms, organisms, and therapeutic areas. The NextBio search engine uniquely combines public datasets of information to identify and rank studies and conditions that most highly correlate with a user-defined query. Different queries can consist of genes, proteins, and pathways of interest, as well as an investigator's own experimental results. With the goal of making all data easily accessible to all researchers, even to those scientists without a rigorous background in statistics or bioinformatics, NextBio strictly adheres to principles of simplicity and follows classical web search conventions familiar to the public. Unlike classical web and text search engines, which at their core are based on text matches, the NextBio search engine is based on "correlation matches" – meaning that the most significantly ranked search results depend on the strength of correlations between a user's query and each dataset within all available data. A sound Web 2.0 semantic structure and expansive content enables fast and flexible usability.

Usability

In a highly collaborative environment, NextBio places real experimental data from a range of high throughput platforms and literature resources within the reach of scientists, clinical researchers and scientific managers of all kinds. NextBio enables integrative biology by organizing and presenting results in a way that diminishes the boundaries between scientific domains.

Program Directors
Assemble knowledge to drive program success and utilize study results from diverse scientific disciplines and technology platforms in their quest for more insightful data interpretation. Thereby, gaining a holistic view for better decision making.

Clinical Scientists
Mine the aggregate experimental and clinical data within NextBio to facilitate translational research efforts and elucidate alternative indications for existing compounds.

Research Scientists – Biologists, Chemists, Toxicologists
Query NextBio's repository to validate or generate novel hypotheses prior to investing in new experiments. Correlate results from high-throughput assays to understand treatment effects, validate biomarker activities and identify new targets.

Computational Experts
Explore experimental results and correlate these to a wealth of public and private data within NextBio in order to assist scientists with the biological interpretation of their studies.

Library Scientists
Share a variety of knowledge resources and search expertise across organizations.

The NextBio search engine enables any set of genes or proteins to be searched against all other studies in order to identify significant correlations with other data, providing an easy-to-understand biological context for the otherwise complex data. A researcher can load a set of genes with or without associated statistics into the NextBio search engine and instantly view significant correlations to all other studies. Rank-based enrichment analysis enables quick identification of studies that show very similar positively or negatively associated results from different platforms, data types, and organisms.

The NextBio correlation-based search strategy can be divided into three parts – 1) A "universal" biological dictionary translates information across different data types, platforms, and organisms; 2) a highly standardized pre-processing pipeline for raw experimental data removes the majority of noise and extracts gene subsets significant for a given condition (e.g. all genes that are differentially expressed between control and treated samples); and 3) algorithms that compute the correlation between diverse types of queries and each experimental result. Correlations across all data can be computed for individual genes, sets of genes with ranks indicating their significant association with a given study, or any unranked set of genes with a related biological function (Figure 1). In essence, the biological activity of genes and proteins in each of the thousands of studies is summarized at the level of ranked signatures (e.g. genes differentially expressed between two conditions and ranked by fold change). Pre-computed correlations enable users to explore this space of genes, studies, and functional groups in real time.

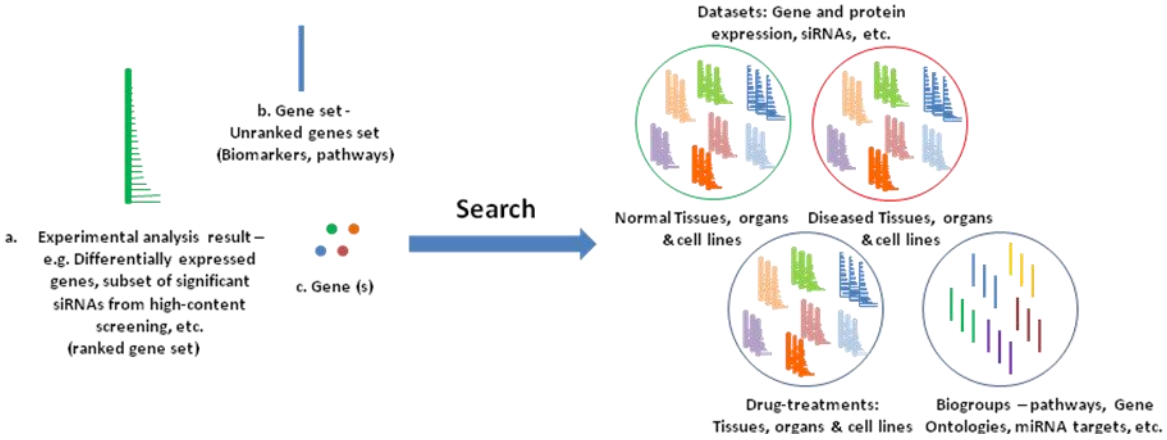


Figure 1 – *Outline of different types of search options available within the NextBio search engine a) with user defined set of data (e.g. differentially-expressed genes ranked by fold change, subset of significant siRNAs with associated scores from a high-content screen of cellular phenotypes); b) with a simple unranked geneset (genes in a pathway, Gene Ontology groups, set of biomarkers or literature-retrieved genes); and c) with a gene or multiple genes of interest. Vertical bars represent a set and ranks of individual elements (e.g. genes or proteins).*

Sources of Public Data

NCBI Gene Expression Omnibus (GEO), Stanford Microarray Database (SMD), and Array Express (AE) are excellent sources of raw microarray and other high-throughput data. The critical distinction between these databases and the NextBio search engine is that the former serve as the repositories of raw data, and not as comprehensive discovery platforms. These repositories do not have a way to allow researchers to postulate questions and to make data correlations based on genes or pathways of interest. For example, the most dynamic result of an NCBI GEO query is to view expression of a single gene across multiple samples within the same study. In the case of a cancer study with hundreds of tumor samples, researchers can see the level of gene expression in each individual sample; however, there is no good way to consider questions that concern i) whether a gene is differentially expressed between certain clinical subsets within that study, ii) what other studies find that gene significantly changing, iii) what other functionally related genes are observed to change with it, and iv) what is the biological context. The NextBio search engine is designed to provide answers to these and other complex questions. Multi-gene expression signatures from one study are automatically compared to all other signatures in the repository to assist researchers to discover, connect, and make sense of related biology and underlying mechanisms.

Cross-platform Comparisons

To compare microarray data from different platforms in the NextBio search engine, an index of microarray platforms was compiled. The index enables recognition of commonly used public gene identifiers as well as specific vendor identifiers, and provides a standardized mapping of individual identifiers from over 25 sources of reference identifiers such as NCBI Entrez Gene, UniGene, Ensembl, RefSeq, or GenBank accession numbers. The conversion of platform-specific identifiers into reference identifiers allows gene comparison across different platforms. Different studies from different platforms also exhibit variable expression profiles and statistics with different dynamic ranges, distributions of fold-changes, and p-values that reflect the technologies used. To allow inter-study comparability, a non-parametric approach was established so that, for each study, ranks are assigned to each feature based on a selected statistical data type, such as fold-change, p-value, and log ratio (Shi et al. 2004). Ranks are then further normalized to eliminate any bias due to various platform sizes.

Cross-species Comparisons

To enable seamless comparison across different species, orthologs are identified for each pair of organisms and are grouped into ortholog clusters. Ortholog information was derived from Mouse Genome Informatics (MGI) at Jackson Lab (www.informatics.jax.org), HomoloGene at NCBI (www.ncbi.nlm.nih.gov), and Ensembl (www.ensembl.org). Ortholog clusters were generated as follows: 1) the manually curated pairwise ortholog data among human, mouse, and rat from MGI were retrieved and clustered to form initial ortholog clusters. 2) The homology group data among human, mouse, rat, fly, and worm were analyzed to remove those in conflict with MGI data. The filtered homology group data were then entered into the ortholog clusters. 3) The whole genome pairwise sequence similarity data from Ensembl were processed to identify reciprocal best hits as candidate orthologs for all pairwise organisms among human, mouse, rat, fly, worm, and yeast. The candidate orthologs were prioritized based on the percentage sequence identity and examined against the existing ortholog cluster. Qualified ortholog candidates were then entered into the ortholog cluster. Once the organism-specific gene identifiers were translated into universal ortholog cluster identifiers, studies performed on different species could then be compared based on the cross-platform principles as previously described for “Cross-platform comparisons”.

Overview of NextBio Content

NextBio content is interactively engineered into over 60 million dynamic web pages.

Data/Biosets	> 2 billion data points and 30,000 study results
	> 4,000 studies (~80% of publically available data)
Biogroups	~ 34,000 gene sets
Diseases	> 65,000 concepts
Tissues	> 9,000 organs & tissues
Compounds	> 8 million compound clusters
Literature	> 18 million articles
Clinical Trials	> 56,000 trials

The NextBio search engine contains data that was collected and preprocessed from several major public and user generated sources.

Public Sites (partial list)
NCBI GEO
EBI Array Express
Stanford Microarray Database (SMD)
Japan's GeMDBJ
Japan's NIBIO
caBIG Consortium
Sanger Center
Jackson Lab Mouse Informatics
dbGAP
NCBI PubChem
Broad Cancer Genomics
International Genomics Consortium

User Generated Content (examples)
Stem Cell Consortium data
Private company data
Burnham Institute data
Others

The predominant publically available experimental data type is from microarray-based gene expression experiments. As new types of platform data become more widely available they will continue to be integrated into the NextBio search engine: such as SNP, ChIP, DNA methylation, genome-wide association, comparative genomic hybridization, proteomic and phenotype data.

The NextBio search engine comprises over 60 million concept term and study specific HTML Home Pages and covers a significant portion of the world's public data across a growing subset of organisms. Figure 2 provides a summary of NextBio content according to organisms and different research areas. As the availability of public high-throughput data and platform technologies evolve, so will the distributions.

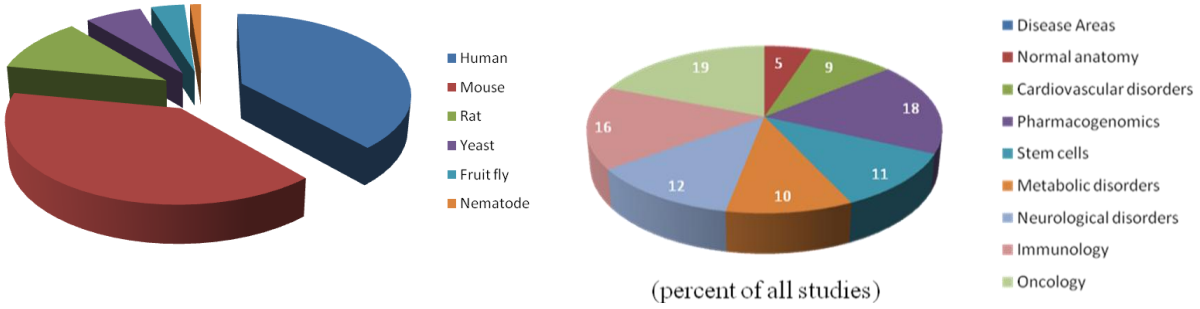


Figure 2 – Public data searchable by the NextBio search engine according to organism and study areas. *http://www.nextbio.com/b/corp/content.nb*

NextBio Indexes, Vocabularies and Ontologies

Gene Index – integrates millions of synonyms and reference IDs, developed by NextBio as indicated above - SNP Index: > 20 million reference SNPs, ~6 million SNP-Gene associations NextBio assembly of resources

SNOMED CT (disease)	www.ihtsdo.org/snomed-ct
FMA (Foundational Model of Anatomy)	http://sig.biostr.washington.edu/projects/fm/AboutFM.html
ATCC cell lines	www.atcc.org
Gene Ontology (functional attributes)	www.geneontology.org

Compound Cluster Index - developed by NextBio from 11 sources, including:

PubChem Compounds	http://pubchem.ncbi.nlm.nih.gov
PubChem Substances	http://pubchem.ncbi.nlm.nih.gov
PubChem InChI/IUPAC	http://pubchem.ncbi.nlm.nih.gov
PubChem smiles/formulas	http://pubchem.ncbi.nlm.nih.gov
DrugBank	www.drugbank.ca
ChemIDPlus	www.nlm.nih.gov/pubs/factsheets/chemidplusfs.html
Stanford Compound Ontology (beta release)	

Authoritative Data Sources

NCBI's Online Mendelian Inheritance in Man (OMIM): 179 OMIM Studies and 1,435 Biosets. 2,730 distinct genes associated with 1,435 diseases with a total of 3,798 associations.
http://www.ncbi.nlm.nih.gov/Omim/getmorbid.cgi

Wellcome Trust Sanger Institute's Cancer Gene Census: 226 diseases linked to 366 gene mutations with 688 distinct associations.
www.sanger.ac.uk/genetics/CGP/Census

Jackson Labs Mouse Genome Informatics (MGI): 587 diseases linked to 905 genes with 1,120 distinct associations. www.informatics.jax.org/phenotypes.shtml

University of Alberta's DrugBank: 945 drugs targeting 442 genes with 1,700 distinct associations.
www.drugbank.ca

NIH's GeneTests data will soon be released into the NextBio system.
www.geneclinics.org

Biogroup Content

Biogroup	Count	Live Link to Example in NextBio	Source Site
Broad MSigDB - Canonical Pathways*	473	GenMAPP: Prostaglandin Synthesis Regulation	www.broad.mit.edu/gsea/msigdb/index.jsp
Broad MSigDB - Positional Gene Sets	331	chr19p13	www.broad.mit.edu/gsea/msigdb/index.jsp
Broad MSigDB - Regulatory Motifs	500	PPARA binding site geneset 1	www.broad.mit.edu/gsea/msigdb/index.jsp
Gene Ontology	23,807	GO: angiogenesis	www.geneontology.org
InterPro	8,573	InterPro: Olfactory receptor	www.ebi.ac.uk/interpro
PMAP Peptidase Inhibitors	20	PMAP: All Peptidase Inhibitors	http://pmap.burnham.org
PMAP Proteases	7	PMAP: Metallo proteases	http://pmap.burnham.org
TargetScan miRNA targets DB	163	TargetScan: miR-23	www.targetscan.org

* Compilation of 12 pathway gene sets including KEGG, Reactome, BioCarta, GenMAPP and others.

NextBio Quality Control Process

Assessing Experimental Design

Data sets pre-processed by NextBio go through extensive quality control processes, about 20% of public data fails and is discarded. Similarly, if the data is contributed to NextBio by individual users or groups for public consumption, then it will go through an initial review by our scientific staff to make sure of the following:

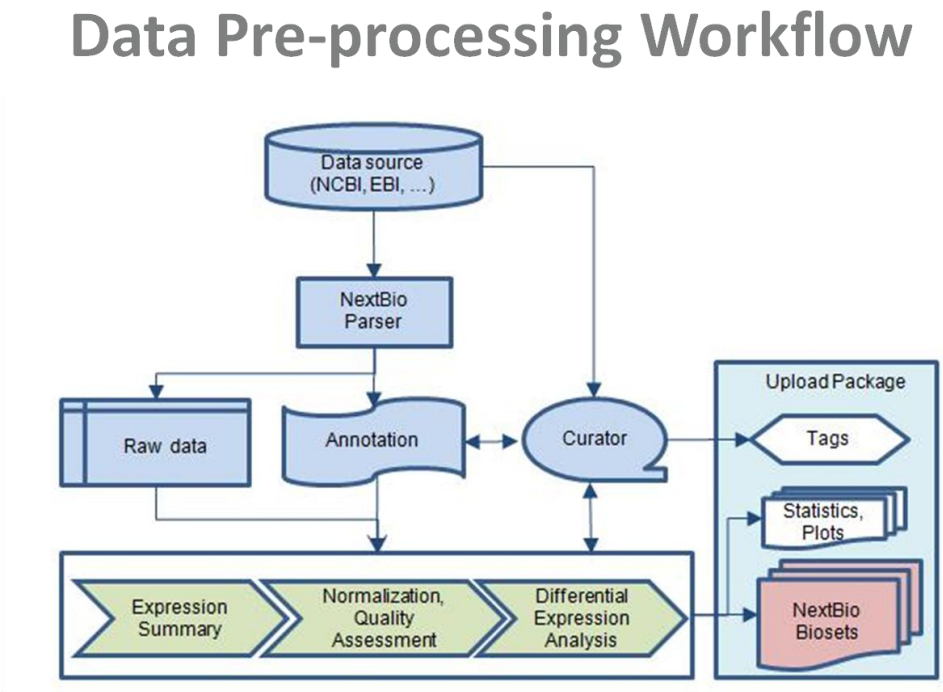
- Experimental design is sound (including replication)
- Analysis approach is reasonable
- Extensive sample annotations and experimental details are provided
- Analysis statistics is available within a result set
- Links to raw data is available for download

Assessing Statistical Significance

All of the raw data goes through rounds of quality control, normalization, and statistical analysis that depend on the data type and on the platform and experimental design used to generate these data. For example, for Affymetrix-based studies in which CEL files are available RMA normalization is applied (Parrish and Spencer 2004). If only MAS5 data is available, then per chip normalizations are applied to reduce inter-chip variability. Although not a prerequisite, the vast majority of processed data in the system falls into the category of case-control experimental design — a set of "test" samples is compared to a set of "control" samples using Welsh or similar t-test type. Basic cutoffs are applied: *i*) one of the two compared sample varieties must exceed a 20th percentile intensity cutoff, *ii*) p-values less than 0.05, and *iii*) fold changes greater than 1.2 or less than -1.2. These are typically the lowest sensitivity thresholds of commercial microarray platforms. Features passing these cutoffs are maintained as the final differentially-expressed gene set, thereby constituting what we call a Bioset with pre-processed directional values. Additional quality control checkpoints are applied at different places depending on the type, platform, and source of experimental data. In cases of other data types, such as cell-based siRNA or proteomics, assay-specific pre-processing steps are applied. Another variety of pre-processing is applied to the "Atlas"-type studies – experiments in which multiple normal tissues and organs are profiled. Since there is no obvious control and "artificial" filtering out of genes from a tissue profile is not desirable (each gene has some "justifiable" low or high level in each tissue), we use a different strategy: For each gene its median across all samples in that study is computed. This serves as the "normal" expression level. Next, each gene's expression value in a given tissue is divided by its median across all tissues in that Study to compute fold changes relative to "normal" atlas-wide expression. Thus, each sample variety (tissue) is represented in a Bioset by all measured genes and their relative fold changes.

As part of the NextBio analysis protocol, the data is initially examined using diagnostic plots such as box-plots for each study. Per-chip normalization is applied when necessary to remove global chip-to-chip variability. Boxplots of post-normalized data are supplied with all public studies for user review. To further examine data quality and experimental design assumptions, hierarchical clustering of samples is used to visually assess separation of samples according to the treatment or test factors. If good separation is not seen for each factor of interest (i.e. factors the experiment was designed to test), it is an indicator that the data may not be good enough to construct reliable Biosets. These analyses, as well as the boxplot analysis, are also used to look for "bad" or "outlier" samples, with the determination on whether to exclude any samples and reanalyze is made on a case-by-case basis. To construct a clustering tree, a statistical test (e.g. ANOVA) is applied to identify a set of genes differentially expressed among different experimental conditions of interest. Clustering of samples using Pearson correlation and based on the statistically significant set of genes identified above is then performed. This is not a Bioset generation step and serves specifically to assess separation of samples based on the parameters tested in the experiment. As well, at a minimum, the number of genes passing the ANOVA (for all comparison parameters) must exceed the number of genes expected by chance. A histogram of the ANOVA p-value distribution from all elements is also generated and

examined for concordance with expected distribution for a study with statistically significant results (i.e. disproportionately lower p-values). This figure is also associated with each public dataset for users to validate study results significance.



Essentials of Data Security

NextBio is the leading search engine organizing the world's life sciences information. Our customers include the top pharmaceutical, biotechnology, medical, academic and institutional research organizations in the world. Our business demands state of the art security at every level, and security of customer data is NextBio's number one priority.

Our commitment to security is reflected in the enormous investments that we have made in our security infrastructure, our internal security policies, and our Privacy Policy. NextBio strives to attain the highest levels of information security as defined by the ISO (International Organization of Standards), PCI (Payment Card Industry), and the AICPA (American Institute of Certified Public Accountants). Using extensive firewall protection, intrusion detection systems, SSL/TLS-based encryption and proprietary security products, NextBio gives you the peace of mind that only a world-class security infrastructure can provide.

We further recommend that users take these steps to secure their own systems prior to using our service:

- Use current Web browsers that support strong SSL/TLS encryption, such as Internet Explorer 6.0 or above and FireFox 2.0 or above.
- Use an anti-virus application with updated engines and definitions.
- Use up-to-date patches of your computer's operating system and all local applications.

Security is paramount in everything that we do, and NextBio utilizes some of the most advanced technology for Internet security available today. To ensure the protection of your data, we have implemented a multi-faceted security policy that includes:

- Protection at the Application Level
- Protection at the Network Level
- Protection at the Facilities Level

Essentials of NextBio Enterprise Domain

NextBio is a Software-as-a-Service (SaaS) provider. As with most SaaS companies it is built on the principle of "single instance, multiple tenants" -- the same hardware is used to serve multiple tenants. The application consists of multiple domains with all information within a domain being available to only individuals within their domain. All data is logically partitioned with a domain or user id associated with each data item. NextBio uses the industry-standard Spring Security (ACEGI) framework to manage the access controls between users, groups and data items.

- Enterprise users will access NextBio through a secure private domain at https://companyname.nextbio.com. Along with this HTTPS private domain, a number of settings that greatly enhance the level of security can be configured. These include things such as password strength and policies, restricting certain IP addresses, session time-outs after a specified time of non-usage, and others.

- Inclusion of corporate data in "Individual Study Results" – The significance of your private data is factored in and correlated with all other data in NextBio. In the results, you'd see data from your various projects interspersed with the public study data.

- Inclusion of corporate data in "Summary Results" – Depending on query type, NextBio presents the most significant concepts within Tissue, Disease, Gene, Treatment and Biogroup columns in the Summary Results section. Intensive, sophisticated meta-analysis algorithms prioritize the most significant concepts from all studies specified by a query. Enterprise customer data will be included and influence the public contributions to Summary Results within their domain.

- Advanced Query Capability – Enterprise customers can access interfaces that allow them to build multi-Bioset queries for direct side-by-side comparisons. Expression signatures can be aggregated and correlated gene by gene to discover highly relevant activities. For instance, to validate hypotheses one may want to combine results from several entirely independent studies measuring gene expression levels from clinical p53 status in breast cancer. Alternatively, one may have a comprehensive study with many clinical variables with a desire to compare gene expression levels within that study alone to tease apart clinical progression.

- Enterprise customers can access APIs, including the following:
  - Search API
  - Data Import API
  - Autocomplete API
  - Single Sign-On

  These are expanded upon in the section titled 'Essentials of Integration via APIs'.

Essentials of Gene Query Processing and Results

Understanding the functional significance of genes requires the ability to investigate expression activity in a systematic way – in different organisms, across different tissues, disease states, and chemical and environmental perturbations. Absolute values, such as fold change, correlation coefficient, p-value, or any other statistical parameter representing gene activity in different types of measurements are converted into a normalized rank, using a combination of the gene's absolute value ranked relative to other measured genes and additional factors, such as the platform size. The final normalized rank indicates the significance of a gene's activity in a given condition within a study. The search function uses this rank to prioritize studies in which different disease states, compound treatments, or other biological factors have the most significant effect on that gene's activity. If orthologs for genes of interest exist, the search will automatically be performed across studies from different organisms. You will often find that the highest ranking study result from a query would have that gene as the #1 highest ranking gene in a Bioset within that study.

As shown under the 'experiments' tab in the ESR1 gene search example below, Pie Charts and Meta-categorization results are displayed directly below the gene description section. All relevant studies retrieved from a Gene Search are shown in the bottom 'Individual Study Results' section. In this case, for a Study to be present it would have to have a significant association to ERS1 changes. These Study results are then categorized into various associations in the 'sources of data and associations for experiments' section. Interactive Pie Charts divide the studies into Organisms and Data Types. Click on any slice of pie to specifically filter down to particular experiments. Not only are the 'individual study results' filtered to the selection, but recalculation of the top 'Normal tissue', 'Disease' and 'Treatment' categories occurs based on that subset of Studies. For instance, it is fascinating to see Treatment categories change when selecting different organisms. For mouse you are likely to see more gene treatments (knock out, conditional expression and the sort) than for rat where you'll likely see more compound treatments. This reflects the historical fact that the mouse became the model of choice for the first mammalian geneticists, while the rat became the model of choice for physiologists, nutritionists and toxicologists.

'Normal tissue', 'Disease' and 'Treatment' categories are derived from meta-analysis of the experimental design tags. Tags generally describe the essence of the experimental design (tissue, disease, treatment, etc), they are described in detail in another section. Meta-categorization for "normal tissues" and "diseases" exclude data from cell lines, as it is assumed that cell lines don't represent true models of normal tissues or diseases.

Gene Query Example, link out to:
http://www.nextbio.com/b/home/home.nb?q=esr1

Interactive Pie Charts act as filters

Interactive Meta-Categories derived from experiments tag correlations

Level 1 results

Level 2 results

Level 3 results

Organism Pie - human subset selected:

Selecting a wedge of pie will subset the experiments and recalculate meta-categories, as with human shown here. Selecting a category class or individual concepts also act as filters to show relevant studies.

Meta-Categories recalculated according to pie selction

Essentials of Bioset Query Processing and Results

The NextBio search engine enables any set of genes or proteins to be searched against all other studies in order to identify significant correlations with other data, providing an easy-to-understand biological context for the otherwise complex data. A researcher can load a set of genes with or without associated statistics into the NextBio search engine and instantly view significant correlations to all studies. Rank-based enrichment analysis prioritizes all Bioset-Bioset correlations to the queried Bioset.

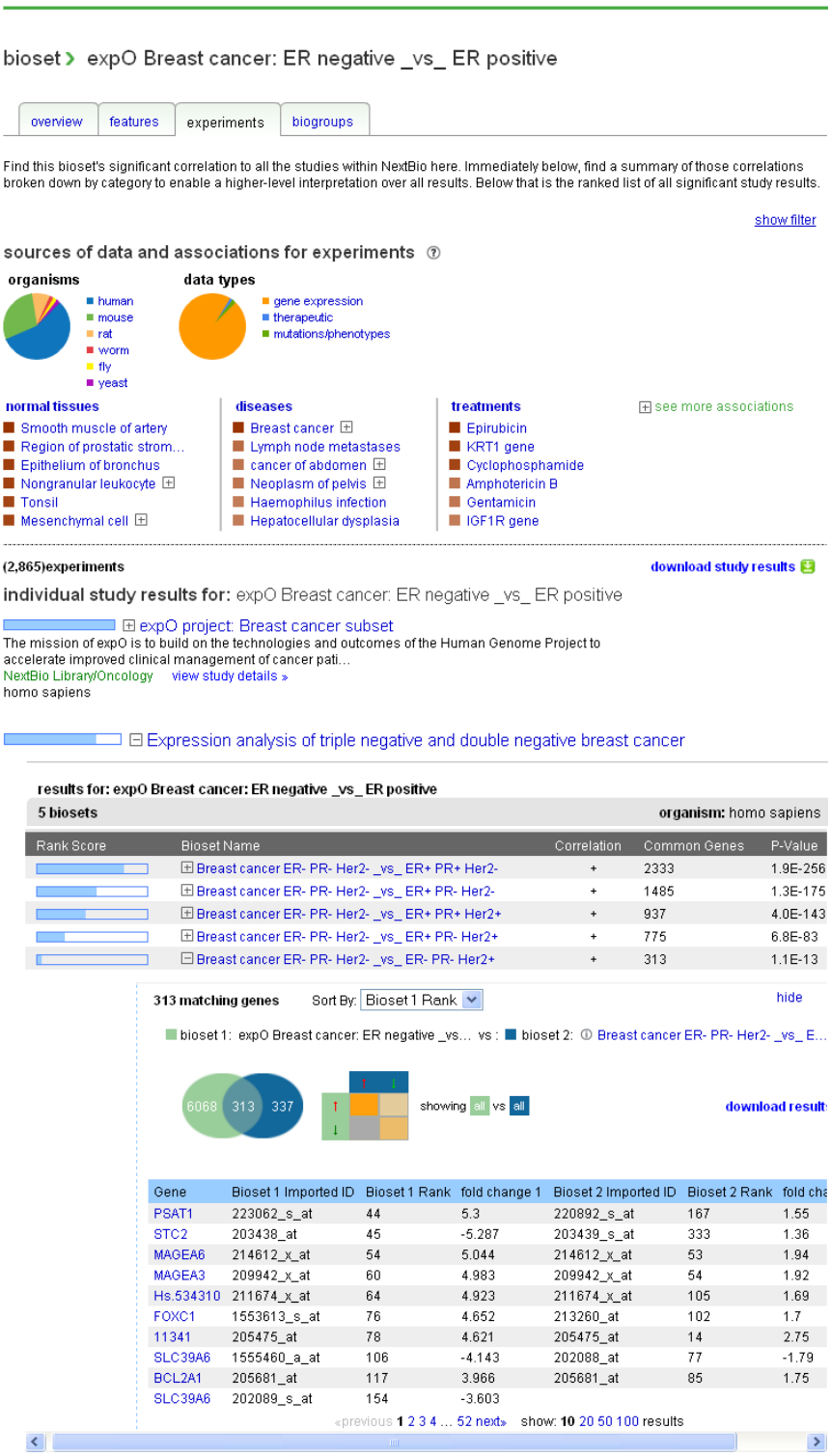
> The correlation score of one gene expression data set B1 (Bioset 1) with another data set B2 (Bioset 2) is determined by evaluating the significance of the overlapping genes between B1 and B2, in contrast with the random distributions within their respective platforms P1 and P2. The null hypothesis is that the distribution of overlapping genes in B1 and B2 are not significantly different from what would be expected randomly given the platforms P1 and P2. The alternative hypothesis is that there are more overlapping genes between B1 and B2 in the upper rankings than would be expected in random samplings from their platforms P1 and P2. To evaluate the statistical significance of B1B2 overlap contrasting against P1P2 overlap, a top to bottom scan is performed in the rank ordered gene lists B1 and B2. The overall correlation score is represented by the maximum score, adjusted by a multiple testing correction. Few assumptions are made about the Bioset data structure. General comparisons are enabled between data sets with small-tens or large-thousands of genes, with or without rankings, between data sets with or without directionalities, and between data sets from different data types, platforms or different species.

If a data set contains directional subsets, up-regulated and down-regulated genes, in the differentially expressed gene sets, then subsets with distinct directionality are automatically recognized and the analysis would indicate whether a positive or negative correlation is discovered. The positive or negative correlation call in Level 2 results (see below) is a generalized one, as you will often see subsets of genes positively and negatively correlating. Statistics are calculated for each pair-wise possibility. A positive correlation would show the same gene(s) up-regulated in both Biosets, or the same gene(s) down-regulated. A negative correlation would show the same gene(s) having opposite expression directionality, like up-regulated in one Bioset and down-regulated in the other. The 2x2 Correlation Table, found at the Level 3 results (see below), is interactive: filter by clicking on a particular quadrant, for say up-regulation across both Biosets will filter the gene table to those select genes.

As shown in the Bioset Search example below, meta-categorization results are displayed directly below the Bioset description section. Bioset-Bioset correlations are calculated for all data sets in the system and presented in the bottom ‘Individual Study Results’ section. Studies are prioritized by significance of Bioset-Bioset correlations. For instance, the Study containing the Bioset having the lowest p-value will be at the top of the list. Individual Study results are categorized into various associations in the ‘sources of data and associations for experiments’ section. ‘Normal tissue’, ‘Disease’ and ‘Treatment’ categories are derived from meta-analysis of the experimental design tags specific to each associated Bioset. Meta-categorization for “normal tissues” and “diseases” exclude data from cell lines, as it is assumed that cell lines don’t represent true models of normal tissues or diseases. Meta-categorization for “Biogroups” is determined from associations to the gene set data within the single queried Bioset.

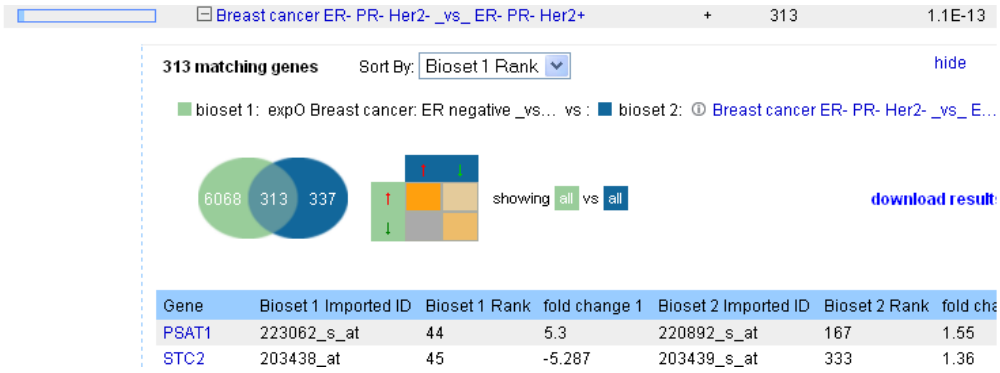
Bioset Query Example, link out to:
http://www.nextbio.com/b/home/home.nb?id=14568&type=bioset&currentTab=experiments

Bioset search results show ranked correlations to all other studies in NextBio. In this example, the correlations from a user selected study on clinical estrogen receptor status are shown contrasted to an independent study of expression changes across combinations of estrogen receptor, progesterone receptor and Her2 status.



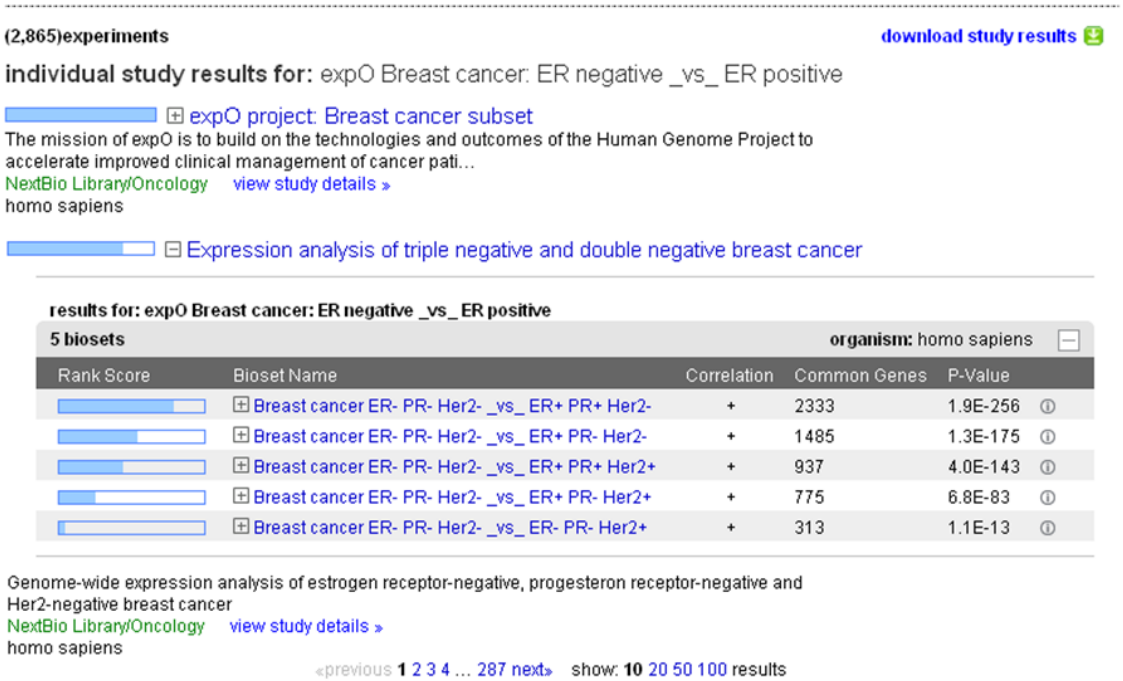
Venn Diagram and Correlation Table Explanation

Level 3 results show two micrographs that visually illustrate Bioset –Bioset relationships. The Venn Diagram outlines the overlap of common genes between Biosets. The directional relationships in the correlation table are computed individually and colored by the strength of the p-value; lower p-values yield darker shades of orange and gray is null. Quadrants of the correlation table are interactive: mouse-overs reveal individual statistics and clicking on one will accordingly subset the results table below the micrographs. Enterprise customers can download tabular results accordingly.



Rank Score Explanation

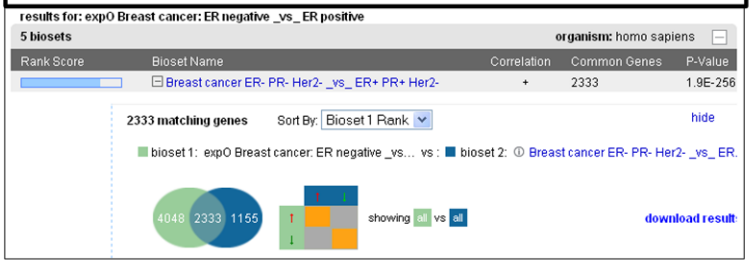
Blue bars visually represent Rank Scores decreasing with relevance according to correlations between the query and the highest ranking bioset in a study. The magnitude is an aggregate score using the –Log(p-value) of the directional combinations, as explained below.

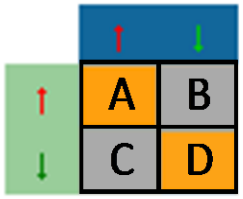


Bioset-Bioset Correlation Scoring

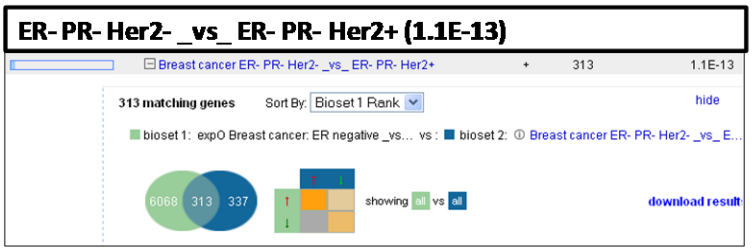
Correlations to Bioset 'expO Breast cancer: ER negative _vs_ ER positive'				
	ER- PR- Her2- _vs_ ER+ PR+ Her2- (1.9E-256)		ER- PR- Her2- _vs_ ER- PR- Her2+ (1.1E-13)	
Quad	**p-value**	**-LOG(p-value)**	**p-value**	**-LOG(p-value)**
A	2.10E-131	130.68	9.60E-11	10.02
B	no overlap		3.12E-02	1.51
C	no overlap		no overlap	
D	9.30E-125	124.03	3.40E-05	4.47
Aggregate Score = A-B-C+D		254.71	**Aggregate Score = A-B-C+D**	12.98
10E-AgScore		1.95E-255	**10E-AgScore**	1.05E-13

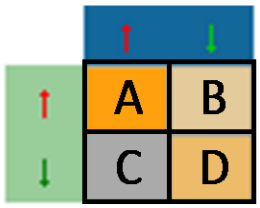
ER- PR- Her2- vs ER+ PR+ Her2- (1.9E-256)





ER- PR- Her2- _vs_ ER- PR- Her2+ (1.1E-13)





Essentials of Biogroup Query Processing and Results

In an expression profile generated from a DNA microarray study, significant alterations in gene expression might manifest at the level of biological pathways or co-regulated gene sets, rather than at the level of individual genes. A gene set enrichment analysis strategy can be used to detect coordinate changes in the expression of groups of functionally related genes, such as metabolic pathways, transcriptional programs, and stress responses, that are distributed across an entire expression profile and subtle at the level of individual genes (Mootha et al. 2003; Subramanian et al. 2005; Sweet-Cordero et al. 2005). A Biogroup, in NextBio terms, represents a functionally related set of genes or proteins, such as a pathway, a gene ontology group or a protein family.

> To evaluate the significance of a given gene set G (Biogroup) in a gene expression result Bioset B, we identify gene set G members and contrast their specific rankings in B against the background distribution in the platform P that is used in the gene expression experiment. Genes differentially expressed under the experimental design are ranked according to a selected criteria such as magnitude of changes resulting in a rank-ordered gene list in the result set B. Significance scores are evaluated as follows: the null hypothesis is that the distribution of gene set G members is not significantly different from what would be expected randomly given the platform P; the alternative hypothesis is that the gene set G members are enriched in the upper rankings among result set B compared with what would be expected in random samplings from the platform P. A top to bottom scan is performed in the rank-ordered gene list B, to compute the significance score of B contrasting against P at each rank R. The overall significance score is represented by the maximum score, adjusted by a multiple testing correction.

Direction of gene or protein expression changes within a data set is automatically recognized to form subsets for analysis (different directions are analyzed separately). A general approach was developed that makes few assumptions about the data structure, enables comparisons between Biogroups and Biosets with small (tens) or large (thousands) numbers of genes, does or does not include rankings, is compared against data with or without directionalities, and originates from different data types, platforms or across different species.

Meta-categorization results across all relevant studies retrieved from a Biogroup search are categorized into top associations in the section ‘sources of data and associations for experiments’. ‘Normal tissue’, ‘Disease’ and ‘Treatment’ categories are derived from the tags assigned to data in the ‘individual study results’. The color intensity indicates the significance relative to the top-ranked result. Click on any source concept (‘category’) to filter ‘individual study results’ to that concept. Meta-categorization for “normal tissues” and “diseases” exclude data from cell lines, as it is assumed that cell lines don’t represent true models of normal tissues or diseases.

Biogroup Query Example, link out to:

http://www.nextbio.com/b/home/home.nb?q=angiogenesis

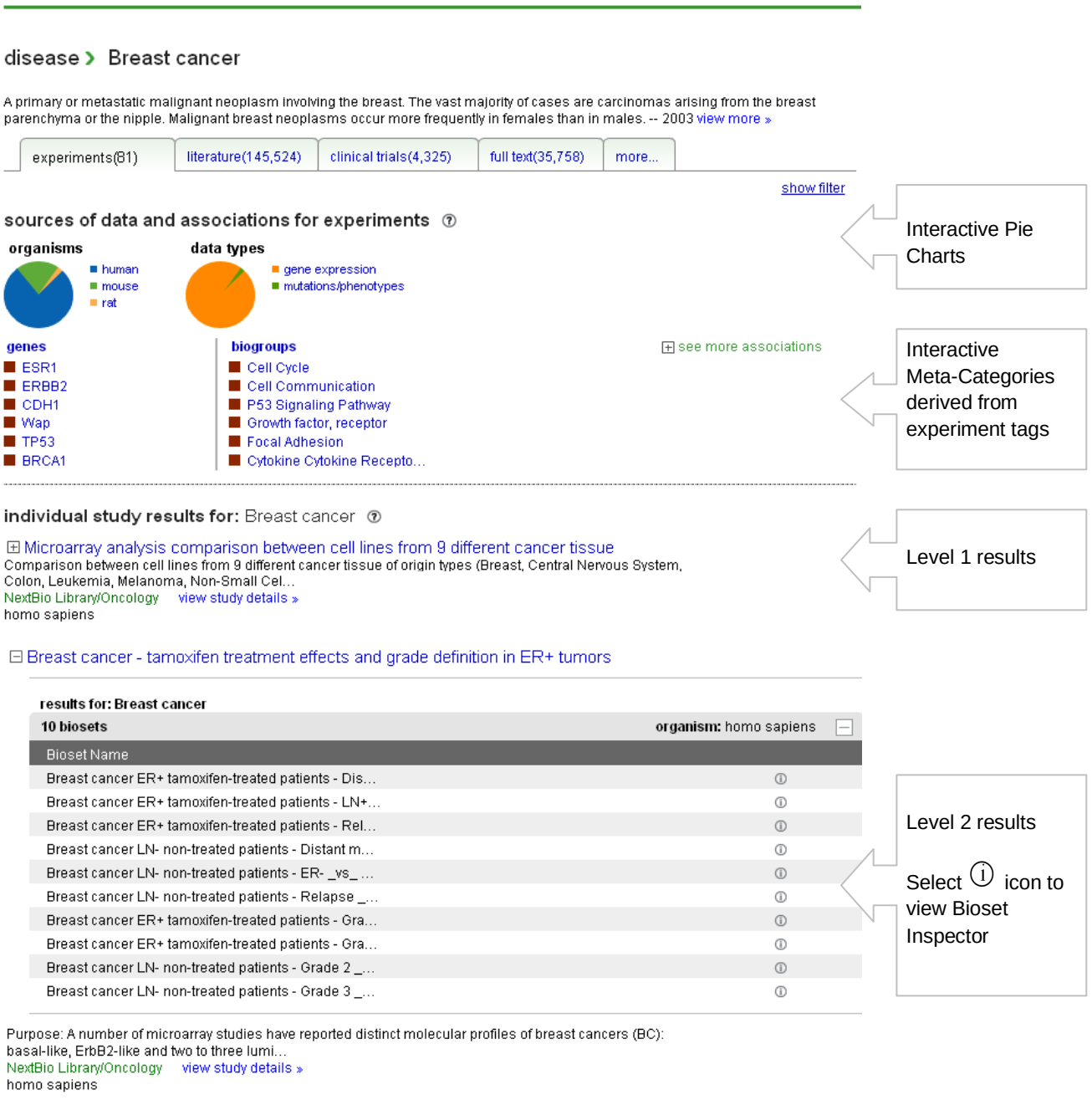
'Angiogenesis' is a Gene Ontology concept comprised of 114 human, 134 mouse, 123 rat and 2 fly genes. Biogroup search results show ranked correlations to studies in NextBio. In this example, the correlations from a user selected study on breast carcinoma are shown contrasted to 'Angiogensis' under the 'experiments' tab.

Essentials of Disease Query Processing and Results

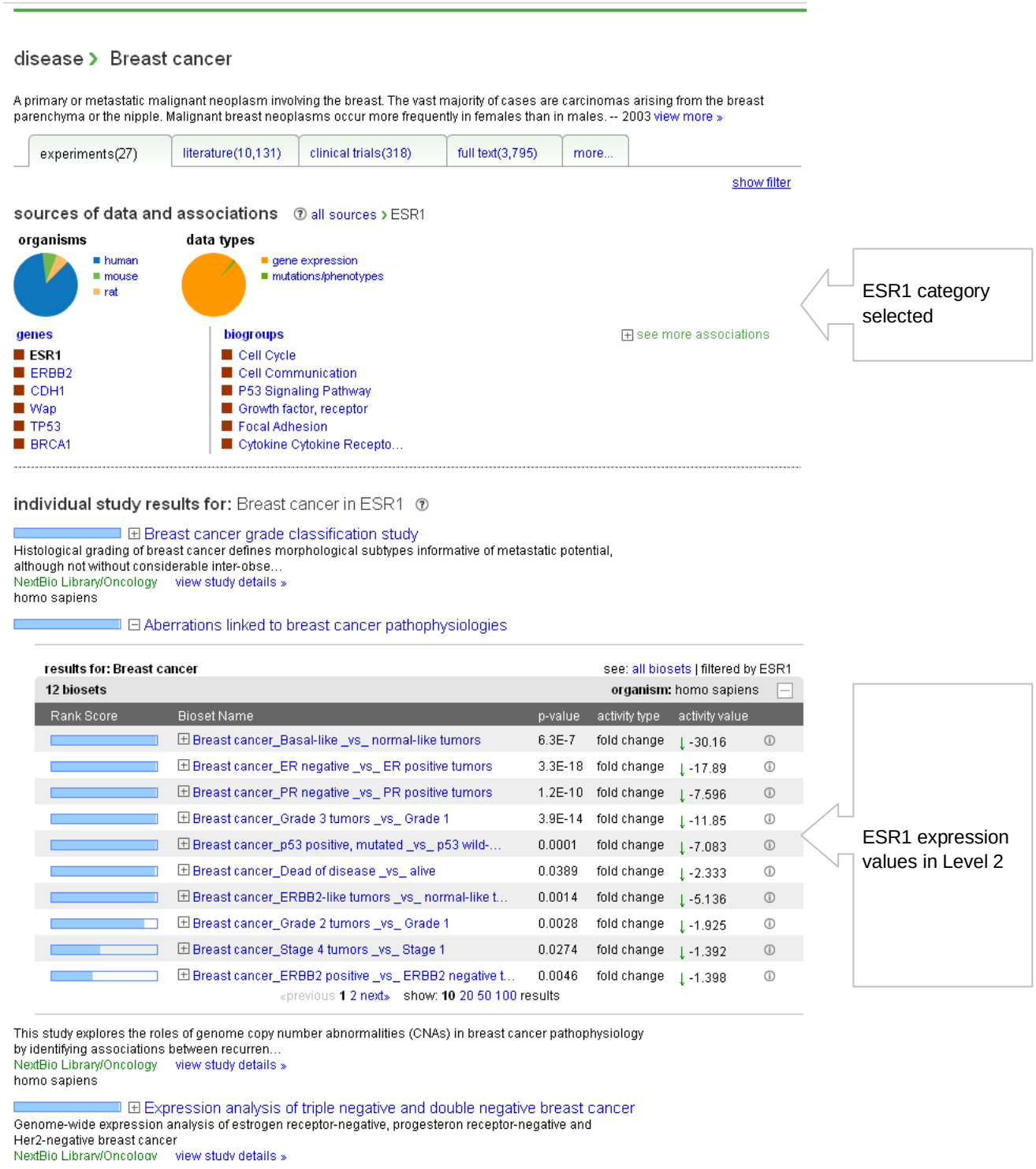
A disease query will return 'individual study results' based on tags under the 'experiments' tab. Meta-Categories are derived from the strength of Gene and Biogroup associations from the individual studies.

Disease Query example, link out to:
http://www.nextbio.com/b/home/home.nb?q=breast%20cancer

The search result shows all Studies where one can scan Study titles and short descriptions, or choose to **view study details**. Level 2 results list the Biosets. Selecting the ⓘ icon across from Bioset names will show a Bioset Inspector page with overview, feature dataset, experimental Study and Biogroup correlations.



Choosing a category from the 'sources of data and associations for experiments' will apply this concept as a filter and display relevant results. This is shown below after selecting the highest ranking, data-driven gene category 'ESR1'. Pie charts will change accordingly to represent only studies where the ESR1 gene is significantly changing or a member of validated data types. Combinations of pie chart and meta-categories are possible, for instance one could choose to limit results to human studies using the organisms pie and 'P53 Signaling Pathway' from the Biogroups meta-categories.



Essentials of Tissue and Organ Query Processing and Results

A tissue query will return 'individual study results' based on tags. Meta-Categories are derived from the strength of Gene and Biogroup associations from the individual studies.

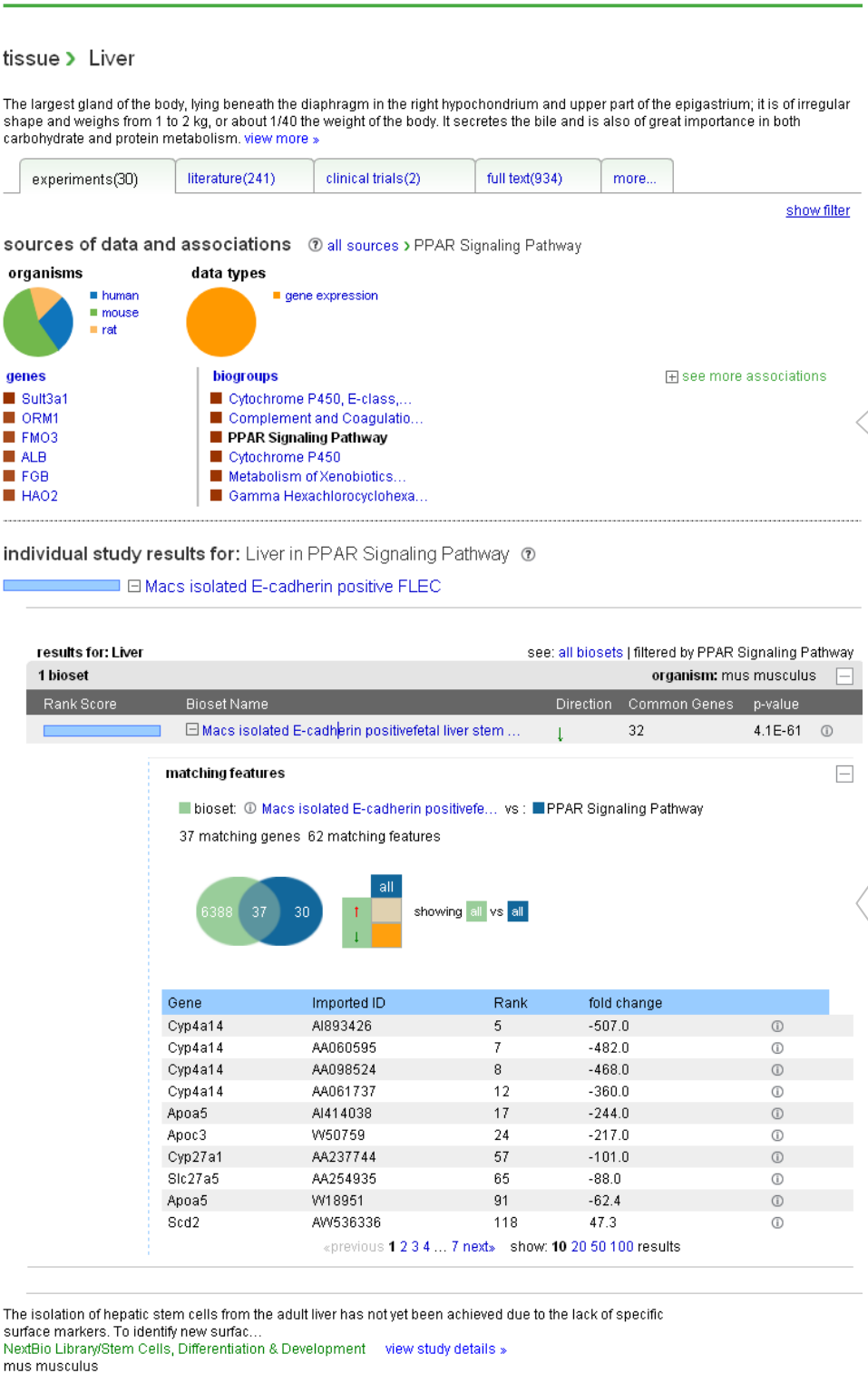
Tissue Query Example, link out to:

http://www.nextbio.com/b/home/home.nb?q=liver

Search results show all liver Studies where one can scan Study titles and short descriptions, and choose to **view study details**. Level 2 results list the Biosets, by selecting ⓘ icon across from Bioset names the complete Bioset description, the dataset and other information can be reviewed.

Choose a category from the ‘sources of data and associations for experiments’ to apply a concept as a filter and display relevant results, as shown below for ‘PPAR Signaling Pathway’.

http://www.nextbio.com/b/home/home.nb?q=liver#bgf=199591&cat=Biogroups&subcat=PPAR%20Signaling%20Pathway&tab=data&tf=PPAR%20Signaling%20Pathway



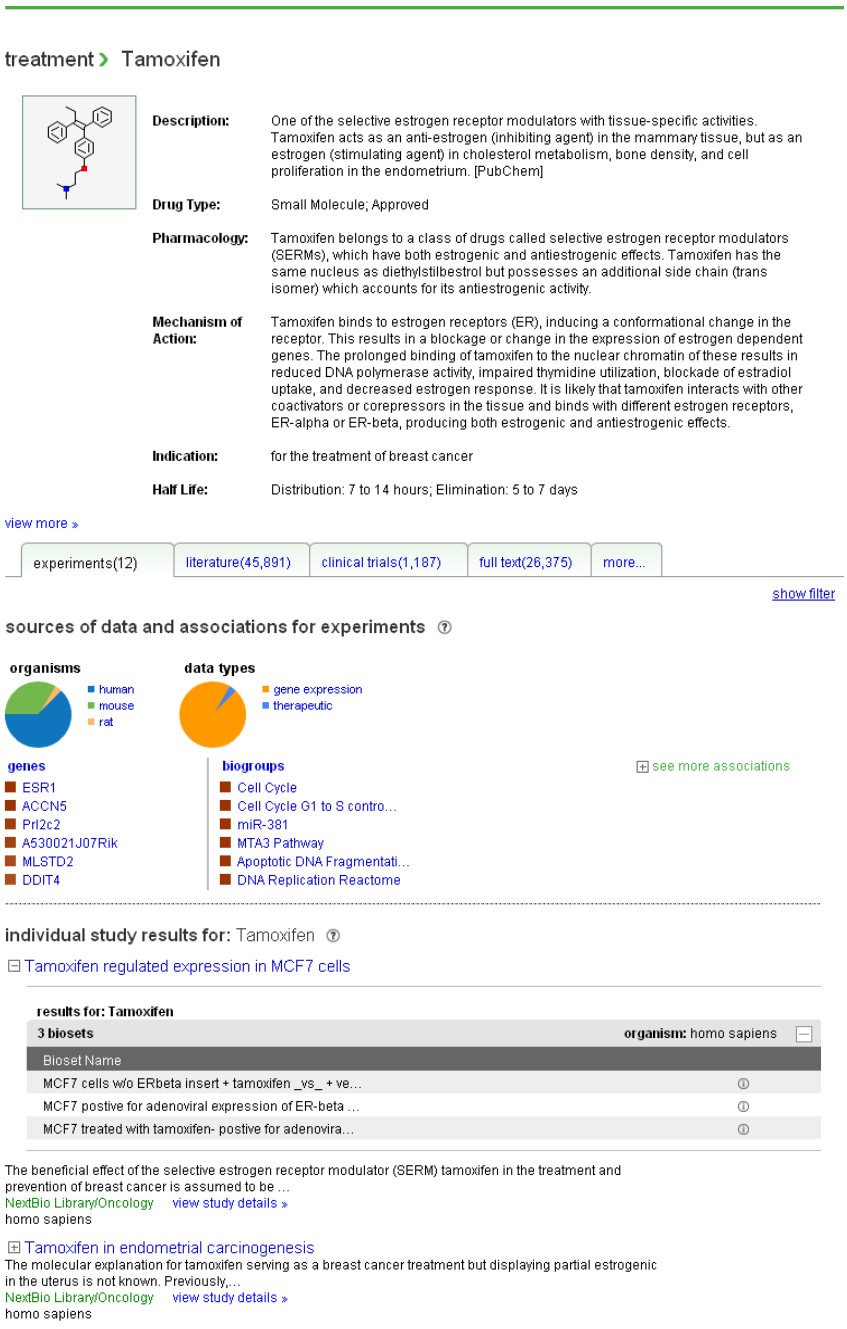
Essentials of Treatment Query Processing and Results

A treatment or drug query will return 'individual study results' based on tags. Meta-Categories are derived from the strength of Gene and Biogroup associations from the individual studies.

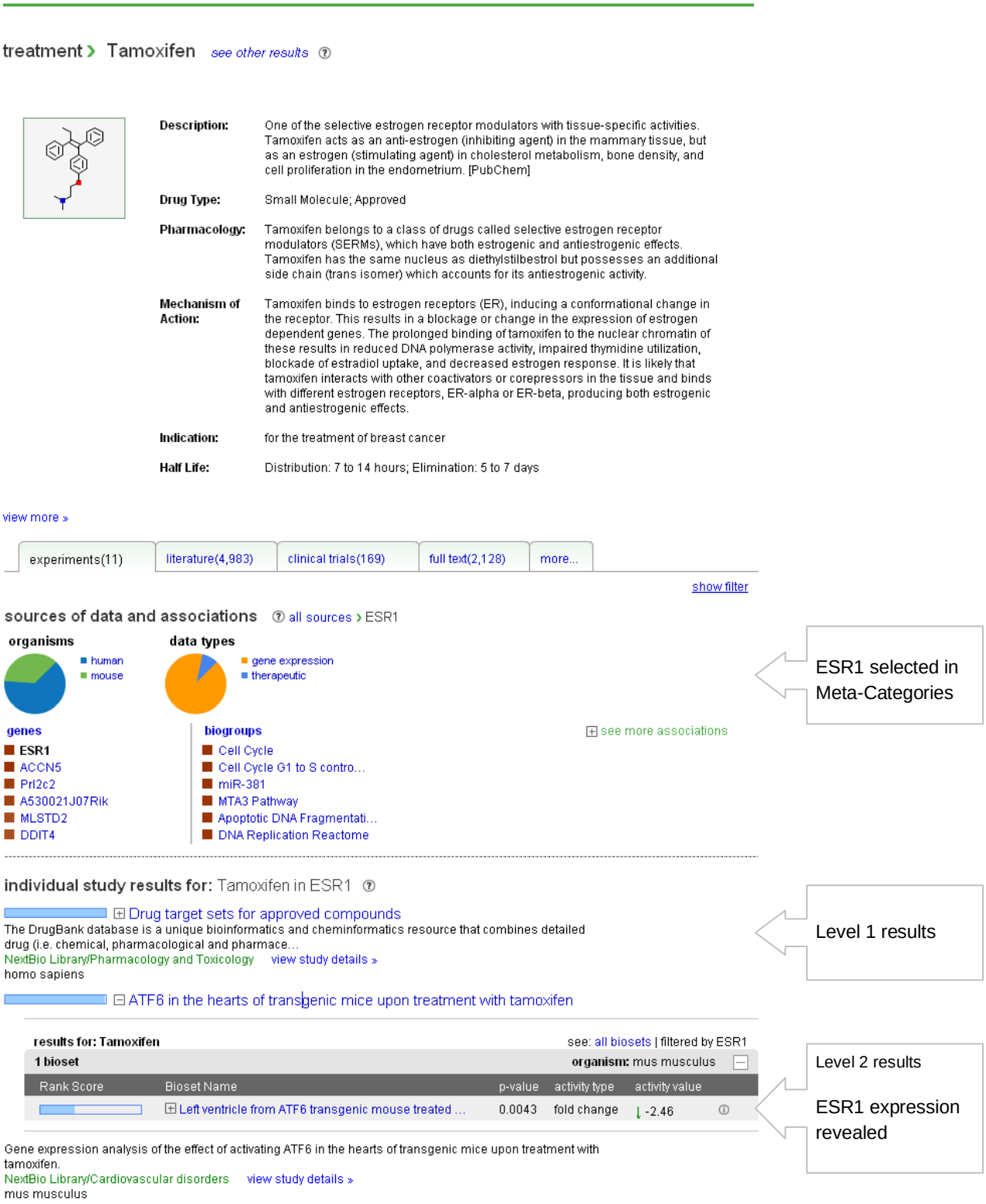
Treatment Query Example, link out to:

http://www.nextbio.com/b/home/home.nb?q=tamoxifen

The search result shows all relevant Studies where one can scan Study titles and short descriptions, and choose to **view study details**. Level 2 results list the Biosets, by selecting ⓘ icon across from Bioset names the complete Bioset description, the dataset and other information can be reviewed.



Choose a category from the ‘sources of data and associations for experiments’ to apply a concept as a filter and display selected results, as shown below for ‘ESR1’.



Essentials of Advanced Search

Advanced Search is a utility only available to Enterprise customers, it provides users the ability to set up custom queries that include a collection of high interest Biosets or Studies for direct comparison. Biosets or Studies can be a mix from Private User, Private Enterprise and NextBio Public Libraries.

Advanced Search can be performed across Genes or Biogroups. By selecting 'query across all genes' users identify the most highly up- or down-regulated genes across the queried set of Biosets or Studies. By selecting 'query across all biogroups' users identify the most highly correlated functionally related gene sets across the queried set of Biosets or Studies.

There are a number of parameters which are used for computing most relevant genes. The most important two parameters are the activity level of a gene in each Bioset and the specificity (the number of Biosets in which the gene is active). The current Advanced Search query limits users to no more than 50 Biosets or 5 Studies at a time

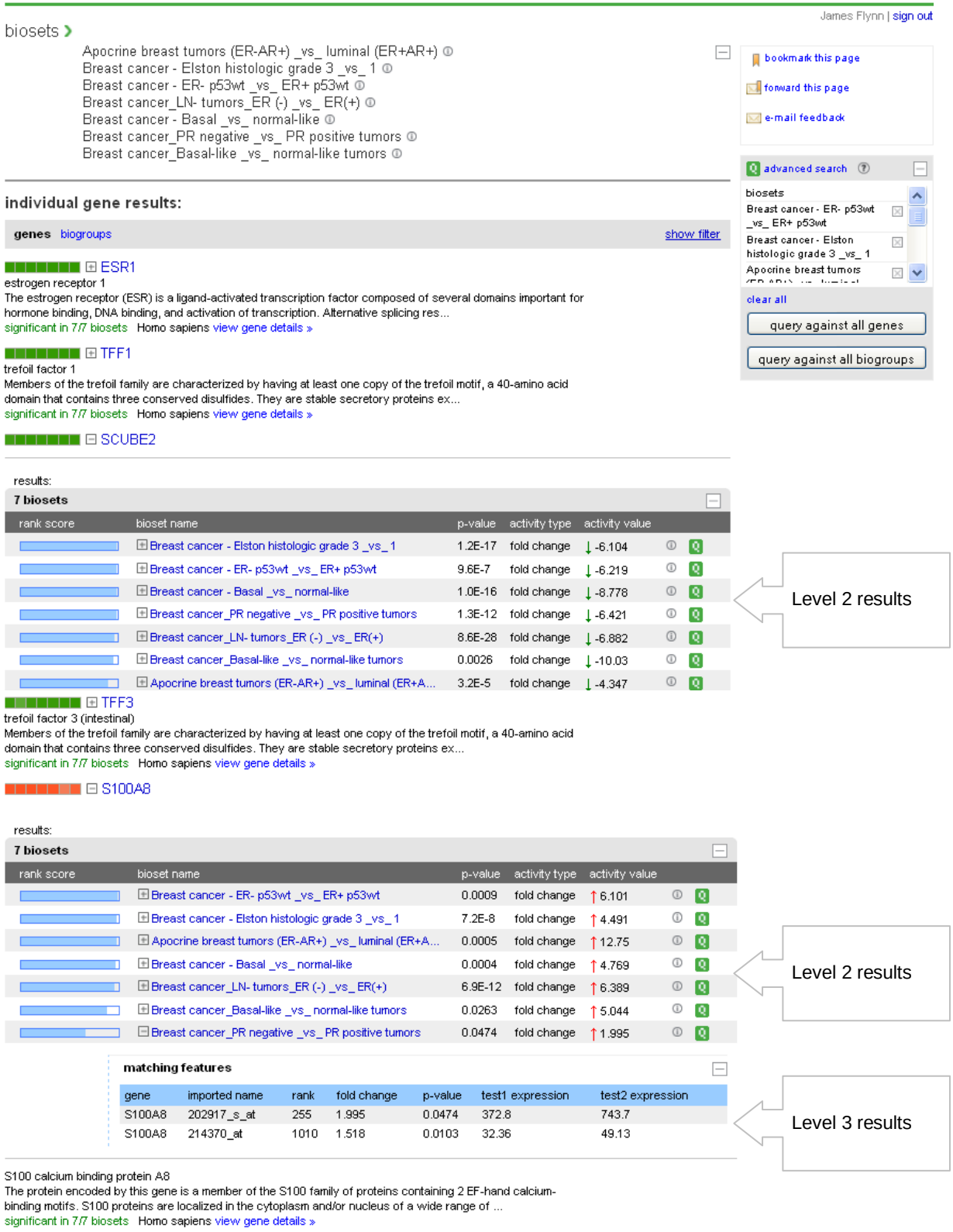
By way of example, let's say I'm a breast cancer researcher that is interested in comparing a number advanced clinical samples for high ranking correlations to important genes and functionally related gene sets. I find and select the following seven Biosets from four independent Studies to broadly validate findings and discover nuances via NextBio Advanced Search.

Bioset Name	**Study ID (content source)**
Breast cancer - ER- p53wt _vs_ ER+ p53wt	GSE4922 (GEO)
Breast cancer - Elston histologic grade 3 _vs_1	GSE4922 (GEO)
Breast cancer_Basal-like _vs_ normal-like tumors	E-TABM-158 (Array Express)
Breast cancer_PR negative _vs_ PR positive tumors	E-TABM-158 (Array express)
Breast cancer - Basal _vs_ normal-like	GSE1456 (GEO)
Apocrine breast tumors (ER-AR+) _vs_ luminal (ER+AR+)	GSE1561 (GEO)
Breast cancer_LN- tumors_ER (-) _vs_ ER(+)	GSE2034 (GEO)

Loading of these Biosets was initiated with a search for 'breast cancer' and viewing individual Studies to determine interest in adding specific Biosets to an Advanced Search. Selecting the Advanced Search  will populate Studies and Biosets with this icon, then selecting the icon associated with the data of interest will place into the bin at right of screen. This is repeated until all data sets of interest are loaded. Then choose whether to do Advanced Search for 'Query Against All Genes' or 'Query Against all Biogroups'. The following screenshots capture the results of the two.

Advanced Search Example - 'Query Against All Genes'

The Matrix Diagram to the left of every 'individual gene result' depicts a box for each of the seven Biosets queried, they are colored and shaded by the general direction and strength of association (red = up-regulated, green = down-regulated). Level 1 results show rank ordering of Genes. Level 2 results show the individual expression results for that gene in each Bioset and Level 3 shows additional statistical information.



Advanced Search Example - 'Query Against all Biogroups'

The Matrix Diagram to the left of every Biogroup in 'individual biogroup results' depicts a box for each of the seven Biosets queried, they are colored and shaded by the general direction and strength of association (red = up-regulated, green = down-regulated). Level 1 results show the breadth of Biogroup source types: Broad MSigDB - Regulatory Motifs, Broad MSigDB - Canonical Pathways, InterPro common protein domains, TargetScan miRNA targets DB and others. Level 2 results show the individual Bioset associations to that Biogroup and Level 3 shows the Venn Diagram, interactive Correlation Table and the common genes.

Essentials of Data Import and Tagging

A detailed exercise describing 'Data Import and Tagging Best Practices' is available upon request.

Import files may consist of a simple list of genes or be more content rich, containing preprocessed data and the results of statistical analysis. Import files can be in .txt, .csv or .xls file formats. NextBio allows users to import standard numerical and statistical columns (fold change, p-value, score, rank, correlation) and a maximum of 5 user-defined columns that will be carried through for reference. The minimum requirement is that your file contains a column name and a list of recognizable identifiers. NextBio recognizes most public and standard commercial platform identifiers: including, NCBI Gene IDs, NCBI Official Gene Symbols, NCBI accession numbers, ENSEMBL IDs, RefSeq IDs, IPI IDs, and platform IDs from Affymetrix, Illumina, Agilent and GE Healthcare. If a custom platform is being used, then we recommend using NCBI Official Gene Symbols or RefSeq IDs if splice variant probes are present. Complete lists of allowed column names and supported platforms are described at the end of this document.

NextBio maps the Imported ID to the Gene Index, recognizes numerical column types and converts one of these to ranks. The primary columns used for ranking are the Gene Name and directional Fold Change columns which are minimally recommended for most purposes. If Fold Change values are not available, then a simple list of feature identifiers can be used, or columns containing other statistical measures can be added that will fully engage the rank-based correlations done by NextBio. If more than one standard statistical column is present, NextBio picks one for ranking in the order:

Rank → Fold change → P-value → Score

Example of a standard NextBio Import format:

Gene Name	Fold Change	P-Value	Test Expression	Control Expression
34756_g_at	-2.50	0.0459	22.9	18.4
39805_at	1.43	0.0357	139.6	113.7
31608_g_at	1.94	0.0202	99.5	69
36976_at	-1.93	0.0203	34.1	17.6
38612_at	1.68	0.0151	50.6	34.2
41724_at	-4.44	0.0117	329.4	266.2

Data tagging is an optional, but very important component of NextBio. Vocabulary sources that support tagging include the following: Gene Index (developed by NextBio), dbSNP (SNP rs#), SNOMEDCT (disease), Foundational Model of Anatomy (tissues/organs), PubChem/DrugBank (compounds) and ATCC cell lines (biosource). Although adding tags is not necessary, it is highly recommended that you add tags where applicable for sample source, experimental design, tissue, disease, treatment and genetic modification. In general, tagging should only describe the main attributes of the experimental design, not the experimental results or observations (don't tag data with a top gene found to be interesting). Tagging is an important process which provides semantic structure to your data. Search results are significantly improved once the data is tagged. Furthermore, tagging can be used to bring up your study within an appropriate context, or for additional computations available to Enterprise users. It also helps your colleagues and collaborators quickly understand the biological background of the experiment. A Tagging Cheat Sheet is on the following page.

NextBio uses proprietary rank-based statistics to compute associations between the data you import and all other experimental data. In that way, you can place your experimental results within the context of the world's experiments to validate your study, discover novel associations and trends, and design new experiments.

Tagging Concepts

Tagging Cheat Sheet

Tags associate terms from biomedical ontologies with studies published to NextBio

Major Tag Categories	
Biosource	(e.g. *primary tissue, MCF7 cell line* or *cell line* [for non-ATCC cell lines])
Biodesign	(e.g. *disease vs. normal*)
Tissue	(e.g. *liver, lung*)
Disease	(e.g. *Alzheimer's disease*)
Compound	(e.g. *Aspirin*)
Gene	(e.g. *ESR1*)
Genemode	(e.g. *Gene knockdown*)
Biogroup	(e.g. *myoblast differentiation*)

Biosource
Blood fraction
Cell culture, primary
Cell line (provide specific cell line name if available)
Cells, primary
Embryo
Laser capture microdissection
Primary tissue - FFPE (formalin-fixed, paraffin-embedded)
Primary tissue - fresh or fresh-frozen
PBMCs
Whole blood
Whole body
Whole organ

Biodesign
circadian time course
developmental time course
disease resistant vs. susceptible
disease vs. disease
disease vs. normal
drug dose response
drug resistant vs. sensitive
mutant vs. mutant
mutant vs. wildtype
normal vs. normal
time course
treated vs. control
treatment vs. treatment

Tissue
Define the most specific tissue/organ/cell type used as a sample for the experiment, vocabulary based on Foundation Model of Anatomy (FMA).

Disease
Assign only where applicable, vocabulary is based on a modified curated version of SNOMEDCT.

Compound
Assign name of compound used. If it's a custom compound, be sure to specify the correct Biodesign tag to indicate that a treatment was applied.

Gene
Assigned when a gene in a given sample was modified or served as the key differentiating marker between experimental groups (e.g. ER+ vs. ER- breast cancer).

Biogroup
Biogroups consist of sets of genes related by function and should only be used as tags when it's appropriate and when no other vocabulary above provides relevant terminology.

If you've tagged with a Gene or Protein Treatment , then add a Genemode Tag:

Genemode
Abnormal gene function
Cell marker (or marker gene)
Conditional expression (or inducible expression)
Gene amplification (or gene duplication)
Gene deficiency
Gene dominant-negative overexpression
Gene Insertion
Gene Inversion
Gene knockdown (or gene silencing or siRNA knockdown)
Gene knockout (or null mutant or gene deletion)
Gene mutation (or point mutation)
Gene overexpression (or transgenic or knock-in)
Protein treatment

My NextBio Functions
(profiles, bookmarks, studies and account)

A pioneering goal of NextBio is to ‘democratize’ data, putting researchers in touch with results that can impact the success of their projects broadly across an organization. Toward this, concepts in ‘social networking’ and ‘collective intelligence’ have been developed for importing data, sharing with colleagues, saving key query results, and communicating interesting details.

My Profiles document a user’s background, projects and publications, so others can reach out to collaborate with professionals in their organization. The more information one adds to their Profile will increase their value to others searching for expertise.

My Inbox is where messages are composed and stored to and from colleagues.

My Contacts is used to find, invite and maintain contacts for a user's personal online scientific community.

My Projects is where communities are developed around data of common interest. Projects are comprised of Studies and can be shared with any number of special interest Groups.

My Studies is where users conveniently store imported data, where they can be edited, correlated to other studies and shared with others.

My Groups is where users collaborate and communicate with a small group of people or a large number of users with similar research objectives. Upon Group creation, Projects are shared that contain specific Studies. Access controls are flexibly set by Group Founders.

My Bookmarks is a storage location where important findings can be saved for future reference and emailed to other NextBio Users, so they can immediately see your discoveries.

My Account is where passwords can be changed and the number of search results per page can be modified.

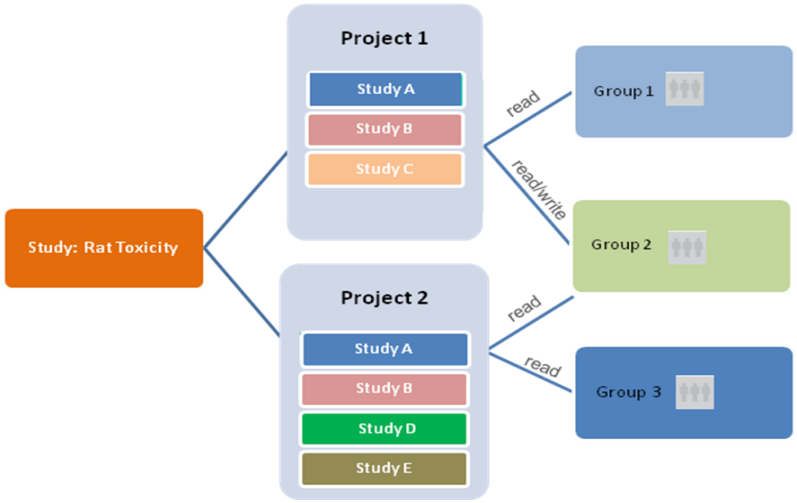
My Privacy is where users control who can see their profile at an Enterprise.

Collaboration & Data Sharing Using Libraries, Projects and Groups

A more detailed document is available upon request.

Combining NextBio's Project and Group functionality offers enterprise users a simple and powerful way to collaborate and share studies of interest with other users. Data sharing occurs at the project level. In order to create focused collections of data, Projects can contain multiple studies selected from one's own, one's enterprise, those shared with you by others and public domain studies. As new Studies are added to a Project already shared with others, they will automatically inherit access to the newly added Studies.

The diagram below outlines the flow of study sharing within NextBio. In this example a user's "Rat toxicity" study can be added to multiple Projects (1 & 2), which in turn can be shared with multiple Groups using different access privileges. As you add more Studies to Project 1 and 2 all users in Groups 1, 2 and 3 will automatically inherit access to the Studies. Studies may also be added to projects created by others as long as this permission has been granted using the group sharing mechanism.



Studies can be added to Projects in two ways:

1. From the **Study Inspector**'s "Overview" page
2. From the **Project Inspector**'s "Studies" page

Study Inspector can be accessed in different ways – by clicking on "Study Details" link on the query results page or by following the Study links from the Library or Project Inspector pages. A blue link "add to project" will enable you add a study into an existing project. If you want to create a new Project you can go to the "Projects" tab of "my nextbio" and select the "create a new project" option.

Project Inspector can be accessed from the "my nextbio/projects" page by selecting a specific project of interest. Single or multiple studies can then be selected and shared using the "add to another project" button or by selecting the sharing icon to the right of each study.

Project Permission Types
All project sharing occurs at the Group level. Within "my nextbio/projects" tab you can select projects that you want to share with other groups. Permissions are applied at the project level. There are two types of permissions:

1. Read - Users can access the content of the project, but can't add new studies to it
2. Add/Read - Users can access the content of the project, as well as add new studies to it

NextBio Libraries and Enterprise Administrators

Administrators control projects and project permissions from within the Enterprise Library which can subsequently be accessed by all users within an organization. Importantly, Studies within an Enterprise Library are used in advanced meta-categorization computations, while those in individual User Libraries are not.

Currently, the basic settings for Enterprise Administrators are controlled by the NextBio Administrator. The NextBio admin can quickly set up projects within the enterprise library and can assign permissions for users to allow them to import data or share data within the Enterprise Library. Note: Within the next few weeks a separate NextBio Administration module will be released. This new module will enable NextBio Enterprise Administrators to define and manage security settings, enterprise library projects, permissions and many other system parameters.

NextBio Libraries are used to browse Public studies or organize, edit and browse your organization's internal data. There are three Library types:

NextBio Public Library contains all public data studies organized into high level projects allowing navigation through all the NextBio public content. You can pick Studies or Biosets of interest and set up Advanced Queries or just browse through available content. Studies and Biosets in the Public Library cannot be edited.

Company Enterprise Library contains studies and projects proprietary to your specific organization. Only users within your company have permission to access it. In order to move data from your private project into this library with organization-wide access users have to have special Administrator permission.

Company User Library hosts your own private Studies. Only you can access these unless you have added them to specified Projects and Groups. When you import your data this is the default library to which your data is saved. If you would like to move your data to the organization-wide Enterprise Library providing access for all Enterprise users discuss it with your NextBio Administrator.

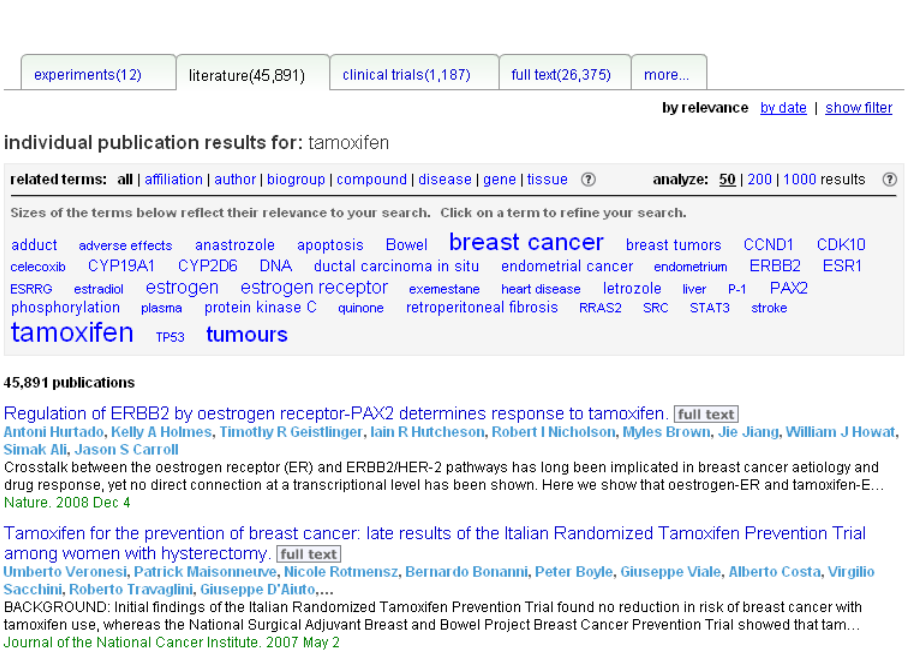
Essentials of Literature Search in NextBio

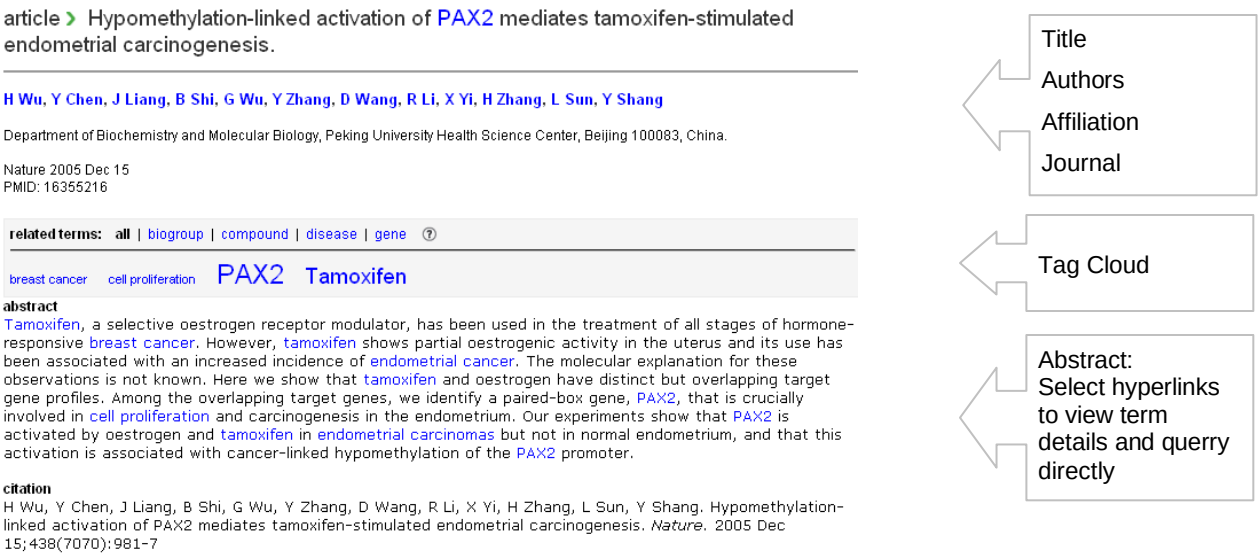
NextBio indexes over 18 million abstracts from the public domain. For literature search, NextBio uses a number of heuristics, including:

- an extensive ontology with relationships between terms and synonyms;
- a customized domain-specific stop word list and analyzer that emphasizes ontology terms;
- the authority of the journal where the paper was published; and
- the date of publication

Embedded hyperlinks to key terms in the citation are semantically linked to NextBio search allowing one to jump out from the citation to search results across the entire NextBio repository. For instance, read what authors say about a gene, then instantly find real experimental data related to its expression across the entire repository. NextBio analyzes literature and clinical trials using all of its ontologies and synonyms, so we are able to, for instance, connect any two trials even though they use different synonyms for the same compound.

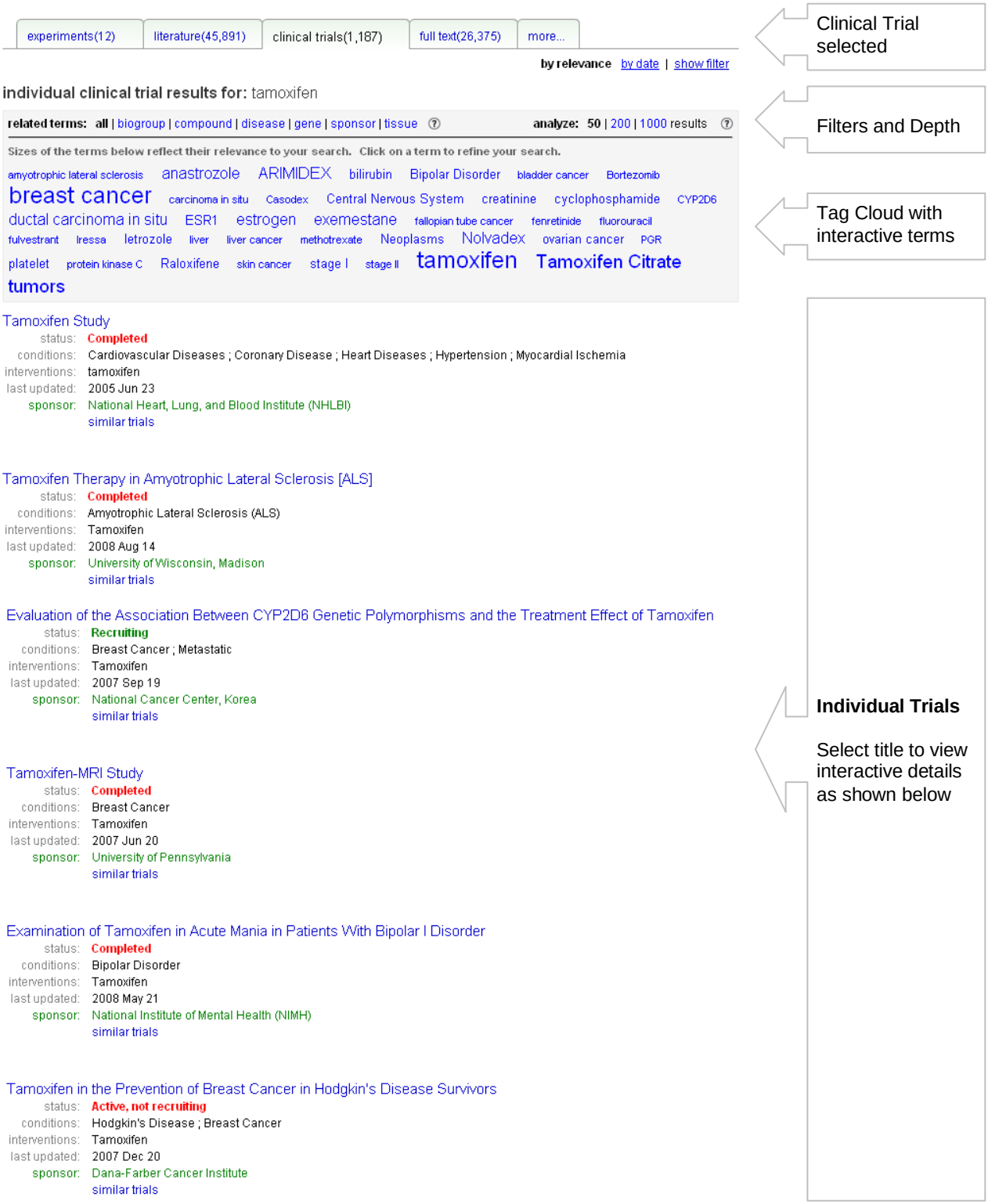
Literature Search Examples





Essentials of Clinical Trials search in NextBio

NextBio indexes over 56,000 clinical trials documented in the public domain via www.ClinicalTrials.gov. A search for tamoxifen clinical trials is shown. The Tag Cloud highlights a list of relevant terms extracted from analyzing the results. Tags are arranged alphabetically, larger font size is used to denote more relevant terms, those that are cited in more trials. Individual trials are listed below the tag cloud. Selecting a trial title will open trial details, as shown in the second screenshot.



Rich details and concept term hyperlinks complement information in NextBio literature and experiments.

clinical trial › Evaluation of the Association Between CYP2D6 Genetic Polymorphisms and the Treatment Effect of Tamoxifen See similar clinical trials »

related terms: all | compound | disease | gene | tissue

Sizes of the terms below reflect their relevance to your search.

albumin AST basal cell carcinoma bilirubin breast cancer creatinine CYP2D6 ESR1 Megace PGR plasma platelet pleural fluid stage IV tamoxifen

purpose

Primary objectives of this study is to evaluate the effects of CYP2D6 genotypes on time to progression after tamoxifen treatment in pre- or postmenopausal women with metastatic breast cancer. Furthermore, we will evaluate the effects of CYP2D6 genotypes on clinical benefit and response duration to tamoxifen administration in pre- or postmenopausal women with metastatic breast cancer and also evaluate the effects of CYP2D6 genotypes on the steady state plasma concentration of tamoxifen and its metabolites

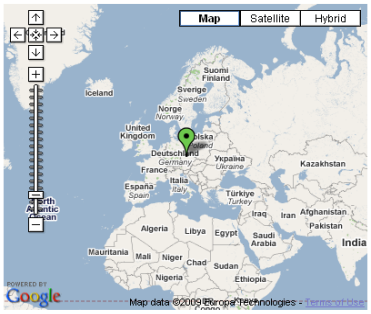
status
Recruiting

conditions
- Breast Cancer
- Metastatic

interventions
- Tamoxifen

phase
Phase 2

study type
Interventional



click on markers to display additional information

study design
Treatment, Non-Randomized, Open Label, Uncontrolled, Single Group Assignment, Efficacy Study

official title
A Clinical Study for the Evaluation of the Association Between CYP2D6 Genetic Polymorphisms and the Treatment Effect of Tamoxifen in Patients With Metastatic Breast Cancer

further study details (as provided by National Cancer Center, Korea)

enrollment
32

eligibility

ages eligible for study
18 Years and older

genders eligible for study
Female

criteria
Inclusion Criteria:
- Histologically or cytologically diagnosed stage IV or recurrent breast cancer patients according to American Joint Committee on Cancer (AJCC)
- Positive estrogen receptor or Positive progesterone receptor.
- Females at least 18 years of age.
- Prior radiation therapy is allowed as long as the irradiated area is not the only source of measurable disease
- Prior hormone therapy less than 2.
- No history of Megace medication for recent 28 days
- Performance status of 0, 1 and 2 on the ECOG criteria
- Clinically measurable disease, defined as bidimensionally measurable lesions with clearly defined margins on x-ray, CT scan, MRI or physical examination. Lesions serving as measurable disease must be at east 1 cm by 1 cm, as defined by x-ray, CT scan, MRI, or physical examination
- Bone only or pleural fluid only disease is included as long as evaluation for clinical benefit is possible
- Estimated life expectancy of at least 12 weeks
- Compliant patient who can be followed-up adequately.
- Adequate hematologic (WBC count 3,000/mm3, platelet count 100,000/mm3), hepatic (bilirubin level 1.8 mg/dL, AST, ALT 1.5xULN, albumin 2.5 g/dL), and renal (creatinine concentration 1.5 mg/dL) function.

- Childbearing women should use non-hormonal contraceptive method

Exclusion Criteria:
- Active or uncontrolled infection.
- Second primary malignancy (except in situ carcinoma of the cervix or adequately treated basal cell carcinoma of the skin or prior malignancy treated more than 5 years ago without recurrence).

contacts and locations

Please refer to this study by its ClinicalTrials.gov identifier: NCT00532454

contacts
Jungsil Ro, MD, PhD 82-31-920-1610 jungsro@ncc.re.kr
locations
Korea, Republic of, Gyeonggi-do - National Cancer Center
status: Recruiting
facility: 809 Madu1-dong, Ilsandong-gu, Goyang-si, Gyeonggi-do, Korea, Republic of, 410-769

sponsors and collaborators

National Cancer Center, Korea

investigators
Principal Investigator: Jungsil Ro, MD, PhD National Cancer Center, Korea

more information

first received
September 19, 2007

last updated
September 19, 2007

ClinicalTrials.gov Identifier
NCT00532454

health authority
Republic of Korea: Institutional Review Board

Essentials of NextBio Integration via APIs

As noted in the section Essentials of NextBio Enterprise Domain, Enterprise customers can access APIs and work with our Engineering Team to:

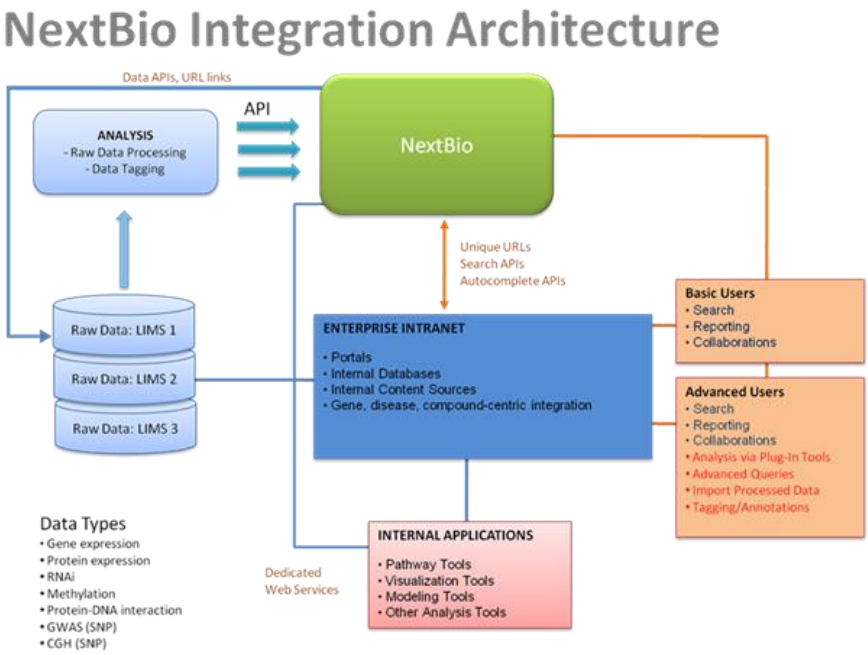
- Provide results of NextBio search within internal apps
- Provide raw/processed data feeds to internal apps/DBs

Search API - Enables data in NextBio to be searchable via a corporate portal or search engine. For instance, a gene search from an internal application/portal links with pertinent results from NextBio will be displayed along with results from internal data.

Data Import API - Automates the uploading of data via a web service or from a secure ftp location. There is no limit to the amount of data that an enterprise can upload.

Autocomplete API - can be included as a service within internal applications as a means to standardize annotations and assists in tag definition when engaging the bulk import process.

Single Sign-On - Seamless access for all Enterprise users without signing-in or registration. User identity is automatically recognized so they could take advantage of personalization features within NextBio.



Supported Platforms

Platform name (77 total)	Example Identifier
Affymetrix GeneChip C. elegans Genome Array	171720_x_at
Affymetrix GeneChip Drosophila Genome 1.0 Array	141200_at
Affymetrix GeneChip Drosophila Genome 2.0 Array	1616608_a_at
Affymetrix GeneChip Human	AA000993_at
Affymetrix GeneChip Human Cancer HC-G110	100_g_at
Affymetrix GeneChip Human Full Length HuGeneFL	A28102_at
Affymetrix GeneChip Human HG-Focus Target Array	1007_s_at
Affymetrix GeneChip Human HG_U133 Plus 2.0	1007_s_at
Affymetrix GeneChip Human HG_U133 Set	1007_s_at
Affymetrix GeneChip Human HG_U133A version [1 or 2]	117_at
Affymetrix GeneChip Human HG_U95A version [1 or 2]	32739_at
Affymetrix GeneChip Human HG_U95A-E version [1 or 2]	1000_at
Affymetrix GeneChip Human Muscle Chip	1001_at
Affymetrix GeneChip Human X3P	1053_3p_at
Affymetrix GeneChip Mouse GNF1M	gnf1m00001_at
Affymetrix GeneChip Mouse MG_430 2.0	1417246_at
Affymetrix GeneChip Mouse MG_430A 2.0	1417246_at
Affymetrix GeneChip Mouse MG_6500A	Msa.14400.0_at
Affymetrix GeneChip Mouse MG_U74	100001_at
Affymetrix GeneChip Mouse MG_U74A	100001_at
Affymetrix GeneChip Mouse Mu11K	aa000148_s_at
Affymetrix GeneChip Rat RAE230 2.0	1368984_at
Affymetrix GeneChip Rat RAE230A	1367452_at
Affymetrix GeneChip Rat RG-U34	A01157cds_s_at
Affymetrix GeneChip Rat RG-U34A	A01157cds_s_at
Affymetrix GeneChip Rat RN-U34	A03913cds_s_at
Affymetrix GeneChip Rat RT-U34	AA108277_at
Affymetrix GeneChip Yeast Genome S98 Array YG-S98	10000_at
Agilent Human 1A G4110A	A_23_P38816
Agilent Human 1A V2 G4110B	A_23_P7262
Agilent Human 1B G4111A	A_32_P838055
Agilent Human CGH 44A G4410A	A_14_P132831
Agilent Human CGH 44B G4410B	A_14_P132831

Agilent Human Whole Genome G4112A/G4112F	A_24_P475014
Agilent Mouse CGH 44A G4414A	| A_53_P168061
Agilent Mouse Development 44K G2519A	A_66_P128131
Agilent Mouse Development G4120A	A_65_P01715
Agilent Mouse G4121A	A_51_P421525
Agilent Mouse Whole Genome G4122A	A_52_P756921
Agilent Rat G4130A	A_43_P22352
Agilent Rat V2 G4130B	A_43_P22352
Agilent Rat Whole Genome G4131A/G4131F	A_44_P292407
Agilent Rhesus Monkey Genome G2519F	| A_01_P000448
Custom C. elegans	171590
Custom D. melanogaster	30970
Custom Human	1
Custom Mouse	11287
Custom Rat	24151
Custom Saccharomyces cerevisiae	850287
GE Healthcare CodeLink ADME Rat 16-Assay	GE1073270
GE Healthcare CodeLink Human Whole Genome	GE479851
GE Healthcare CodeLink Mouse Whole Genome	GE1557164
GE Healthcare CodeLink Rat Whole Genome	GE1100063
GE Healthcare CodeLink UniSet Human 20K	GE85986
GE Healthcare CodeLink UniSet Human I	GE52942
GE Healthcare CodeLink UniSet Mouse 20K	GE110794
GE Healthcare CodeLink UniSet Mouse I	GE36943
GE Healthcare CodeLink UniSet Rat I	GE12207
Illumina Sentrix Human-6 BeadChip	GI_10047089-S
Illumina Sentrix Human-6 v2	4760445
Illumina Sentrix Human-6 v2 (Target ID)	ILMN_89282
Illumina Sentrix HumanRef-8 BeadChip	GI_10047089-S
Illumina Sentrix HumanRef-8 V2	630309
Illumina Sentrix HumanRef-8 V2 (Target ID)	ILMN_25544
Illumina Sentrix HumanRef-8RD BeadChip	GI_10047089-S
Illumina Sentrix Mouse-6 v1	105290026
Illumina Sentrix Mouse-6 v1 (Target ID)	GI_38090455-S
Illumina Sentrix Mouse-6 v1_1	105290026
Illumina Sentrix Mouse-6 v1_1 (Target ID)	GI_38090455-S
Illumina Sentrix MouseRef-8 v1	1450041

Illumina Sentrix MouseRef-8 v1 (Target ID)	scl00227525.1_330-S
Illumina Sentrix MouseRef-8 v1_1	1450041
Illumina Sentrix MouseRef-8 v1_1 (Target ID)	scl00227525.1_330-S
Illumina Sentrix RatRef-12 v1	3170341
Illumina Sentrix RatRef-12 v1 (Target ID)	ILMN_63264
Trex_IAS_RGI3_v	RGICJ31

* Additional platforms are being supported on an ongoing basis as the corresponding data from public and proprietary sources becomes available. If a custom platform is being used, our Engineering Team can be engaged to develop automatic recognition and mapping of those IDs into the system.

THE LIFE SCIENCE SEARCH ENGINE.

NextBio | 3 Results Way Cupertino, CA 95014 | Phone: +1 (408) 861-3610 | Fax: +1 (408) 861-3630 | www.nextbio.com