NCBI resources III: GEO and expression data analysis

Yanbin Yin

Homework assignment 2

- Given the publication http://www.ncbi.nlm.nih.gov/pubmed/19723656, find GEO datasets that are associated with the paper.

- Choose the first dataset and perform a GEO2R analysis.

- Find the top two differentially expressed genes and search their gene symbol or ID at Gene database and explain what their functions are.

- Watch https://www.youtube.com/watch?v=3Wkgr9C9y1U and study my ppt slides.

- Write a report (in word or ppt) to (1) briefly explain why differential gene expression (DGE) is important, the basic idea of DGE analysis, and (2) include all the operations and screen shots

Due on 9/26 (send by email, if there are 2+ files, put them in a zip file; include your last name in the file name)

Gene Expression Omnibus (GEO)

http://www.ncbi.nlm.nih.gov/geo/

GEO is an international public repository that archives and freely distributes microarray, next-generation sequencing, and other forms of high-throughput functional genomics data submitted by the research community.

The three main goals of GEO are to:

Provide a robust, versatile database in which to efficiently store high-throughput functional genomic data

Offer simple submission procedures and formats that support complete and well-annotated data deposits from the research community

Provide user-friendly mechanisms that allow users to query, locate, review and download studies and gene expression profiles of interest (Query and analysis)

Gene expression

- Genome analysis tells us what genes are present, but before we can determine the organism's phenotype, we need to know how those genes are expressed: under what conditions, in what tissues, how much gene product is made, etc.
  - Also, understanding and curing diseases is tied to the analysis of what genes are expressed in disease states.
- There is some progress in deciphering gene control signals, but right now it is more profitable to do experiments and use bioinformatics tools to analyze their results. Later, with some luck, we will be able to make better predictions about gene expression just from sequence data.

Gene Expression and Regulation

Genes are differentially expressed in...

- Different cell types (e.g. muscle cells, fibroblasts)
- Environmental conditions (e.g. heat shock, nutrient deprivation)
- Developmental phases (e.g. embryonic day 10)
- Cell-cycle stages (e.g. G1 phase)
- Disease states (e.g. tumor cells, virus-infected cells)

Gene expression is primarily regulated at the level of transcription

Hence, the number of mRNA copies in a cell for a particular gene is a good indicator of that gene's expression (number of proteins)

Dynamic range of mRNA levels:

- Highly expressed genes can have up to 9400 mRNA copies per cell
- Poorly expressed genes can have < 1 mRNA copies per cell

What Gene Expression Can Do?

- Gene expression is the process by which information from a gene is used in the synthesis of a functional gene product.

- The most fundamental process through which the genotype gives rise to the phenotype.

- For a specific cell at a specific time, only a subset of the genes coded in the genome are expressed.

- Measure of mRNA expression level can
  - Provide a good indicator of corresponding protein expression level
  - Provide insight on the mechanisms of transcriptional and post-transcriptional regulation
  - Shed light on disease mechanisms and treatments

- Transcriptome: The complete set of messenger RNA (mRNA) and noncoding RNA (ncRNA) transcripts produced by a particular cell, cell type, or organism.

High-throughput transcriptome profiling

- Hybridization based approaches: incubating fluorescently labeled cDNA with microarrays. Hybridization signal is measured.
  - Hybridization: the process of establishing a sequence-specific interaction between two strands of nucleic acids through complementary hydrogen bonding between nucleobases.
  - Printed arrays (cDNA microarray)
  - Synthesized arrays (High density oligo arrays)

- Sequencing based approaches: directly determine the cDNA/mRNA sequence. Count is measured.
  - Sanger sequencing-based approaches
    - EST sequencing
    - Serial Analysis of Gene Expression (SAGE)
  - High-throughput sequencing-based approach
    - RNA-Seq

Concepts

Gene expression profiling: the simultaneous assessment of the expression level of a large number of genes, often an entire genome, to obtain a global snapshot of the complete mRNA component of the cell at a given time; helps to distinguish between different cell types, different disease states, and different time points during development

Expressed sequence tag (EST): a single-pass sequencing read from the 3′ or 5′ end of a cDNA clone. In contrast, full-length cDNA (FLcDNA) sequencing involves generation and assembly of sequencing reads spanning the full length of cDNA clones

Microarray: a method for high-throughput gene expression profiling involving hybridization of mRNA to an array of complementary DNA probes corresponding to genes of interest. Hybridization intensity to a particular probe is related to the expression level of the corresponding transcript. The microarray method has dominated expression profiling research for the past decade

More concepts

Resequencing: sequencing of the genome or transcriptome of an individual of a species for which a reference genome sequence is available. In contrast, sequencing and assembly of the reference genome itself is termed de novo sequencing. Resequencing is commonly conducted to gauge sequence diversity within the species

Serial Analysis of Gene Expression (SAGE): the first sequencing-based method for high-throughput gene expression profiling. SAGE involves the generation of short sequence tags from 3′ ends of mRNA transcripts. The tags are then concatenated, sequenced, and counted providing estimates of transcript abundance

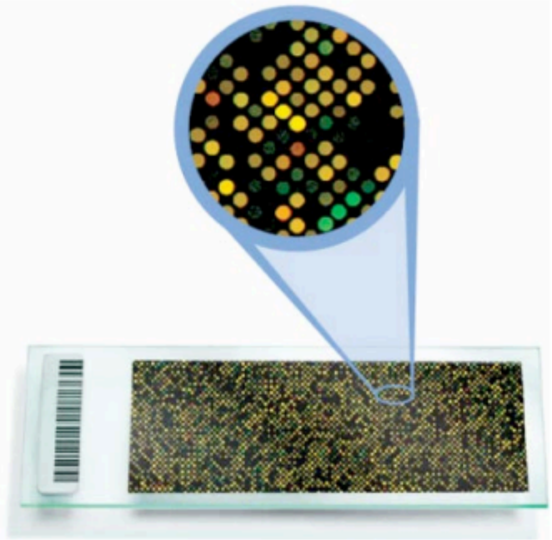
Sequencing depth: the total number of sequencing reads generated from a sequencing library. The higher the sequencing depth the higher the chance of detecting rare transcripts and sequence variants present in the cell

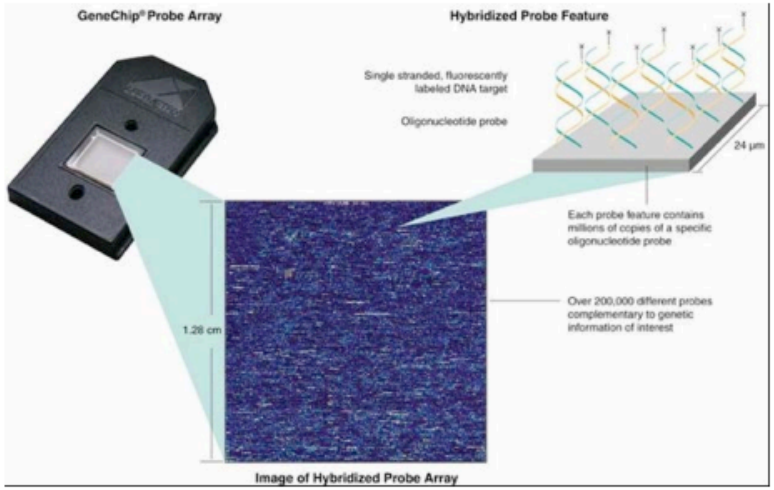
DNA microarrays

DNA microarray: a solid support (glass slide, silicon chip, etc) on which DNA of known sequence is deposited in a regular grid-like array.

- **Spotted or printed arrays (aka cDNA array)**: DNA feature physically transferred from a plate or reservoir to a solid support, typically a chemically modified glass microscope slide. (Agilent, GE, ABI)

- **Synthesized arrays (aka oligonucleotide array)**: DNA features chemically synthesized in-situ on the substrate. (Affymetrix, NimbleGen, Combimatrix)





A simplified protocol for DNA microarray experiment

1. Prepare or purchase DNA microarray

2. Isolate mRNA from cell cultures or tissue samples

3. Reverse transcribe mRNA into cDNA

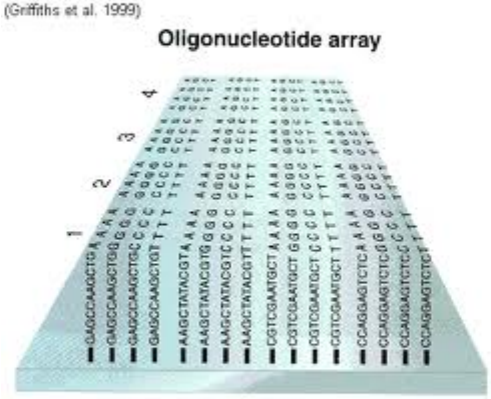
4. Label cDNA by incorporating fluorescently-labeled nucleotides

5. Hybridize labeled cDNA to DNA microarray

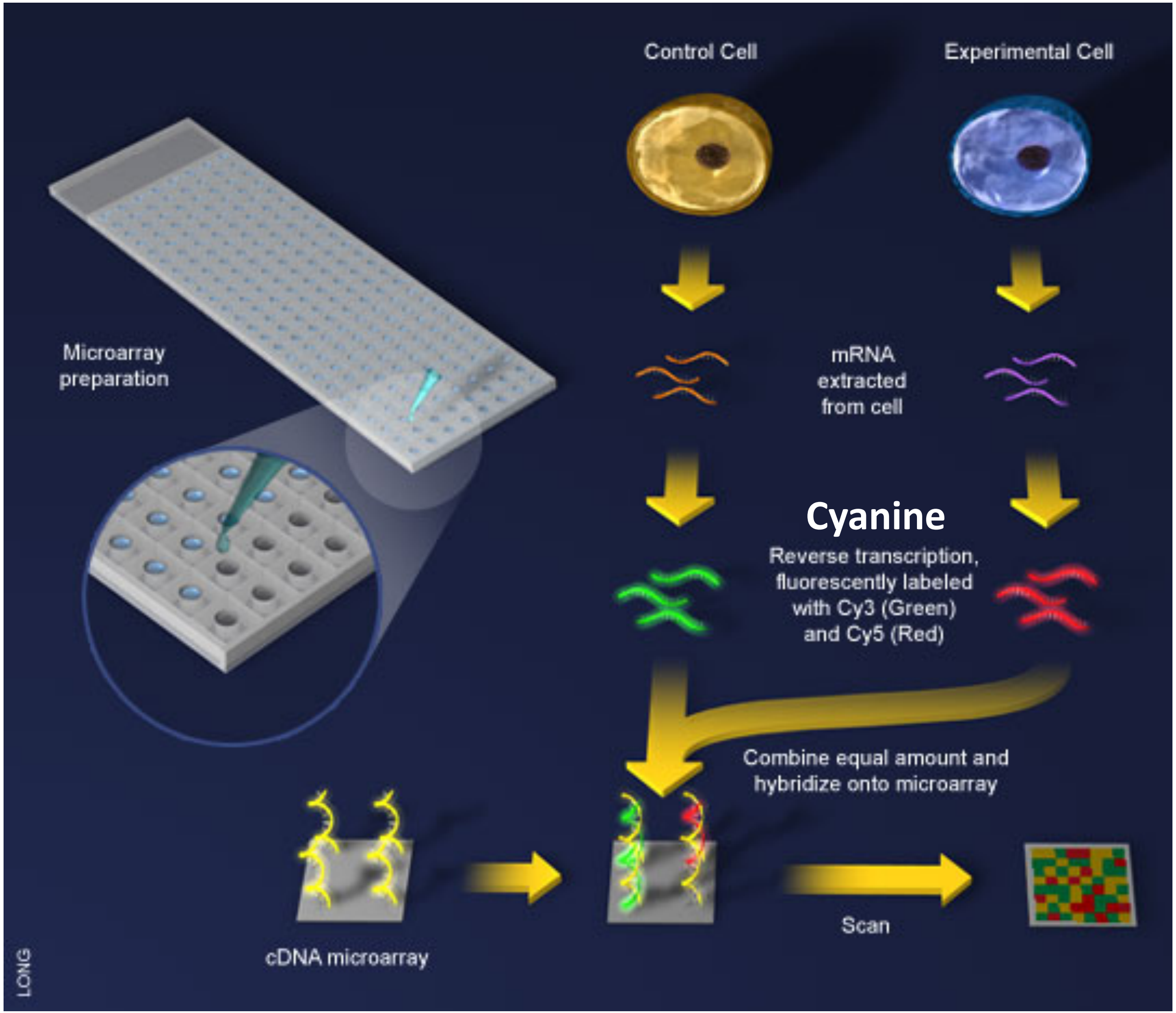
6. Wash and scan microarray in scanner

7. Analyze data

Basic intro to microarray


(Griffiths et al. 1999)

Oligonucleotide array


Control Cell

Experimental Cell

Microarray preparation

mRNA extracted from cell

Cyanine

Reverse transcription, fluorescently labeled with Cy3 (Green) and Cy5 (Red)

Combine equal amount and hybridize onto microarray

cDNA microarray

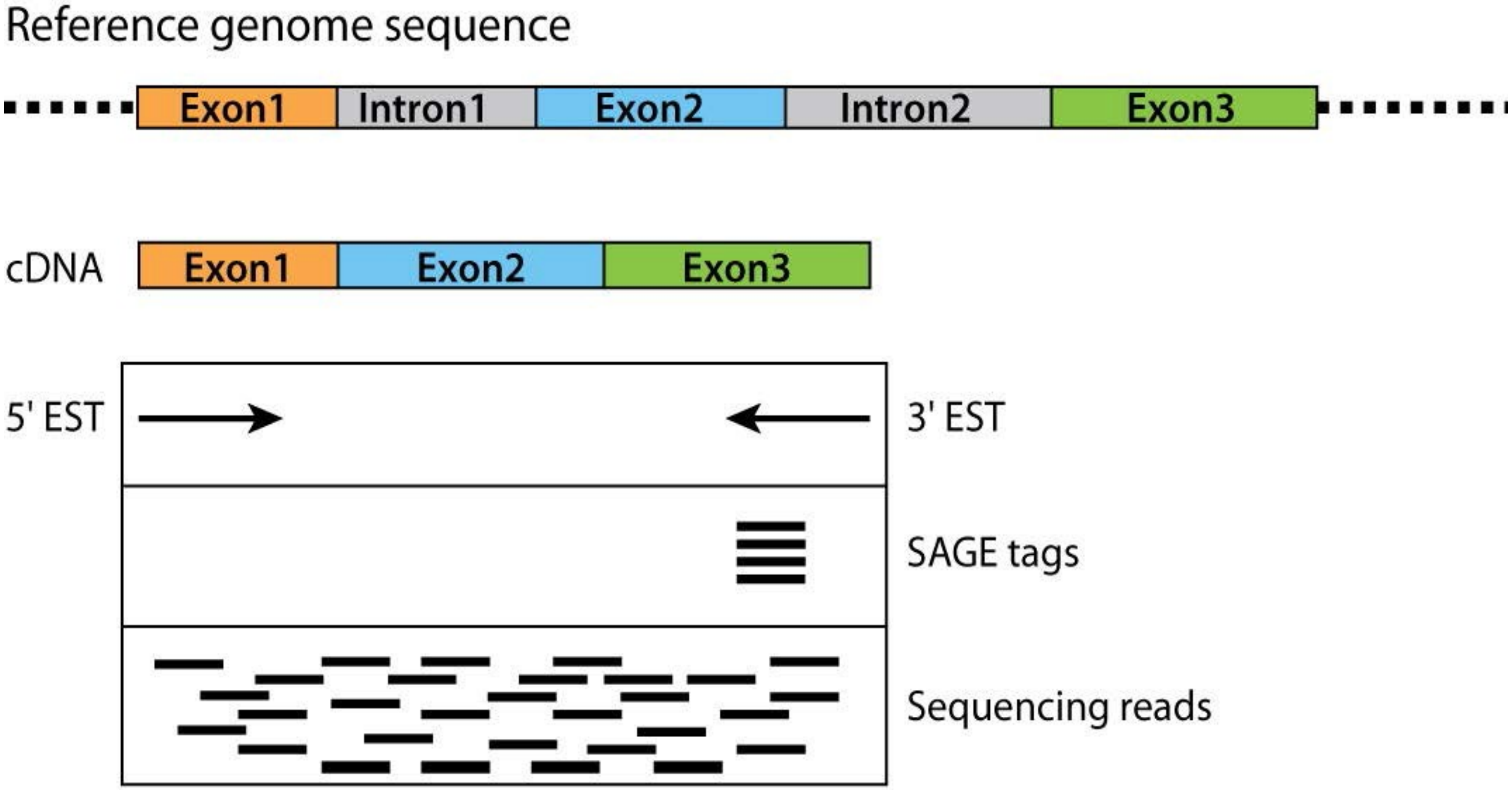
Scan

People are moving from microarray to high throughput sequencing

Technology	Tiling microarray	EST sequencing	RNA-Seq
Technology specifications			
Principle	Hybridization	Sanger sequencing	High-throughput sequencing
Resolution	From several to 100 bp	Single base	Single base
Throughput	High	Low	High
Reliance on genomic sequence	Yes	No	In some cases
Background noise	High	Low	Low
Practical issues			
Required amount of RNA	High	High	Low
Cost for mapping transcriptomes of large genomes	High	High	Relatively low
Application			
Dynamic range to quantify gene expression level	Up to a few hundred-fold	Not practical	>,8000-fold
Simultaneously map transcribed regions and gene expression	Yes	Limited for gene expression	Yes
Ability to distinguish different isoforms	Limited	Yes	Yes
Ability to distinguish allelic expression	Limited	Yes	Yes

Wang et al. Nature Rev Genet, 10:57, 2009

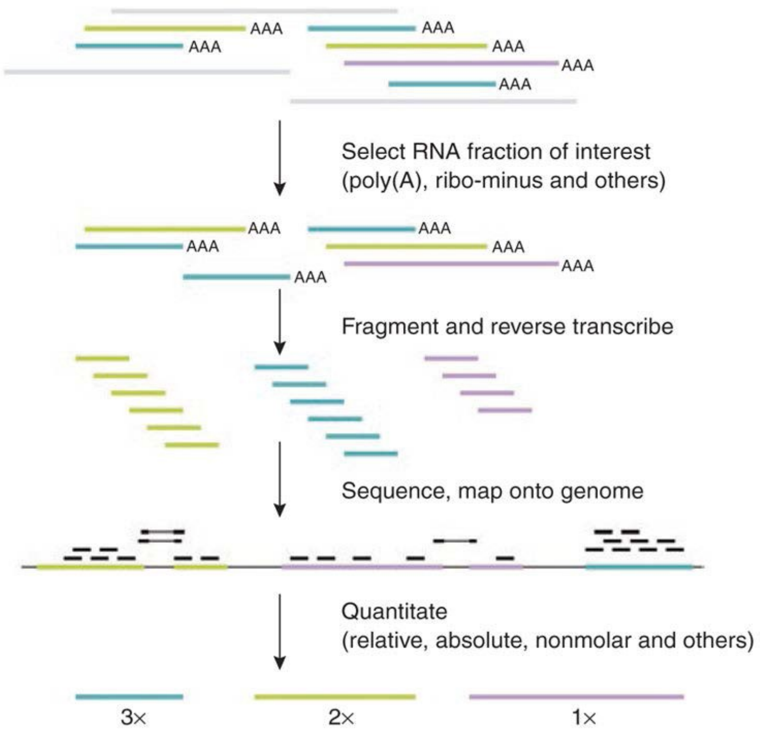
Sequences targeted by different sequencing-based approaches



Morozova et al. Annu Rev Genom Human Genet, 2009

RNA-Seq

- RNA-Seq: the use of next generation sequencing technologies to sequence cDNA in order to get information about a sample's RNA content.
- A RNA fraction of interest is selected, fragmented and reverse transcribed.
- Resulting cDNA are sequenced using one of the next generation sequencing technologies to obtain ten to a hundred million reads.
- The reads are mapped back onto the genome and analyzed to calculate expression levels.



Pepke et al. Nature Methods 6:S22 - S32, 2009

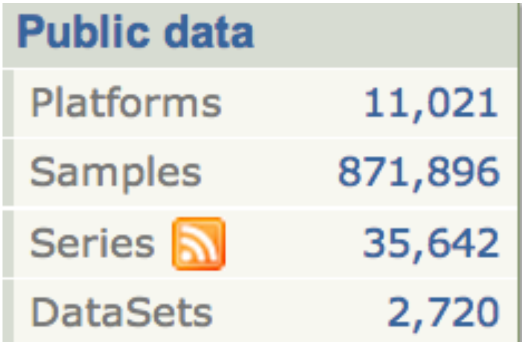
What data does GEO have?

http://www.ncbi.nlm.nih.gov/geo/

- Submitter supplied: Platform, Sample, Series
- NCBI curated: DataSets and Profiles
- Tools: GEO BLAST and GEO2R

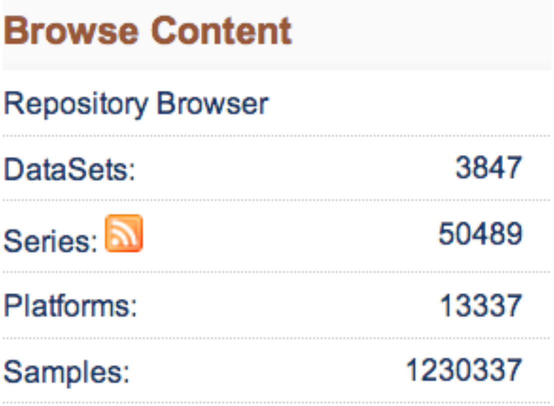
Omics data:

Genomics
Transcriptomics
Epigenomics
Proteomics
…



Public data

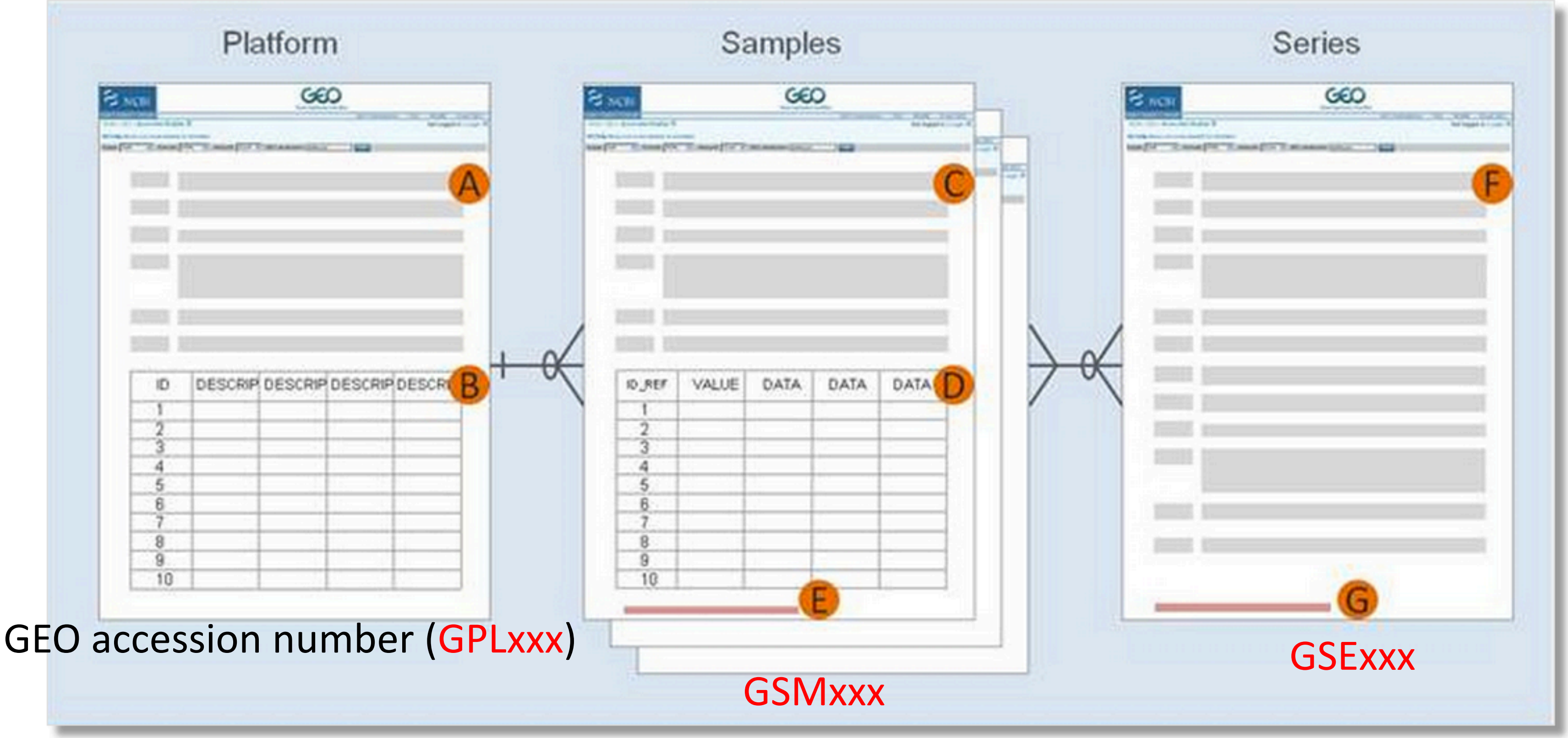
Platforms	11,021
Samples	871,896
Series	35,642
DataSets	2,720



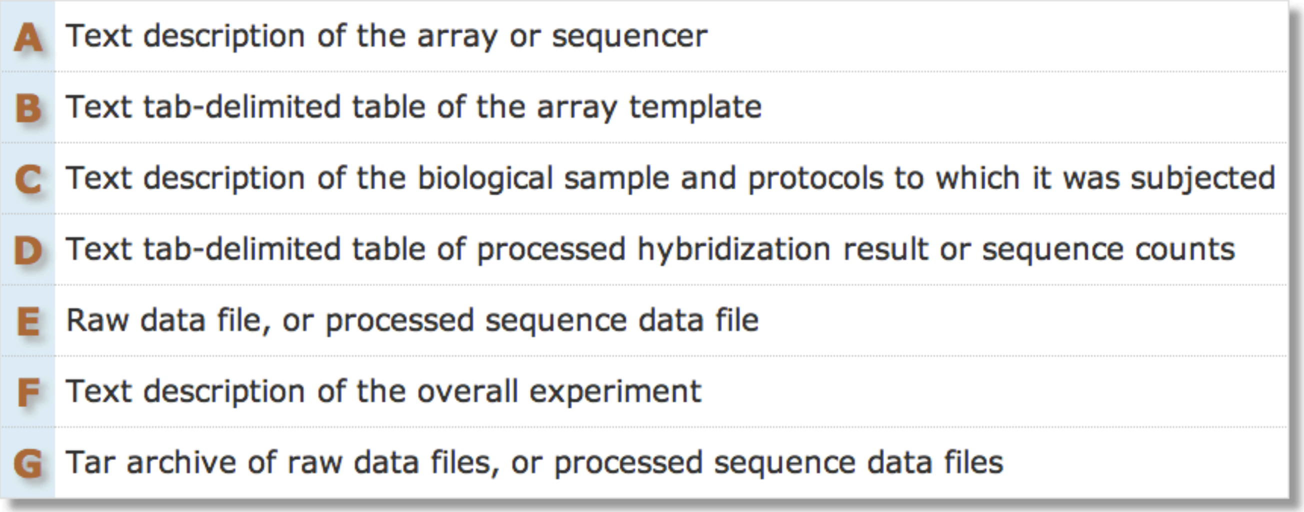
Browse Content

Repository Browser

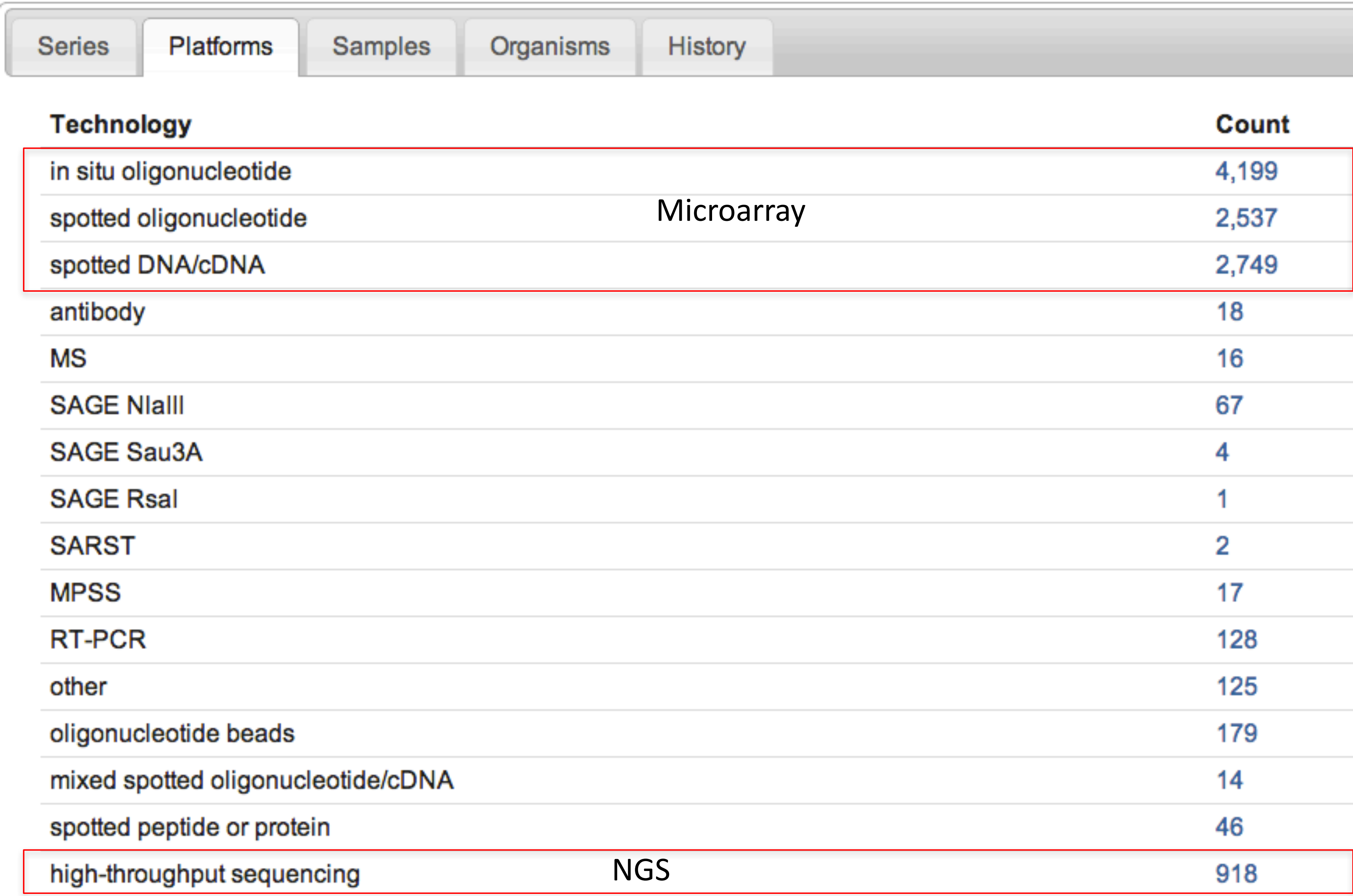
DataSets:	3847
Series:	50489
Platforms:	13337
Samples:	1230337



For almost all array data submissions, you will be asked to provide the following information:



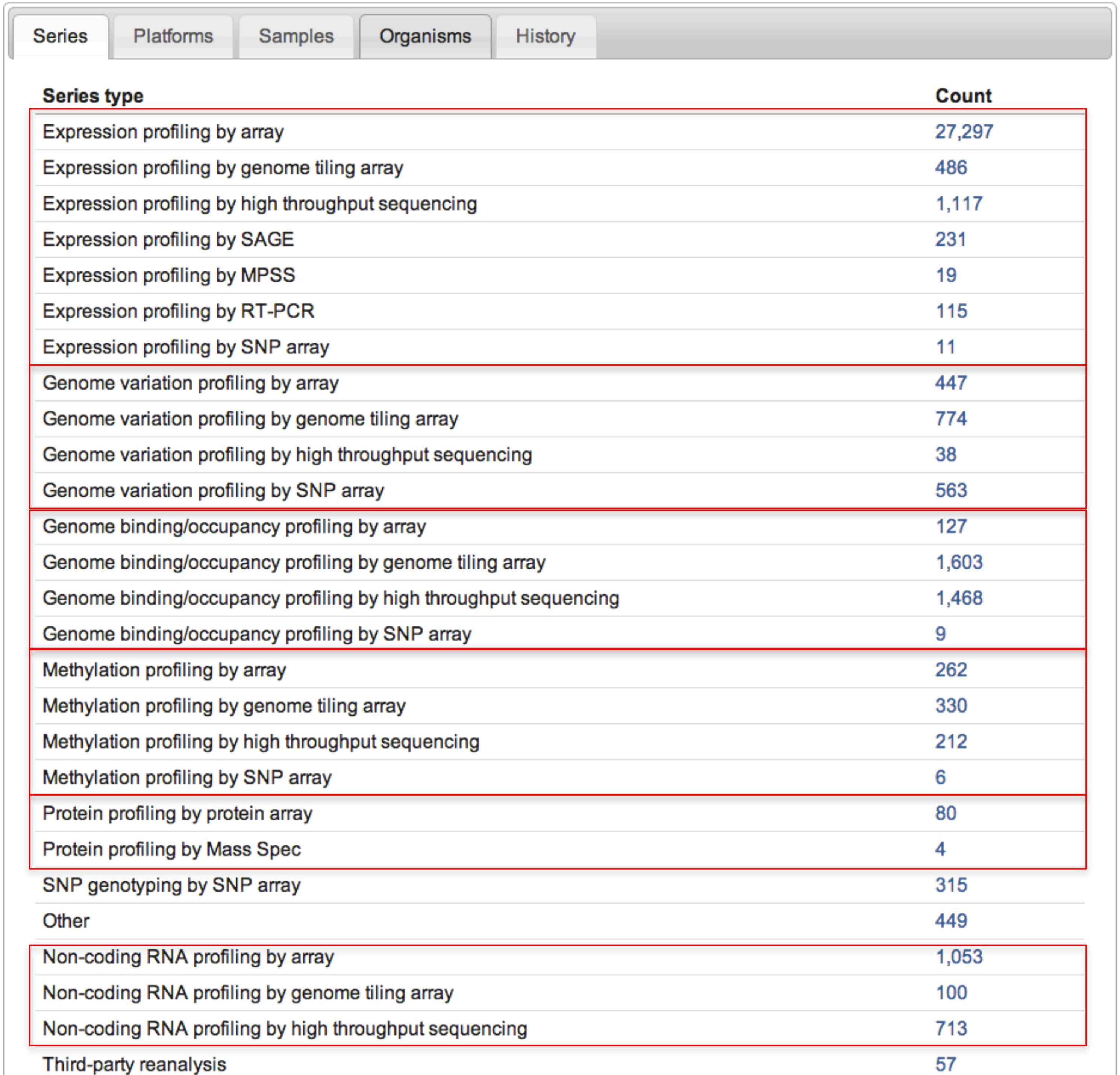
A	Text description of the array or sequencer
B	Text tab-delimited table of the array template
C	Text description of the biological sample and protocols to which it was subjected
D	Text tab-delimited table of processed hybridization result or sequence counts
E	Raw data file, or processed sequence data file
F	Text description of the overall experiment
G	Tar archive of raw data files, or processed sequence data files



Technology	Count
in situ oligonucleotide	4,199
spotted oligonucleotide	2,537
spotted DNA/cDNA	2,749
antibody	18
MS	16
SAGE NlaIII	67
SAGE Sau3A	4
SAGE RsaI	1
SARST	2
MPSS	17
RT-PCR	128
other	125
oligonucleotide beads	179
mixed spotted oligonucleotide/cDNA	14
spotted peptide or protein	46
high-throughput sequencing	918

Microarray

NGS



Series type	Count
Expression profiling by array	27,297
Expression profiling by genome tiling array	486
Expression profiling by high throughput sequencing	1,117
Expression profiling by SAGE	231
Expression profiling by MPSS	19
Expression profiling by RT-PCR	115
Expression profiling by SNP array	11
Genome variation profiling by array	447
Genome variation profiling by genome tiling array	774
Genome variation profiling by high throughput sequencing	38
Genome variation profiling by SNP array	563
Genome binding/occupancy profiling by array	127
Genome binding/occupancy profiling by genome tiling array	1,603
Genome binding/occupancy profiling by high throughput sequencing	1,468
Genome binding/occupancy profiling by SNP array	9
Methylation profiling by array	262
Methylation profiling by genome tiling array	330
Methylation profiling by high throughput sequencing	212
Methylation profiling by SNP array	6
Protein profiling by protein array	80
Protein profiling by Mass Spec	4
SNP genotyping by SNP array	315
Other	449
Non-coding RNA profiling by array	1,053
Non-coding RNA profiling by genome tiling array	100
Non-coding RNA profiling by high throughput sequencing	713
Third-party reanalysis	57

Expression

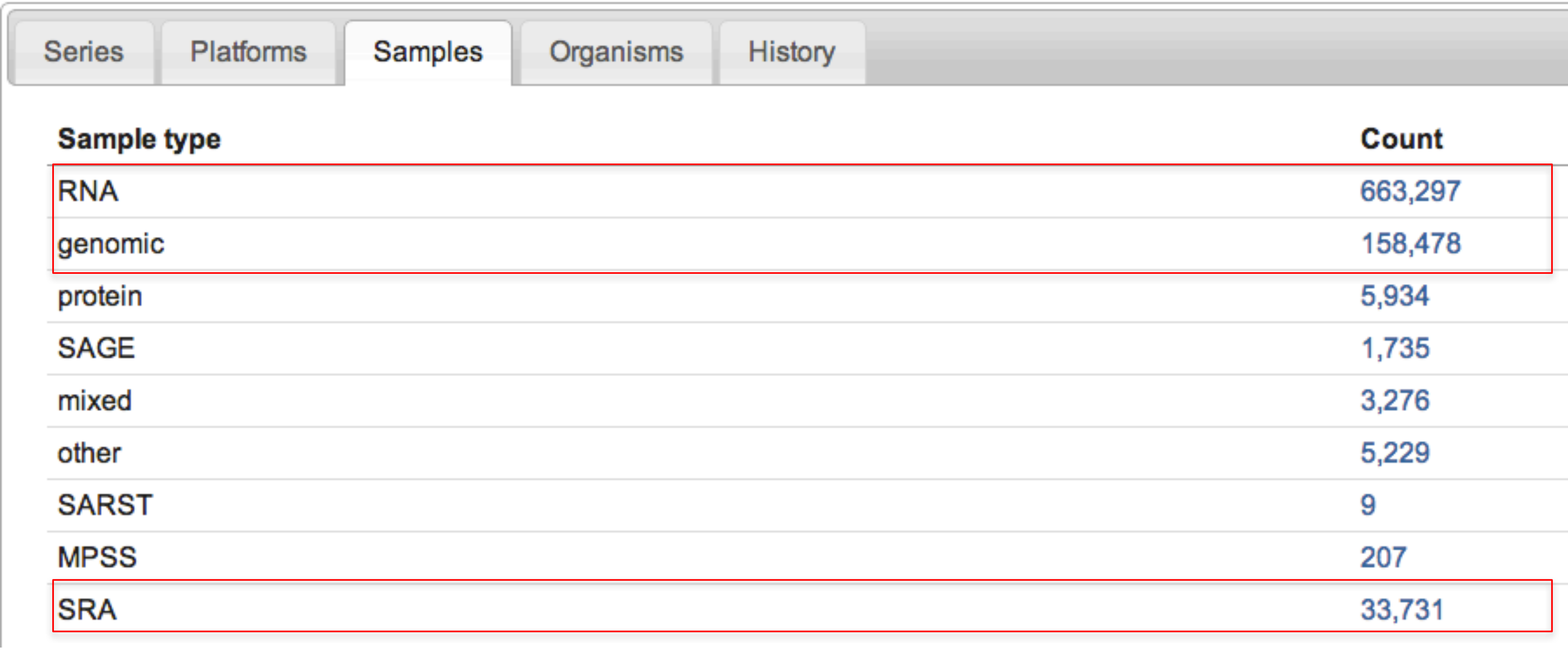
Genome variation

DNA-binding

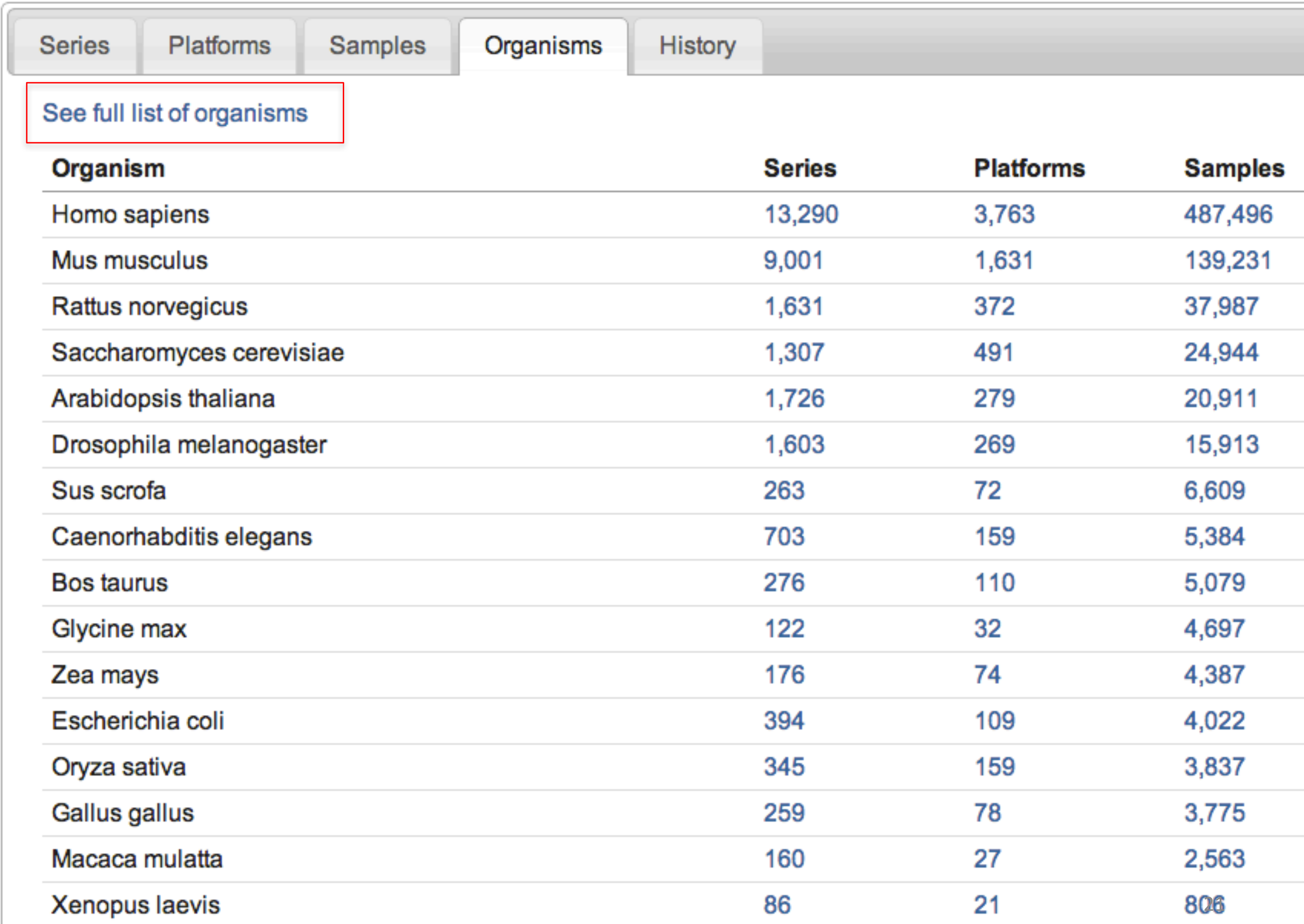
Methylation/
Epigenomics

Protein array

ncRNAs



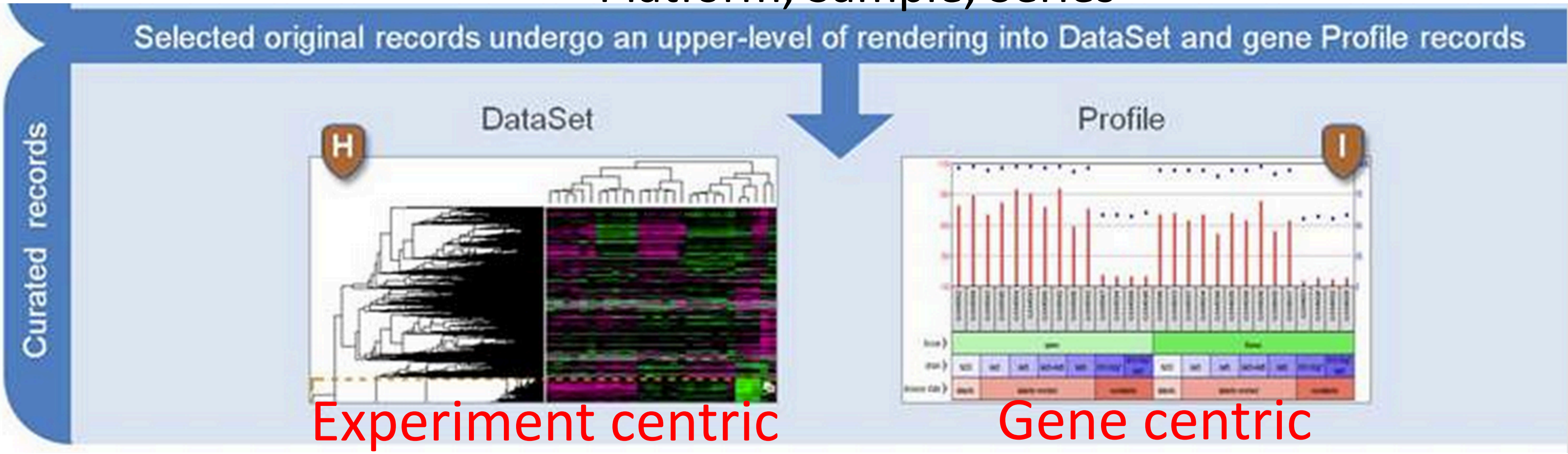
Sample type	Count
RNA	663,297
genomic	158,478
protein	5,934
SAGE	1,735
mixed	3,276
other	5,229
SARST	9
MPSS	207
SRA	33,731



Organism	Series	Platforms	Samples
Homo sapiens	13,290	3,763	487,496
Mus musculus	9,001	1,631	139,231
Rattus norvegicus	1,631	372	37,987
Saccharomyces cerevisiae	1,307	491	24,944
Arabidopsis thaliana	1,726	279	20,911
Drosophila melanogaster	1,603	269	15,913
Sus scrofa	263	72	6,609
Caenorhabditis elegans	703	159	5,384
Bos taurus	276	110	5,079
Glycine max	122	32	4,697
Zea mays	176	74	4,387
Escherichia coli	394	109	4,022
Oryza sativa	345	159	3,837
Gallus gallus	259	78	3,775
Macaca mulatta	160	27	2,563
Xenopus laevis	86	21	806

http://www.ncbi.nlm.nih.gov/geo/info/overview.html

Platform, Sample, Series



Experiment centric

Gene centric

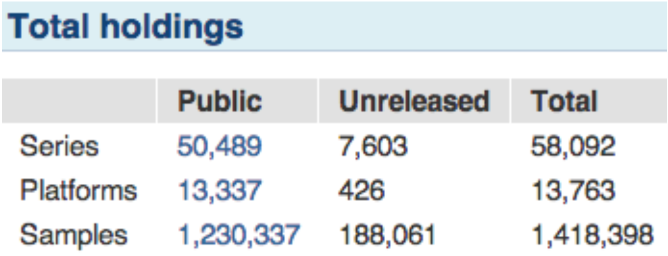
Data of a GEO Series are reassembled by GEO staff into GEO Dataset records (GDSxxx).

A DataSet represents a curated collection of biologically and statistically comparable GEO Samples and forms the basis of GEO's suite of data display and analysis tools.

Not all submitted data are suitable for DataSet assembly, so not all Series have corresponding DataSet record(s).

Profiles are derived from DataSets

A Profile consists of the expression measurements for an individual gene across all Samples in a DataSet.



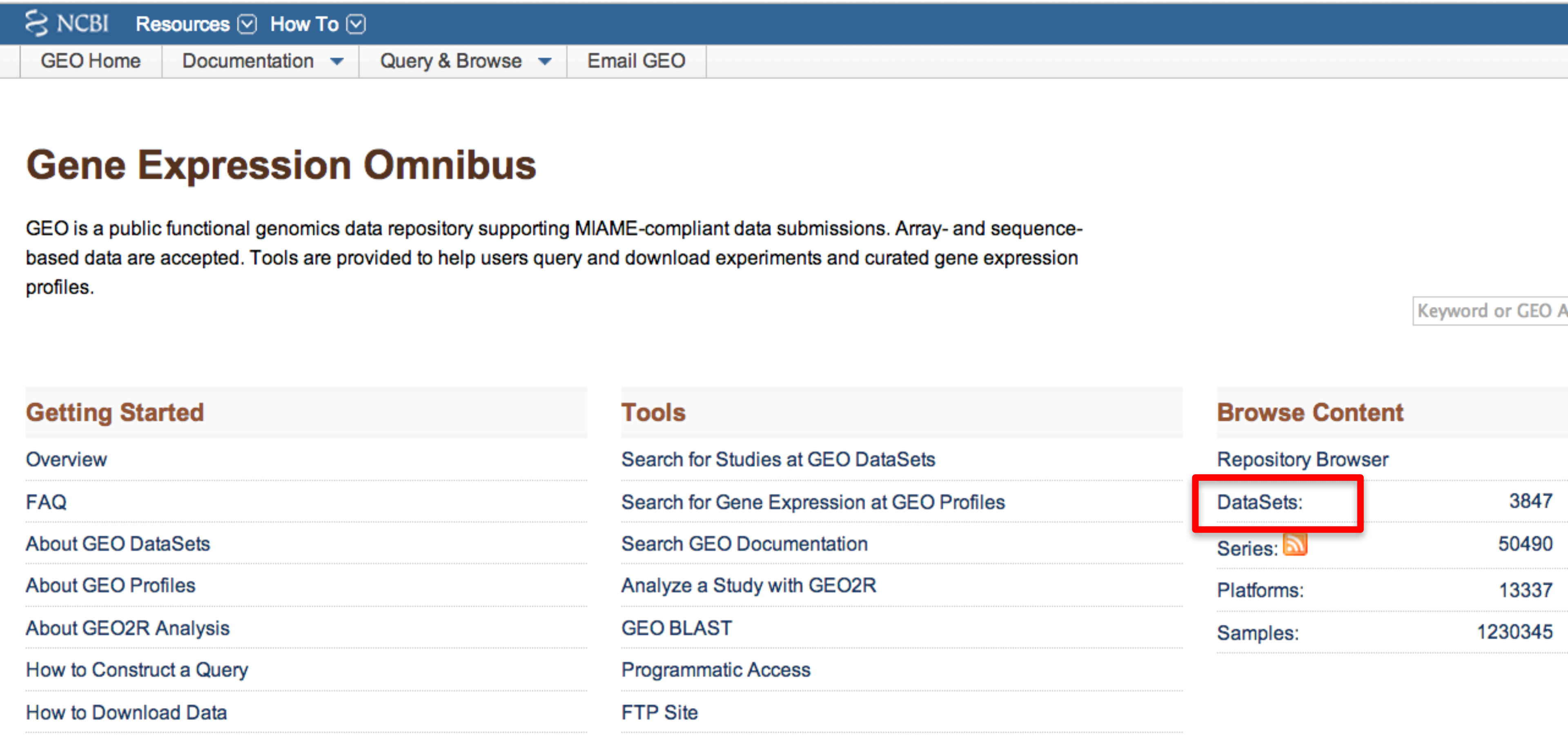
Total holdings

	Public	Unreleased	Total
Series	50,489	7,603	58,092
Platforms	13,337	426	13,763
Samples	1,230,337	188,061	1,418,398

Hands on exercise 1

GEO browse and query

http://www.ncbi.nlm.nih.gov/geo/



Gene Expression Omnibus

GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles.

Getting Started

- Overview
- FAQ
- About GEO DataSets
- About GEO Profiles
- About GEO2R Analysis
- How to Construct a Query
- How to Download Data

Tools

- Search for Studies at GEO DataSets
- Search for Gene Expression at GEO Profiles
- Search GEO Documentation
- Analyze a Study with GEO2R
- GEO BLAST
- Programmatic Access
- FTP Site

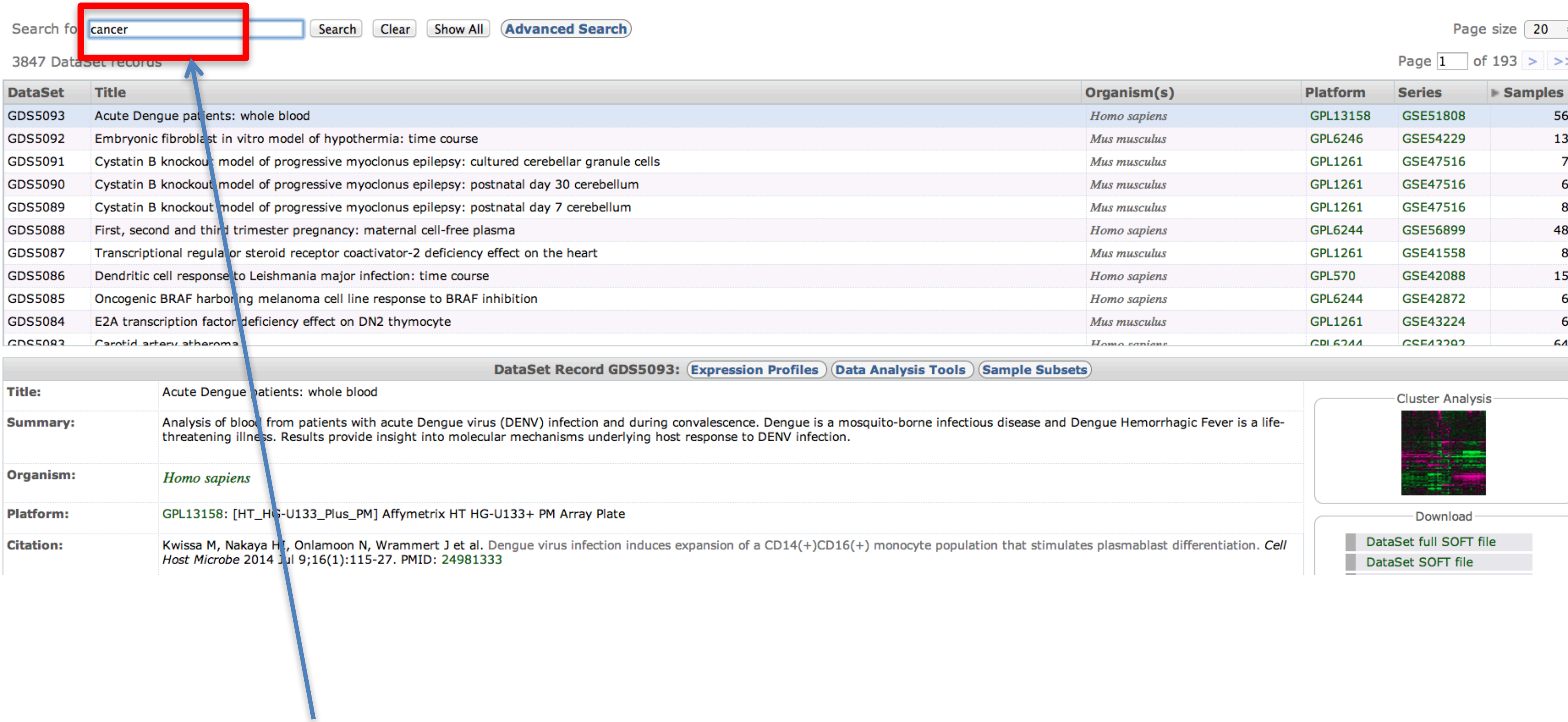
Browse Content

Repository Browser

DataSets:	3847
Series:	50490
Platforms:	13337
Samples:	1230345

Try:
cancer
colon cancer
arabidopsis

These are only DataSets



cancer

3847 DataSet records

DataSet	Title	Organism(s)	Platform	Series	Samples
GDS5093	Acute Dengue patients: whole blood	*Homo sapiens*	GPL13158	GSE51808	56
GDS5092	Embryonic fibroblast in vitro model of hypothermia: time course	*Mus musculus*	GPL6246	GSE54229	13
GDS5091	Cystatin B knockout model of progressive myoclonus epilepsy: cultured cerebellar granule cells	*Mus musculus*	GPL1261	GSE47516	7
GDS5090	Cystatin B knockout model of progressive myoclonus epilepsy: postnatal day 30 cerebellum	*Mus musculus*	GPL1261	GSE47516	6
GDS5089	Cystatin B knockout model of progressive myoclonus epilepsy: postnatal day 7 cerebellum	*Mus musculus*	GPL1261	GSE47516	8
GDS5088	First, second and third trimester pregnancy: maternal cell-free plasma	*Homo sapiens*	GPL6244	GSE56899	48
GDS5087	Transcriptional regulator steroid receptor coactivator-2 deficiency effect on the heart	*Mus musculus*	GPL1261	GSE41558	8
GDS5086	Dendritic cell response to Leishmania major infection: time course	*Homo sapiens*	GPL570	GSE42088	15
GDS5085	Oncogenic BRAF harboring melanoma cell line response to BRAF inhibition	*Homo sapiens*	GPL6244	GSE42872	6
GDS5084	E2A transcription factor deficiency effect on DN2 thymocyte	*Mus musculus*	GPL1261	GSE43224	6
GDS5083	Carotid artery atheroma	*Homo sapiens*	GPL6244	GSE43292	64

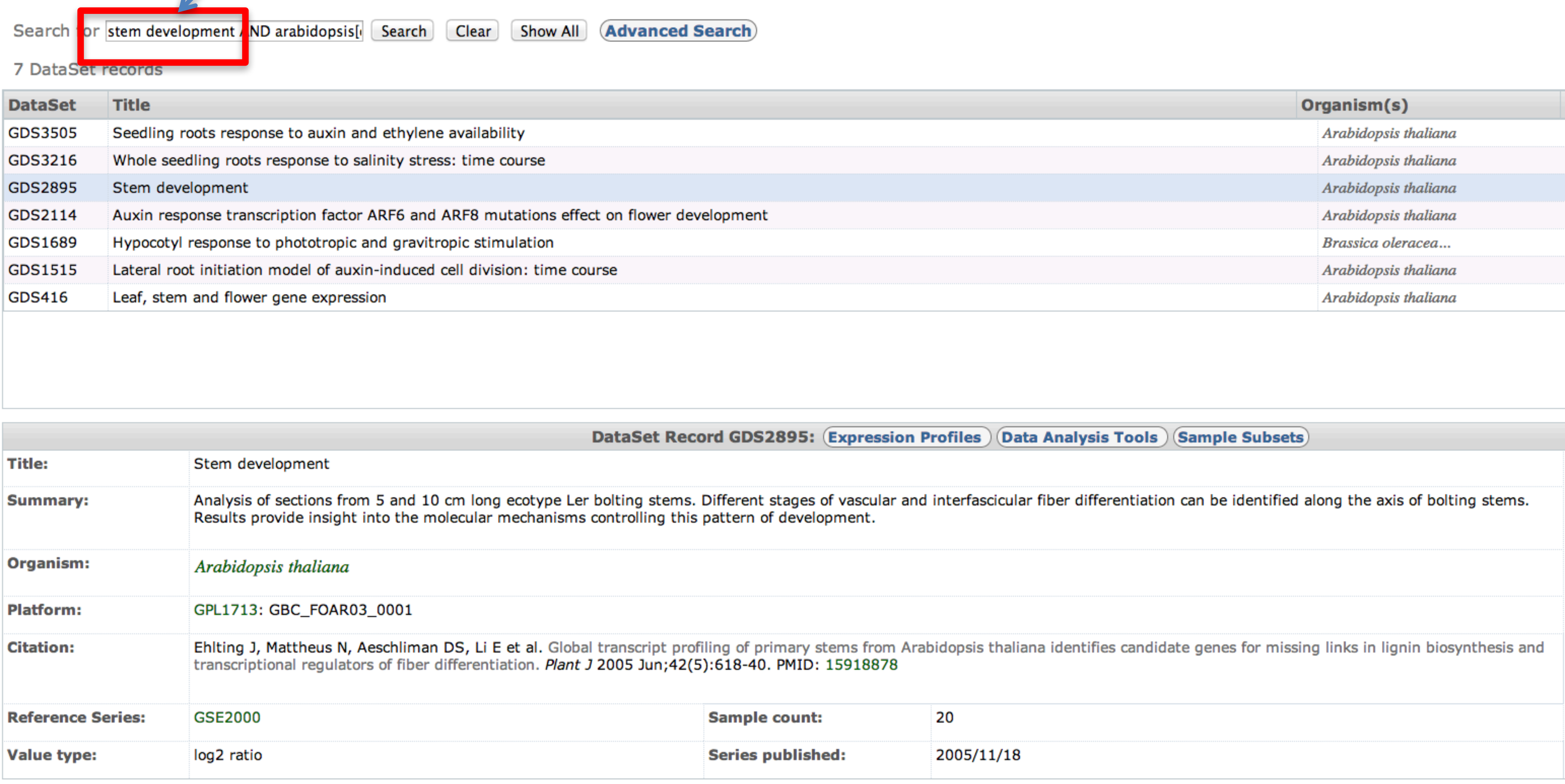
DataSet Record GDS5093: Expression Profiles | Data Analysis Tools | Sample Subsets

Title:	Acute Dengue patients: whole blood
Summary:	Analysis of blood from patients with acute Dengue virus (DENV) infection and during convalescence. Dengue is a mosquito-borne infectious disease and Dengue Hemorrhagic Fever is a life-threatening illness. Results provide insight into molecular mechanisms underlying host response to DENV infection.
Organism:	*Homo sapiens*
Platform:	GPL13158: [HT_HG-U133_Plus_PM] Affymetrix HT HG-U133+ PM Array Plate
Citation:	Kwissa M, Nakaya HI, Onlamoon N, Wrammert J et al. Dengue virus infection induces expansion of a CD14(+)CD16(+) monocyte population that stimulates plasmablast differentiation. *Cell Host Microbe* 2014 Jul 9;16(1):115-27. PMID: 24981333

Type the keyword in the search box and click search

Construct queries to narrow down the results **term [field] OPERATOR term [field]**

stem development AND arabidopsis[organism]



Search for stem development AND arabidopsis[

7 DataSet records

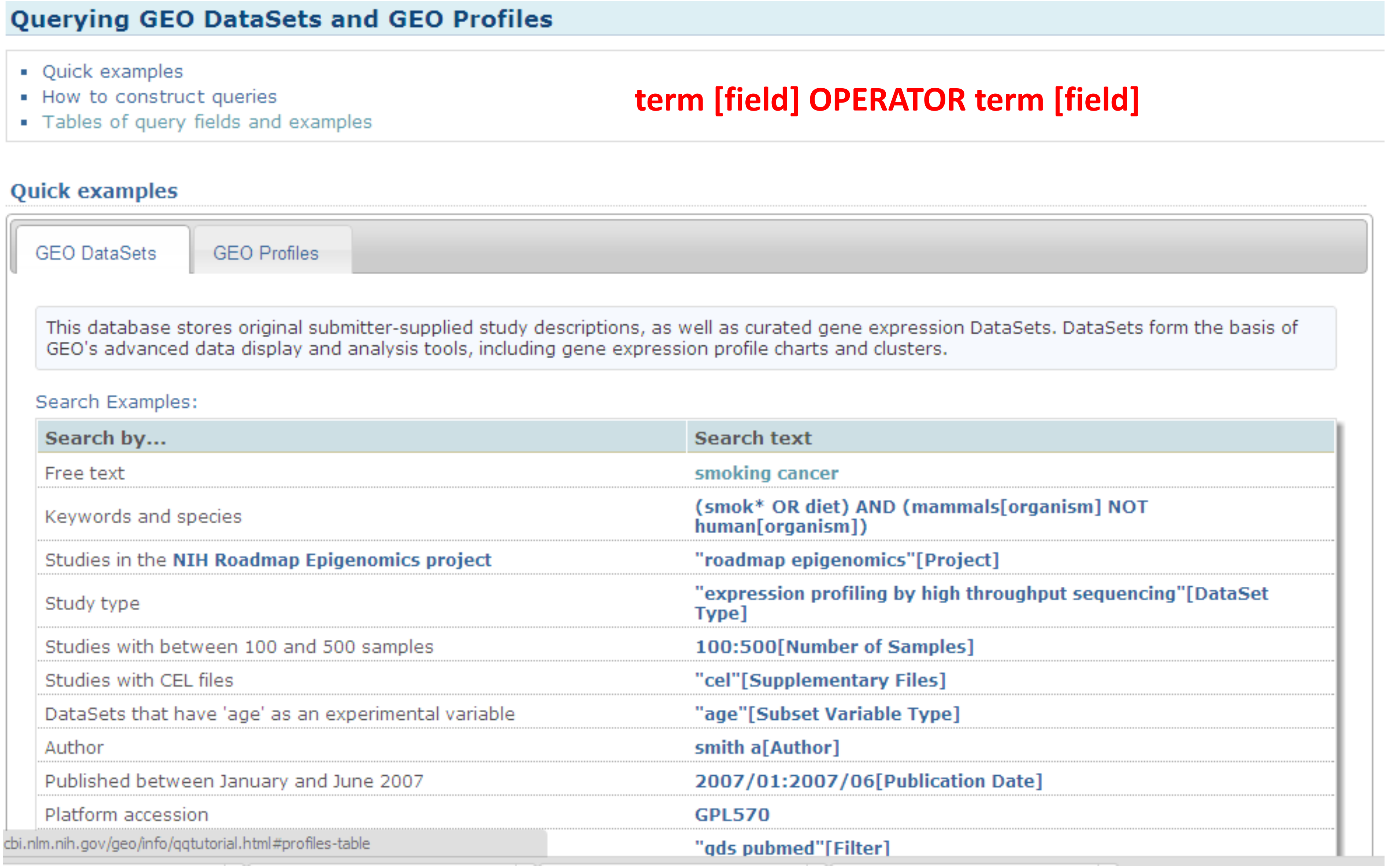
DataSet	Title	Organism(s)
GDS3505	Seedling roots response to auxin and ethylene availability	*Arabidopsis thaliana*
GDS3216	Whole seedling roots response to salinity stress: time course	*Arabidopsis thaliana*
GDS2895	Stem development	*Arabidopsis thaliana*
GDS2114	Auxin response transcription factor ARF6 and ARF8 mutations effect on flower development	*Arabidopsis thaliana*
GDS1689	Hypocotyl response to phototropic and gravitropic stimulation	*Brassica oleracea...*
GDS1515	Lateral root initiation model of auxin-induced cell division: time course	*Arabidopsis thaliana*
GDS416	Leaf, stem and flower gene expression	*Arabidopsis thaliana*

DataSet Record GDS2895:

Title:	Stem development		
Summary:	Analysis of sections from 5 and 10 cm long ecotype Ler bolting stems. Different stages of vascular and interfascicular fiber differentiation can be identified along the axis of bolting stems. Results provide insight into the molecular mechanisms controlling this pattern of development.		
Organism:	*Arabidopsis thaliana*		
Platform:	GPL1713: GBC_FOAR03_0001		
Citation:	Ehlting J, Mattheus N, Aeschliman DS, Li E et al. Global transcript profiling of primary stems from Arabidopsis thaliana identifies candidate genes for missing links in lignin biosynthesis and transcriptional regulators of fiber differentiation. *Plant J* 2005 Jun;42(5):618-40. PMID: 15918878		
Reference Series:	GSE2000	**Sample count:**	20
Value type:	log2 ratio	**Series published:**	2005/11/18

http://www.ncbi.nlm.nih.gov/geo/info/qqtutorial.html

term [field] OPERATOR term [field]



Querying GEO DataSets and GEO Profiles

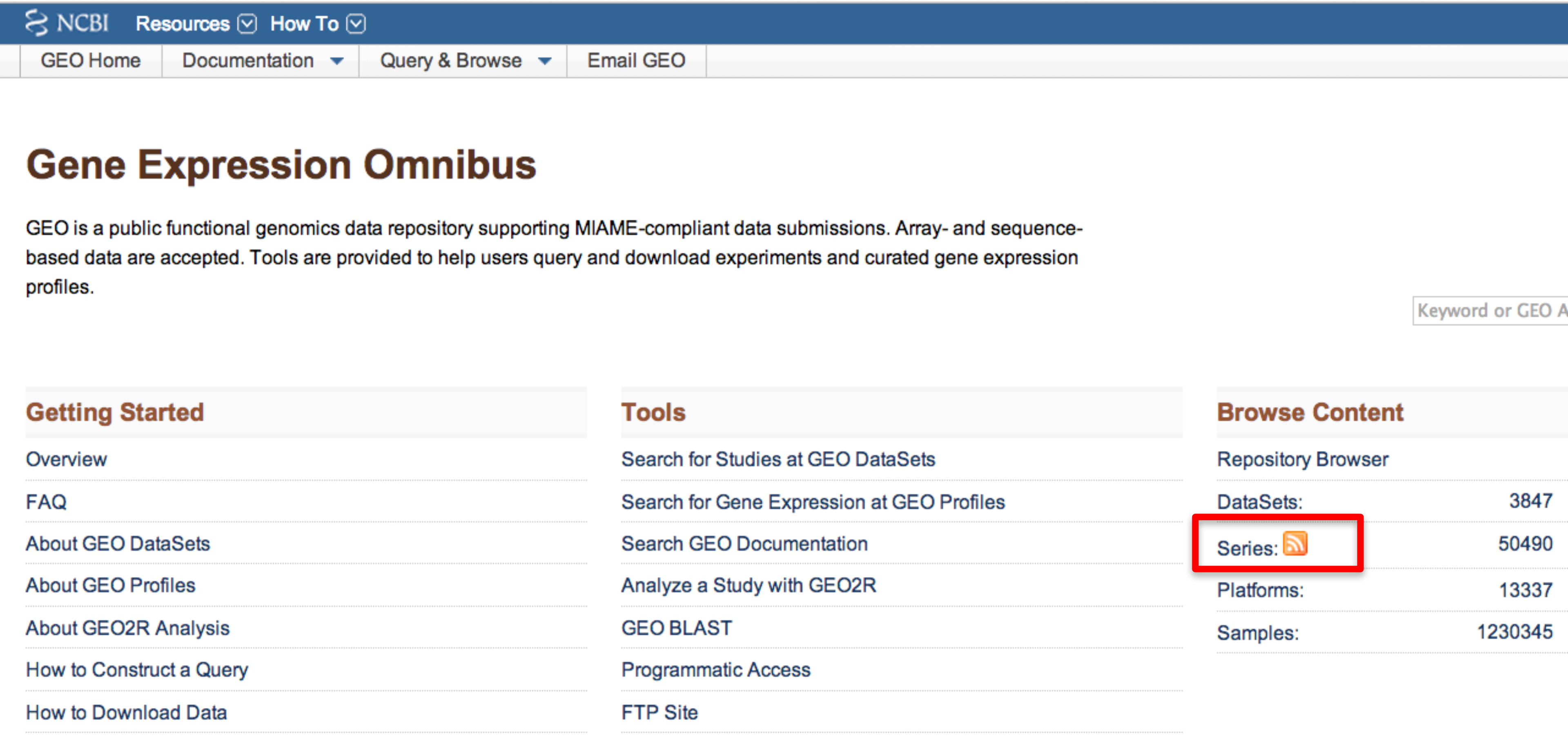
- Quick examples
- How to construct queries
- Tables of query fields and examples

Quick examples

This database stores original submitter-supplied study descriptions, as well as curated gene expression DataSets. DataSets form the basis of GEO's advanced data display and analysis tools, including gene expression profile charts and clusters.

Search Examples:

Search by...	Search text
Free text	smoking cancer
Keywords and species	**(smok* OR diet) AND (mammals[organism] NOT human[organism])**
Studies in the **NIH Roadmap Epigenomics project**	**"roadmap epigenomics"[Project]**
Study type	**"expression profiling by high throughput sequencing"[DataSet Type]**
Studies with between 100 and 500 samples	**100:500[Number of Samples]**
Studies with CEL files	**"cel"[Supplementary Files]**
DataSets that have 'age' as an experimental variable	**"age"[Subset Variable Type]**
Author	**smith a[Author]**
Published between January and June 2007	**2007/01:2007/06[Publication Date]**
Platform accession	**GPL570**
	"gds pubmed"[Filter]



Gene Expression Omnibus

GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles.

Getting Started

- Overview
- FAQ
- About GEO DataSets
- About GEO Profiles
- About GEO2R Analysis
- How to Construct a Query
- How to Download Data

Tools

- Search for Studies at GEO DataSets
- Search for Gene Expression at GEO Profiles
- Search GEO Documentation
- Analyze a Study with GEO2R
- GEO BLAST
- Programmatic Access
- FTP Site

Browse Content

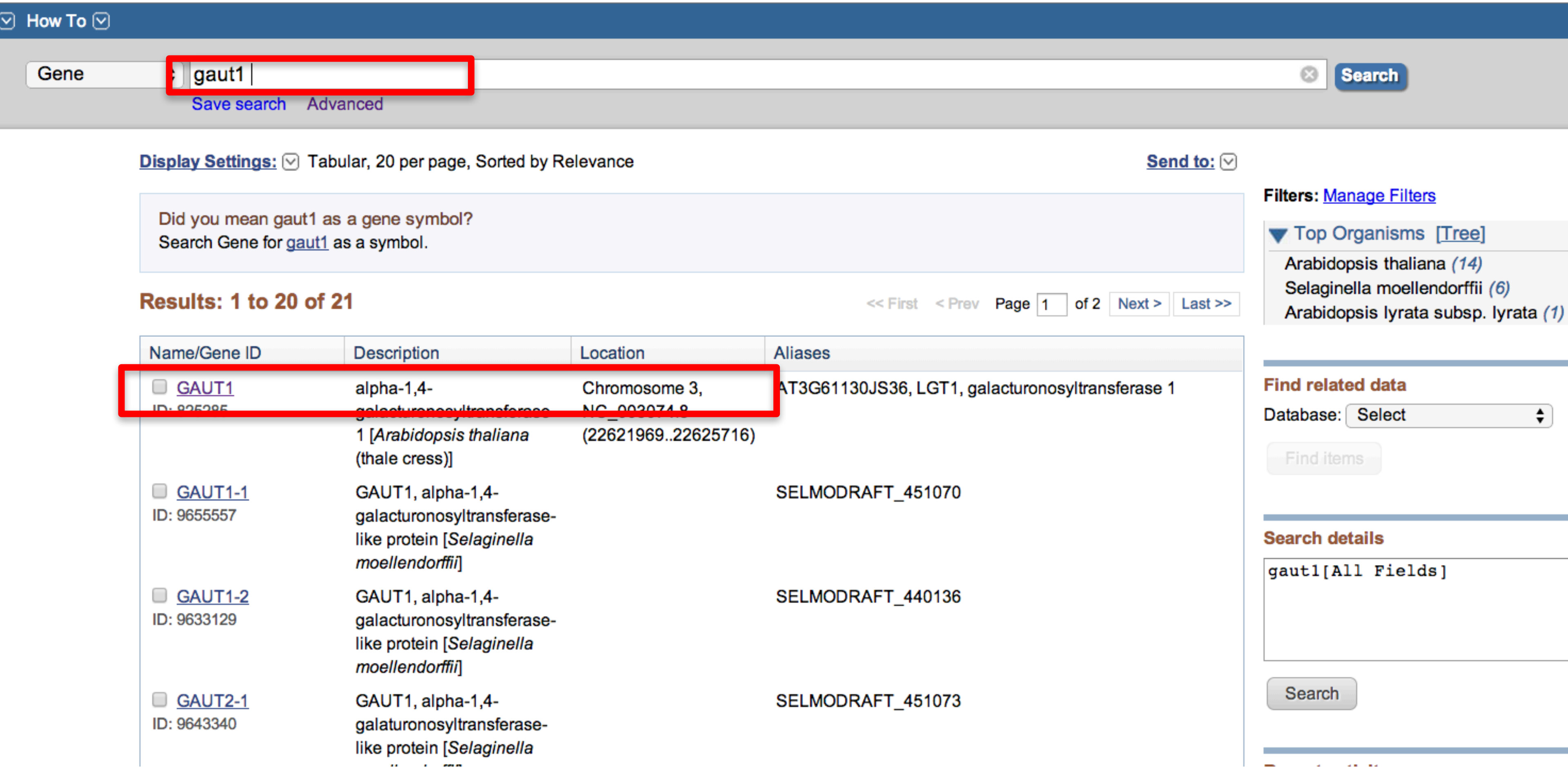
Repository Browser

DataSets:	3847
Series:	50490
Platforms:	13337
Samples:	1230345

Hands on exercise 2

GEO gene profiles

Search for a gene: GAUT1



gaut1

Display Settings: Tabular, 20 per page, Sorted by Relevance

Send to:

Did you mean gaut1 as a gene symbol?
Search Gene for gaut1 as a symbol.

Results: 1 to 20 of 21

<< First < Prev Page 1 of 2 Next > Last >>

Name/Gene ID	Description	Location	Aliases
☐ GAUT1 ID: 825285	alpha-1,4-galacturonosyltransferase 1 [*Arabidopsis thaliana* (thale cress)]	Chromosome 3, NC_003074.8 (22621969..22625716)	AT3G61130JS36, LGT1, galacturonosyltransferase 1
☐ GAUT1-1 ID: 9655557	GAUT1, alpha-1,4-galacturonosyltransferase-like protein [*Selaginella moellendorffii*]		SELMODRAFT_451070
☐ GAUT1-2 ID: 9633129	GAUT1, alpha-1,4-galacturonosyltransferase-like protein [*Selaginella moellendorffii*]		SELMODRAFT_440136
☐ GAUT2-1 ID: 9643340	GAUT1, alpha-1,4-galaturonosyltransferase-like protein [*Selaginella*		SELMODRAFT_451073

Filters: Manage Filters

Top Organisms [Tree]

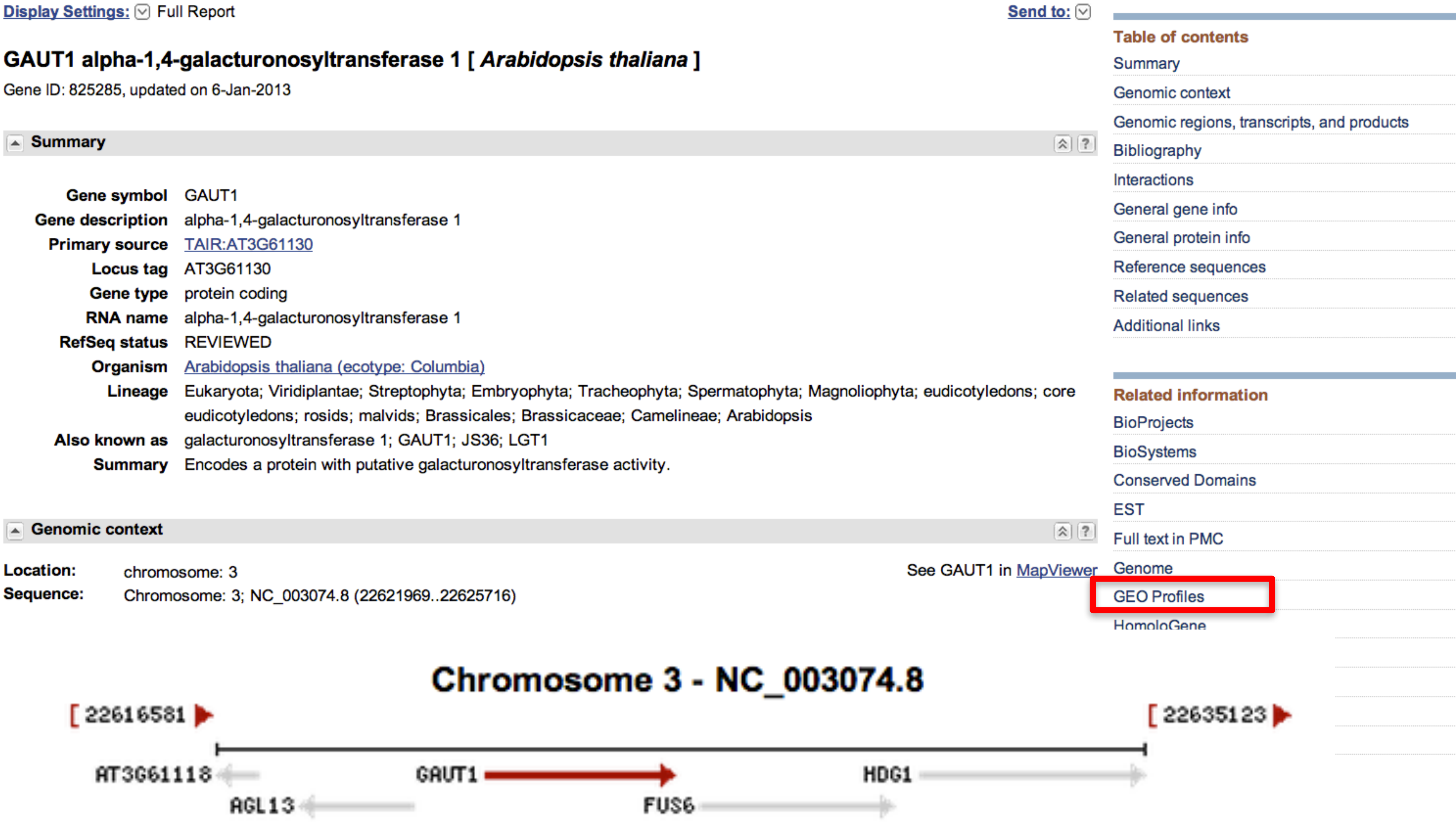
Arabidopsis thaliana	*(14)*
Selaginella moellendorffii	*(6)*
Arabidopsis lyrata subsp. lyrata	*(1)*

Find related data

Database: Select

Search details

```
gaut1[All Fields]
```



Display Settings: Full Report

Send to:

GAUT1 alpha-1,4-galacturonosyltransferase 1 [*Arabidopsis thaliana*]

Gene ID: 825285, updated on 6-Jan-2013

Summary

Gene symbol	GAUT1
Gene description	alpha-1,4-galacturonosyltransferase 1
Primary source	TAIR:AT3G61130
Locus tag	AT3G61130
Gene type	protein coding
RNA name	alpha-1,4-galacturonosyltransferase 1
RefSeq status	REVIEWED
Organism	Arabidopsis thaliana (ecotype: Columbia)
Lineage	Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; eudicotyledons; core eudicotyledons; rosids; malvids; Brassicales; Brassicaceae; Camelineae; Arabidopsis
Also known as	galacturonosyltransferase 1; GAUT1; JS36; LGT1
Summary	Encodes a protein with putative galacturonosyltransferase activity.

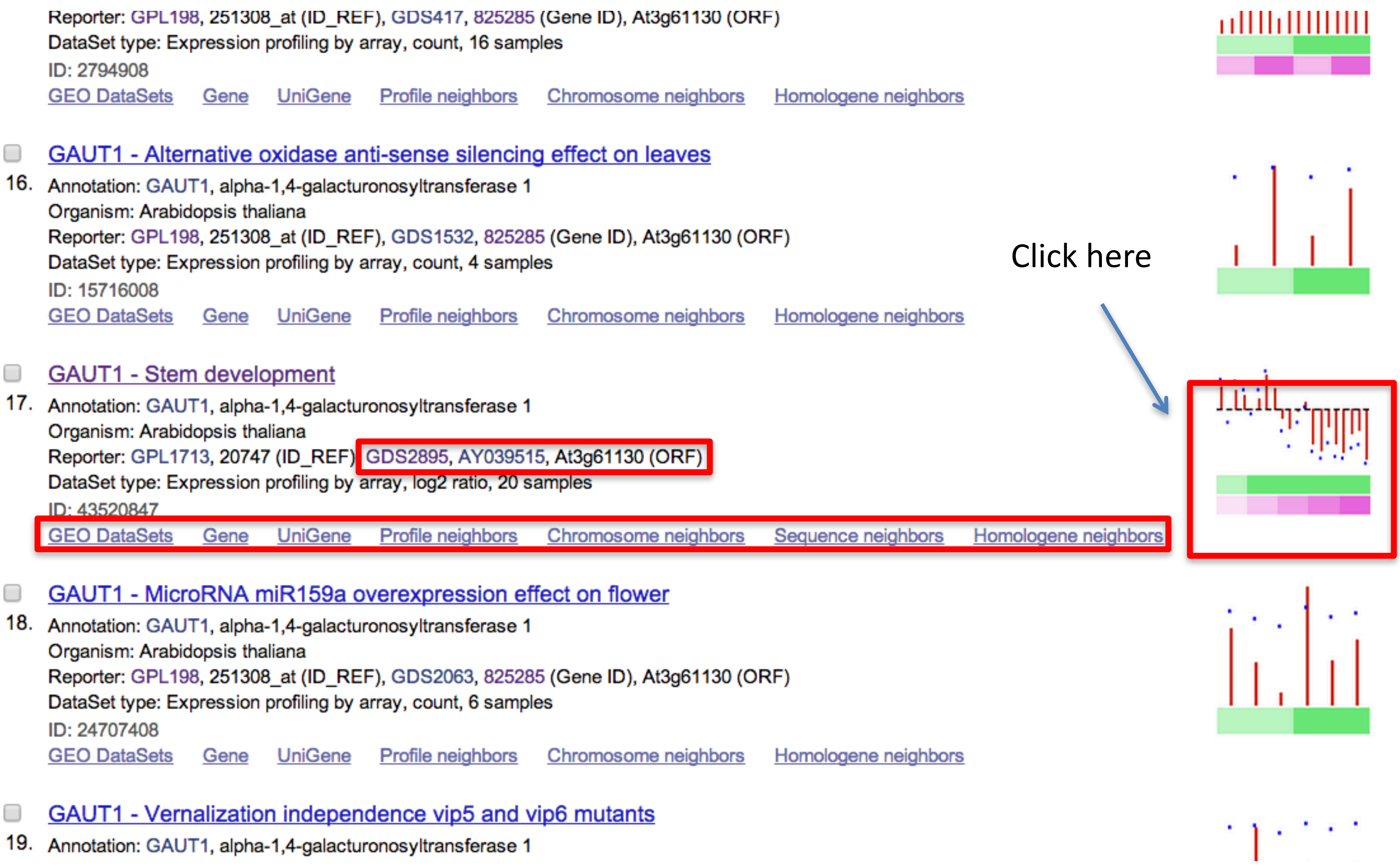
Genomic context

Location: chromosome: 3

Sequence: Chromosome: 3; NC_003074.8 (22621969..22625716)

See GAUT1 in MapViewer

Scroll down to find record 17



Reporter: GPL198, 251308_at (ID_REF), GDS417, 825285 (Gene ID), At3g61130 (ORF)
DataSet type: Expression profiling by array, count, 16 samples
ID: 2794908
GEO DataSets Gene UniGene Profile neighbors Chromosome neighbors Homologene neighbors

☐ GAUT1 - Alternative oxidase anti-sense silencing effect on leaves
16. Annotation: GAUT1, alpha-1,4-galacturonosyltransferase 1
Organism: Arabidopsis thaliana
Reporter: GPL198, 251308_at (ID_REF), GDS1532, 825285 (Gene ID), At3g61130 (ORF)
DataSet type: Expression profiling by array, count, 4 samples
ID: 15716008
GEO DataSets Gene UniGene Profile neighbors Chromosome neighbors Homologene neighbors

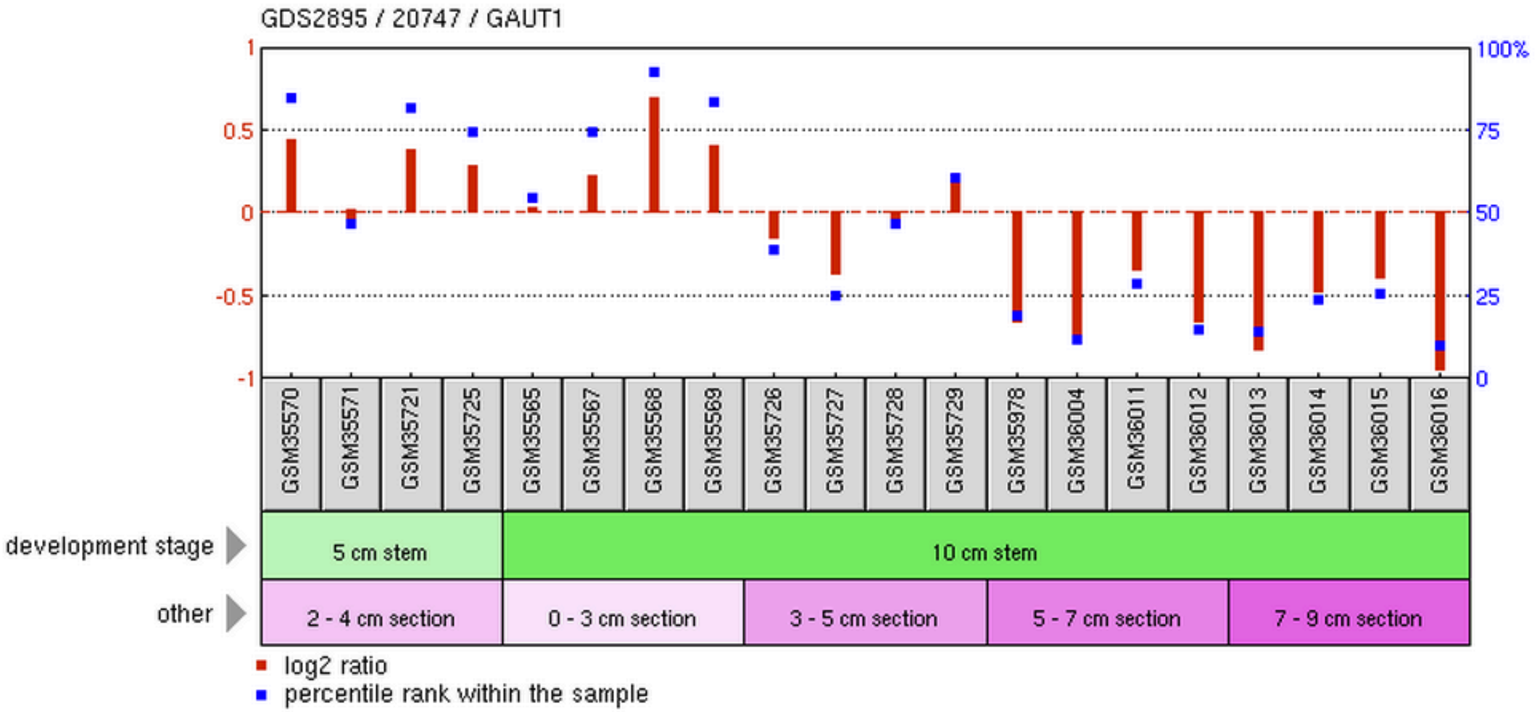
☐ GAUT1 - Stem development
17. Annotation: GAUT1, alpha-1,4-galacturonosyltransferase 1
Organism: Arabidopsis thaliana
Reporter: GPL1713, 20747 (ID_REF), GDS2895, AY039515, At3g61130 (ORF)
DataSet type: Expression profiling by array, log2 ratio, 20 samples
ID: 43520847
GEO DataSets Gene UniGene Profile neighbors Chromosome neighbors Sequence neighbors Homologene neighbors

☐ GAUT1 - MicroRNA miR159a overexpression effect on flower
18. Annotation: GAUT1, alpha-1,4-galacturonosyltransferase 1
Organism: Arabidopsis thaliana
Reporter: GPL198, 251308_at (ID_REF), GDS2063, 825285 (Gene ID), At3g61130 (ORF)
DataSet type: Expression profiling by array, count, 6 samples
ID: 24707408
GEO DataSets Gene UniGene Profile neighbors Chromosome neighbors Homologene neighbors

☐ GAUT1 - Vernalization independence vip5 and vip6 mutants
19. Annotation: GAUT1, alpha-1,4-galacturonosyltransferase 1

Click here

Profile GDS2895 / 20747 / GAUT1
Title Stem development
Organism Arabidopsis thaliana

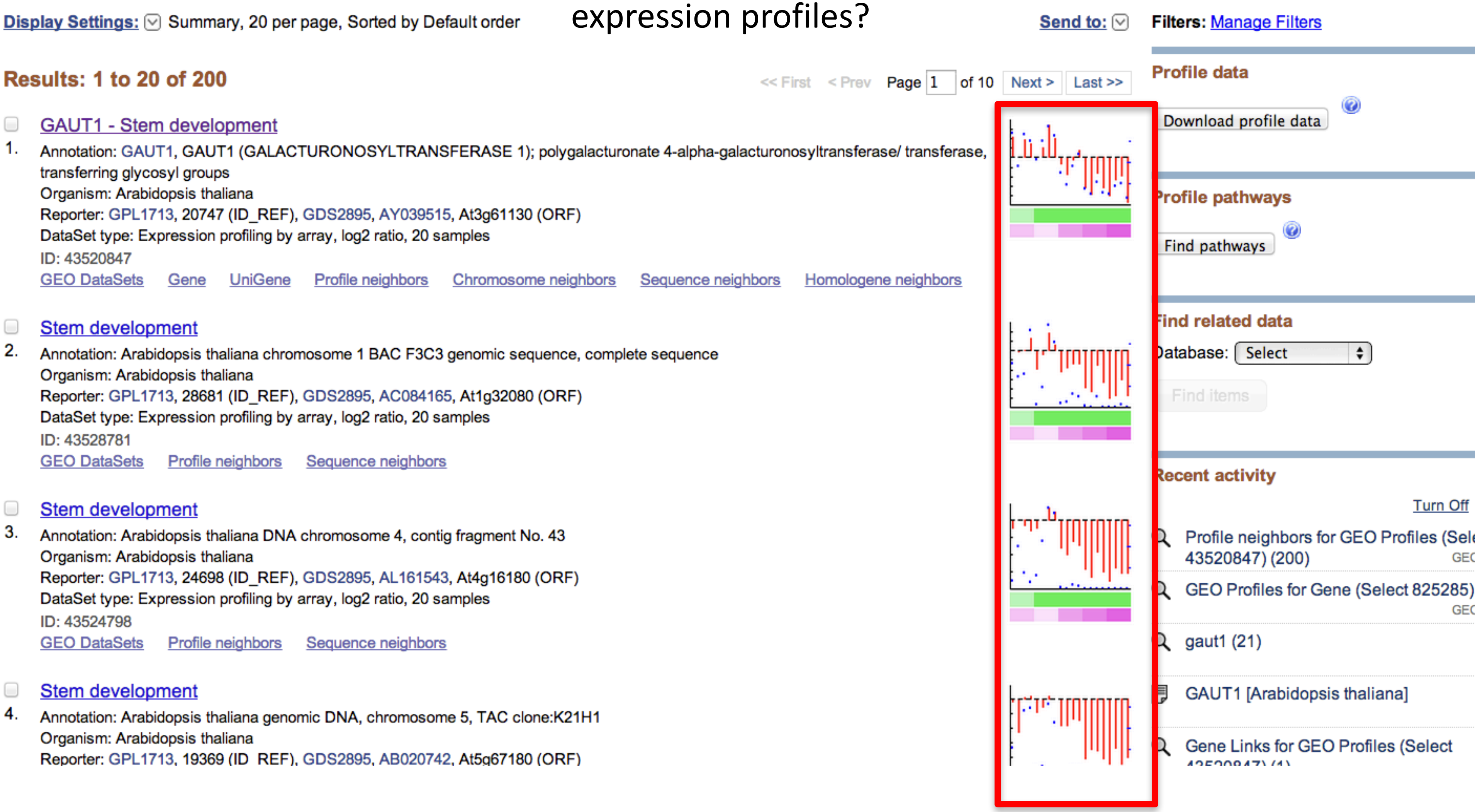


Graph caption help

Sample	Value	Rank
GSM35570	0.456367	85
GSM35571	-0.00232472	47
GSM35721	0.388293	82
GSM35725	0.297085	75
GSM35565	0.0420401	55
GSM35567	0.23765	75
GSM35568	0.7	93
GSM35569	0.421382	84

Go back to result page

Profile neighbors: what are the co-expressed genes sharing similar expression profiles?



Display Settings: Summary, 20 per page, Sorted by Default order

Send to:

Filters: Manage Filters

Results: 1 to 20 of 200

<< First < Prev Page 1 of 10 Next > Last >>

☐ 1. GAUT1 - Stem development

Annotation: GAUT1, GAUT1 (GALACTURONOSYLTRANSFERASE 1); polygalacturonate 4-alpha-galacturonosyltransferase/ transferase, transferring glycosyl groups
Organism: Arabidopsis thaliana
Reporter: GPL1713, 20747 (ID_REF), GDS2895, AY039515, At3g61130 (ORF)
DataSet type: Expression profiling by array, log2 ratio, 20 samples
ID: 43520847
GEO DataSets Gene UniGene Profile neighbors Chromosome neighbors Sequence neighbors Homologene neighbors

☐ 2. Stem development

Annotation: Arabidopsis thaliana chromosome 1 BAC F3C3 genomic sequence, complete sequence
Organism: Arabidopsis thaliana
Reporter: GPL1713, 28681 (ID_REF), GDS2895, AC084165, At1g32080 (ORF)
DataSet type: Expression profiling by array, log2 ratio, 20 samples
ID: 43528781
GEO DataSets Profile neighbors Sequence neighbors

☐ 3. Stem development

Annotation: Arabidopsis thaliana DNA chromosome 4, contig fragment No. 43
Organism: Arabidopsis thaliana
Reporter: GPL1713, 24698 (ID_REF), GDS2895, AL161543, At4g16180 (ORF)
DataSet type: Expression profiling by array, log2 ratio, 20 samples
ID: 43524798
GEO DataSets Profile neighbors Sequence neighbors

☐ 4. Stem development

Annotation: Arabidopsis thaliana genomic DNA, chromosome 5, TAC clone:K21H1
Organism: Arabidopsis thaliana
Reporter: GPL1713, 19369 (ID_REF), GDS2895, AB020742, At5g67180 (ORF)

Profile data

rofile pathways

ind related data

atabase: Select

ecent activity

Profile neighbors for GEO Profiles (Sele
43520847) (200) GE

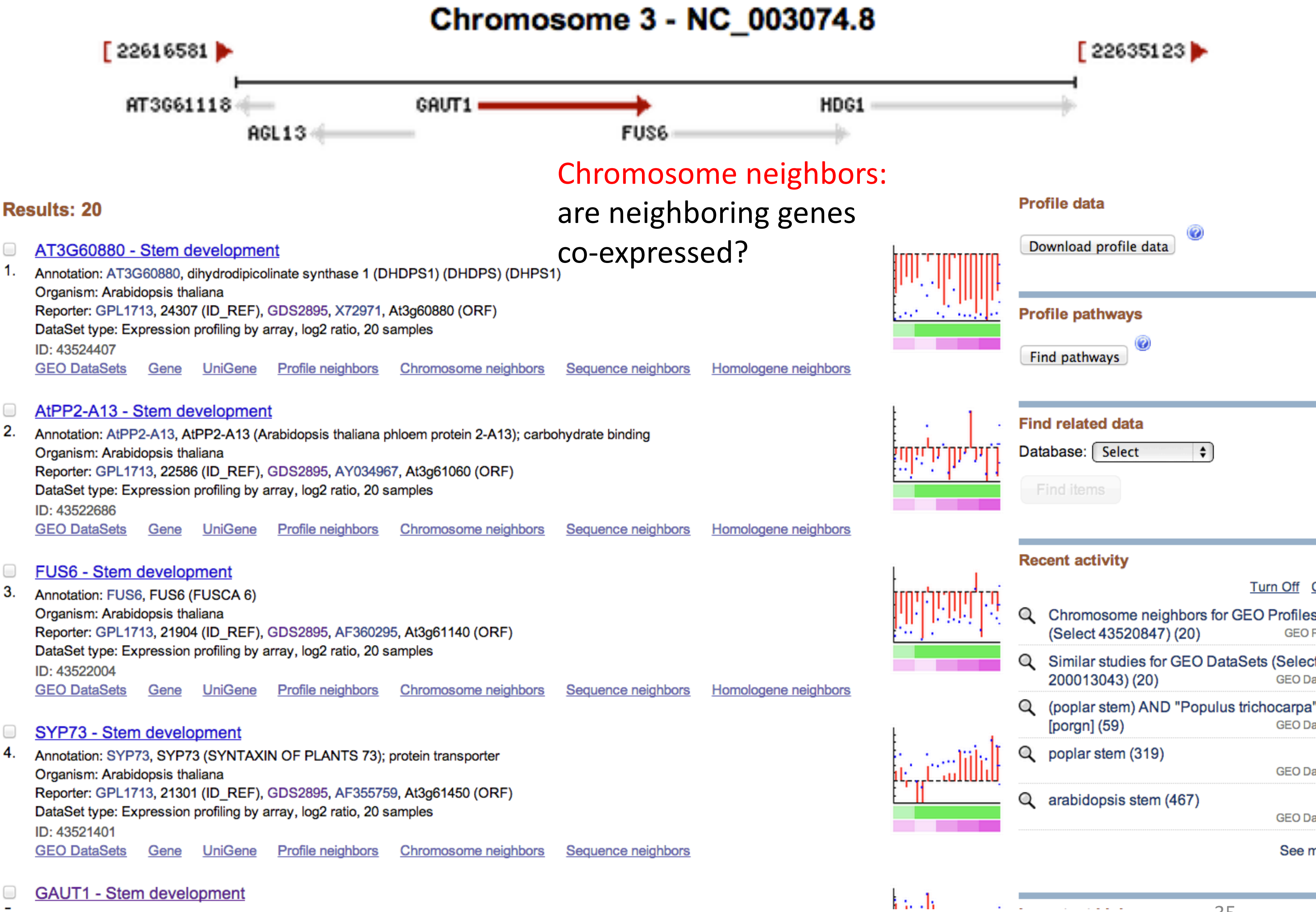
GEO Profiles for Gene (Select 825285)
GE

gaut1 (21)

GAUT1 [Arabidopsis thaliana]

Gene Links for GEO Profiles (Select

Chromosome neighbors: are neighboring genes co-expressed?



Results: 20

☐ 1. [AT3G60880 - Stem development]
Annotation: AT3G60880, dihydrodipicolinate synthase 1 (DHDPS1) (DHDPS) (DHPS1)
Organism: Arabidopsis thaliana
Reporter: GPL1713, 24307 (ID_REF), GDS2895, X72971, At3g60880 (ORF)
DataSet type: Expression profiling by array, log2 ratio, 20 samples
ID: 43524407
GEO DataSets Gene UniGene Profile neighbors Chromosome neighbors Sequence neighbors Homologene neighbors

☐ 2. [AtPP2-A13 - Stem development]
Annotation: AtPP2-A13, AtPP2-A13 (Arabidopsis thaliana phloem protein 2-A13); carbohydrate binding
Organism: Arabidopsis thaliana
Reporter: GPL1713, 22586 (ID_REF), GDS2895, AY034967, At3g61060 (ORF)
DataSet type: Expression profiling by array, log2 ratio, 20 samples
ID: 43522686
GEO DataSets Gene UniGene Profile neighbors Chromosome neighbors Sequence neighbors Homologene neighbors

☐ 3. [FUS6 - Stem development]
Annotation: FUS6, FUS6 (FUSCA 6)
Organism: Arabidopsis thaliana
Reporter: GPL1713, 21904 (ID_REF), GDS2895, AF360295, At3g61140 (ORF)
DataSet type: Expression profiling by array, log2 ratio, 20 samples
ID: 43522004
GEO DataSets Gene UniGene Profile neighbors Chromosome neighbors Sequence neighbors Homologene neighbors

☐ 4. [SYP73 - Stem development]
Annotation: SYP73, SYP73 (SYNTAXIN OF PLANTS 73); protein transporter
Organism: Arabidopsis thaliana
Reporter: GPL1713, 21301 (ID_REF), GDS2895, AF355759, At3g61450 (ORF)
DataSet type: Expression profiling by array, log2 ratio, 20 samples
ID: 43521401
GEO DataSets Gene UniGene Profile neighbors Chromosome neighbors Sequence neighbors

☐ [GAUT1 - Stem development]

Profile data

Profile pathways

Find related data

Database: Select

Recent activity

Turn Off C

Chromosome neighbors for GEO Profiles (Select 43520847) (20) GEO P

Similar studies for GEO DataSets (Selec 200013043) (20) GEO Da

(poplar stem) AND "Populus trichocarpa" [porgn] (59) GEO Da

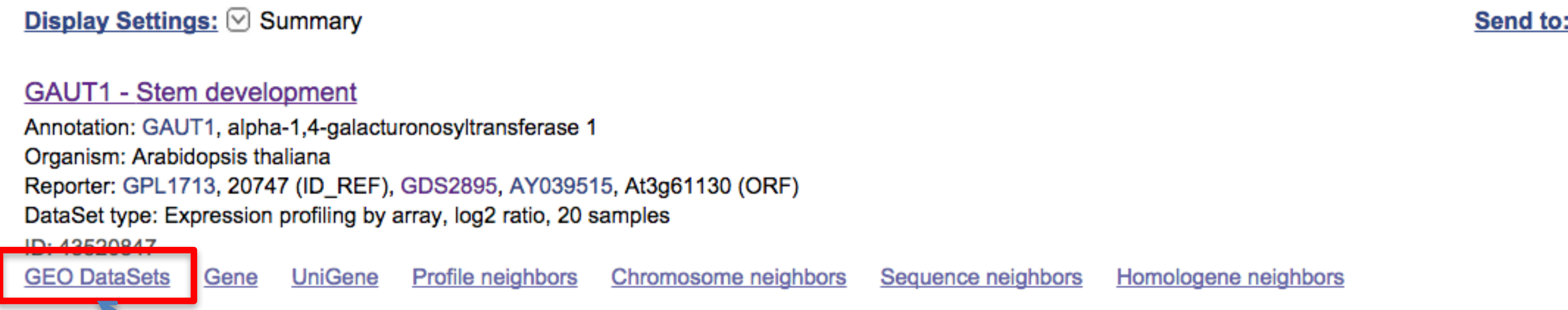
poplar stem (319) GEO Da

arabidopsis stem (467) GEO Da

See m

Hands on exercise 3

GEO DataSets analysis tool



Display Settings: Summary

Send to:

GAUT1 - Stem development

Annotation: GAUT1, alpha-1,4-galacturonosyltransferase 1

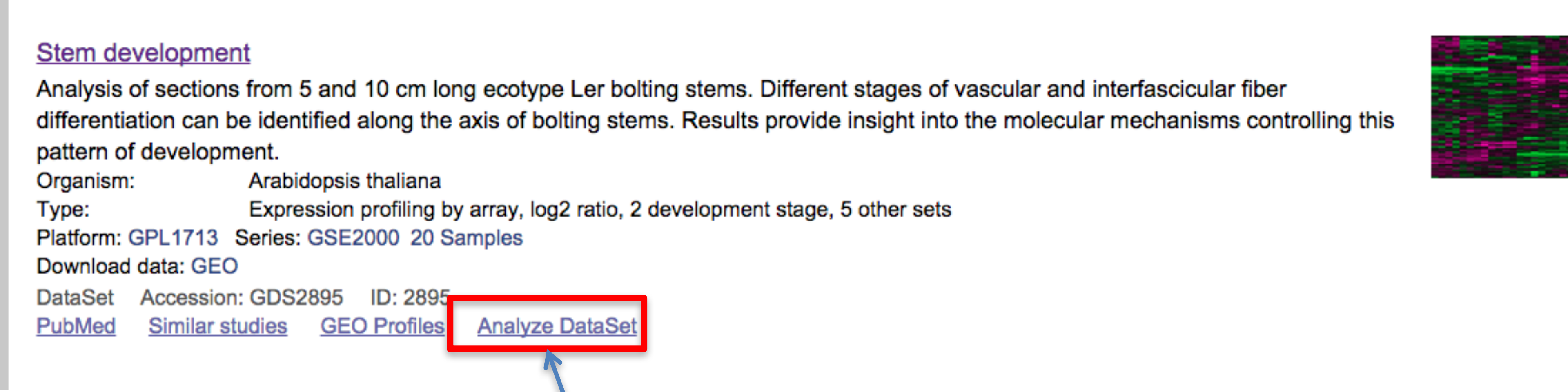
Organism: Arabidopsis thaliana

Reporter: GPL1713, 20747 (ID_REF), GDS2895, AY039515, At3g61130 (ORF)

DataSet type: Expression profiling by array, log2 ratio, 20 samples

ID: 43520847

Click



Stem development

Analysis of sections from 5 and 10 cm long ecotype Ler bolting stems. Different stages of vascular and interfascicular fiber differentiation can be identified along the axis of bolting stems. Results provide insight into the molecular mechanisms controlling this pattern of development.

Organism:	Arabidopsis thaliana
Type:	Expression profiling by array, log2 ratio, 2 development stage, 5 other sets

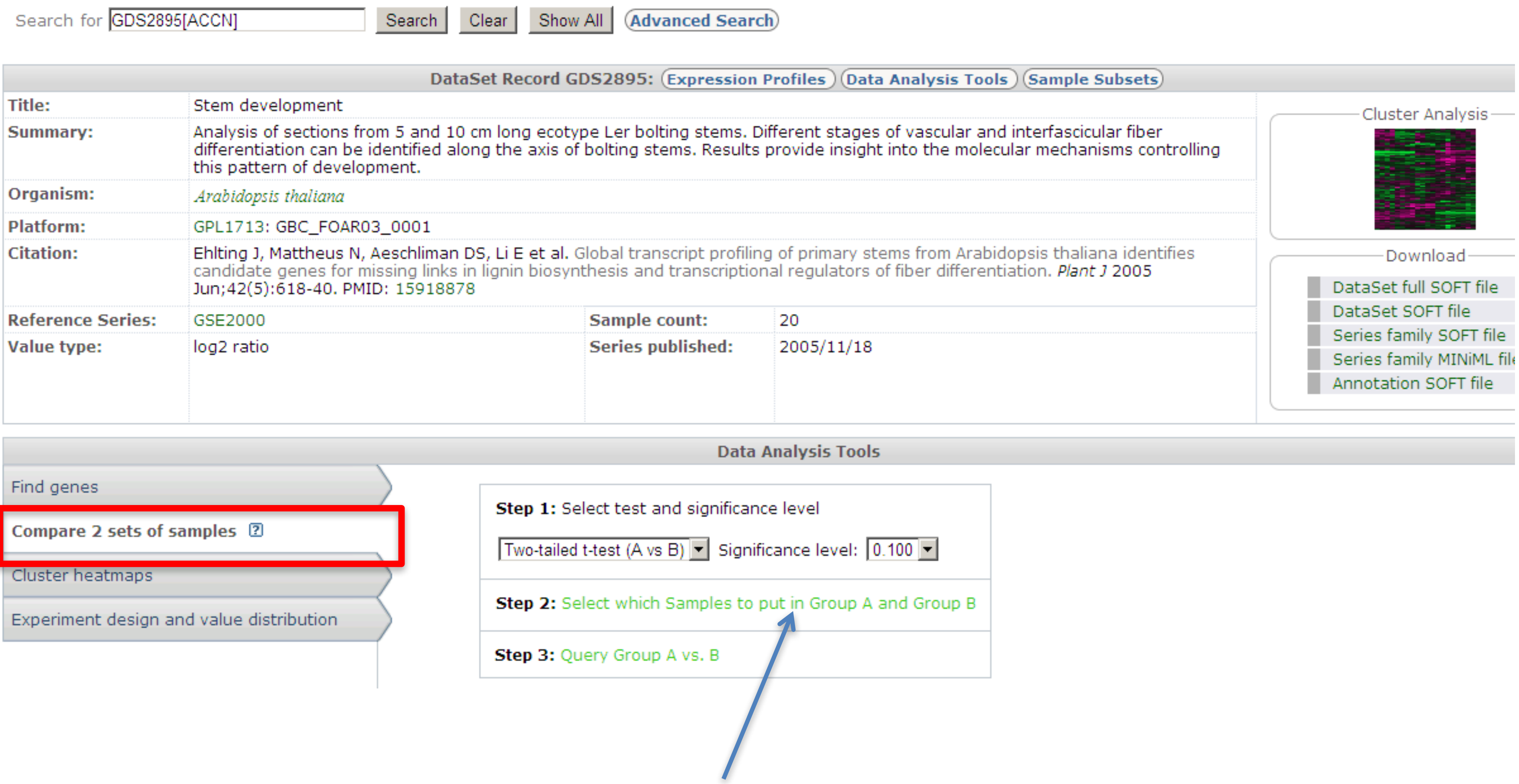
Platform: GPL1713 Series: GSE2000 20 Samples

Download data: GEO

DataSet Accession: GDS2895 ID: 2895

Click

We want to use this DataSet to identify differentially expressed genes in stem development
How: define two groups of samples and run two sample t test



Search for GDS2895[ACCN]

DataSet Record GDS2895:

Title:	Stem development		
Summary:	Analysis of sections from 5 and 10 cm long ecotype Ler bolting stems. Different stages of vascular and interfascicular fiber differentiation can be identified along the axis of bolting stems. Results provide insight into the molecular mechanisms controlling this pattern of development.		
Organism:	*Arabidopsis thaliana*		
Platform:	GPL1713: GBC_FOAR03_0001		
Citation:	Ehlting J, Mattheus N, Aeschliman DS, Li E et al. Global transcript profiling of primary stems from Arabidopsis thaliana identifies candidate genes for missing links in lignin biosynthesis and transcriptional regulators of fiber differentiation. *Plant J* 2005 Jun;42(5):618-40. PMID: 15918878		
Reference Series:	GSE2000	**Sample count:**	20
Value type:	log2 ratio	**Series published:**	2005/11/18

Data Analysis Tools

Step 1: Select test and significance level

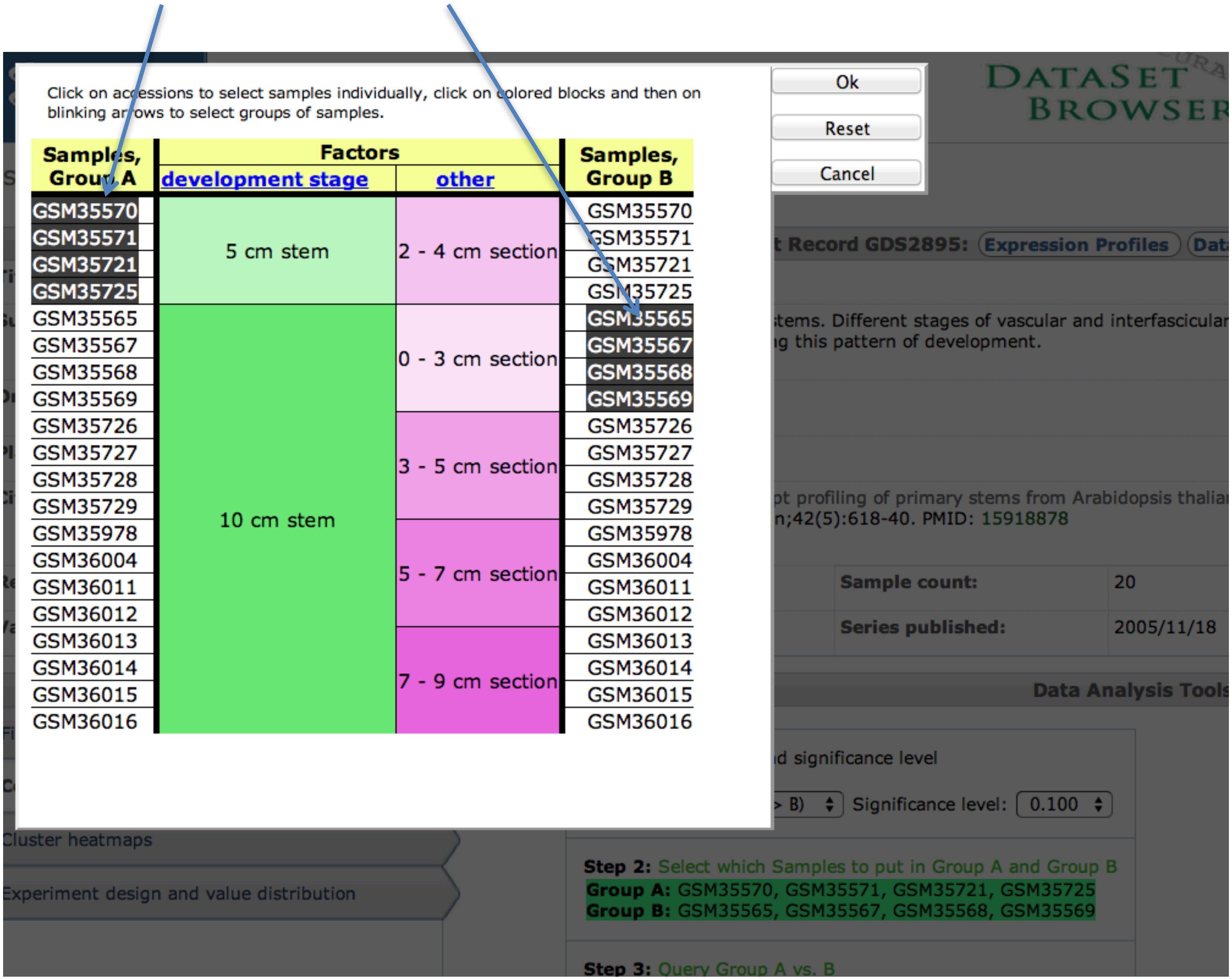
Two-tailed t-test (A vs B) Significance level: 0.100

Step 2: Select which Samples to put in Group A and Group B

Step 3: Query Group A vs. B

Click on step 2 to define two groups of samples

Click samples to select



Click on accessions to select samples individually, click on colored blocks and then on blinking arrows to select groups of samples.

Samples, Group A	Factors		Samples, Group B
	development stage	other	
GSM35570	5 cm stem	2 - 4 cm section	GSM35570
GSM35571			GSM35571
GSM35721			GSM35721
GSM35725			GSM35725
GSM35565	10 cm stem	0 - 3 cm section	GSM35565
GSM35567			GSM35567
GSM35568			GSM35568
GSM35569			GSM35569
GSM35726		3 - 5 cm section	GSM35726
GSM35727			GSM35727
GSM35728			GSM35728
GSM35729			GSM35729
GSM35978		5 - 7 cm section	GSM35978
GSM36004			GSM36004
GSM36011			GSM36011
GSM36012			GSM36012
GSM36013		7 - 9 cm section	GSM36013
GSM36014			GSM36014
GSM36015			GSM36015
GSM36016			GSM36016

t Record GDS2895: Expression Profiles Dat

tems. Different stages of vascular and interfascicular
g this pattern of development.

pt profiling of primary stems from Arabidopsis thalia
n;42(5):618-40. PMID: 15918878

Sample count:	20
Series published:	2005/11/18

Data Analysis Tools

d significance level

> B) Significance level: 0.100

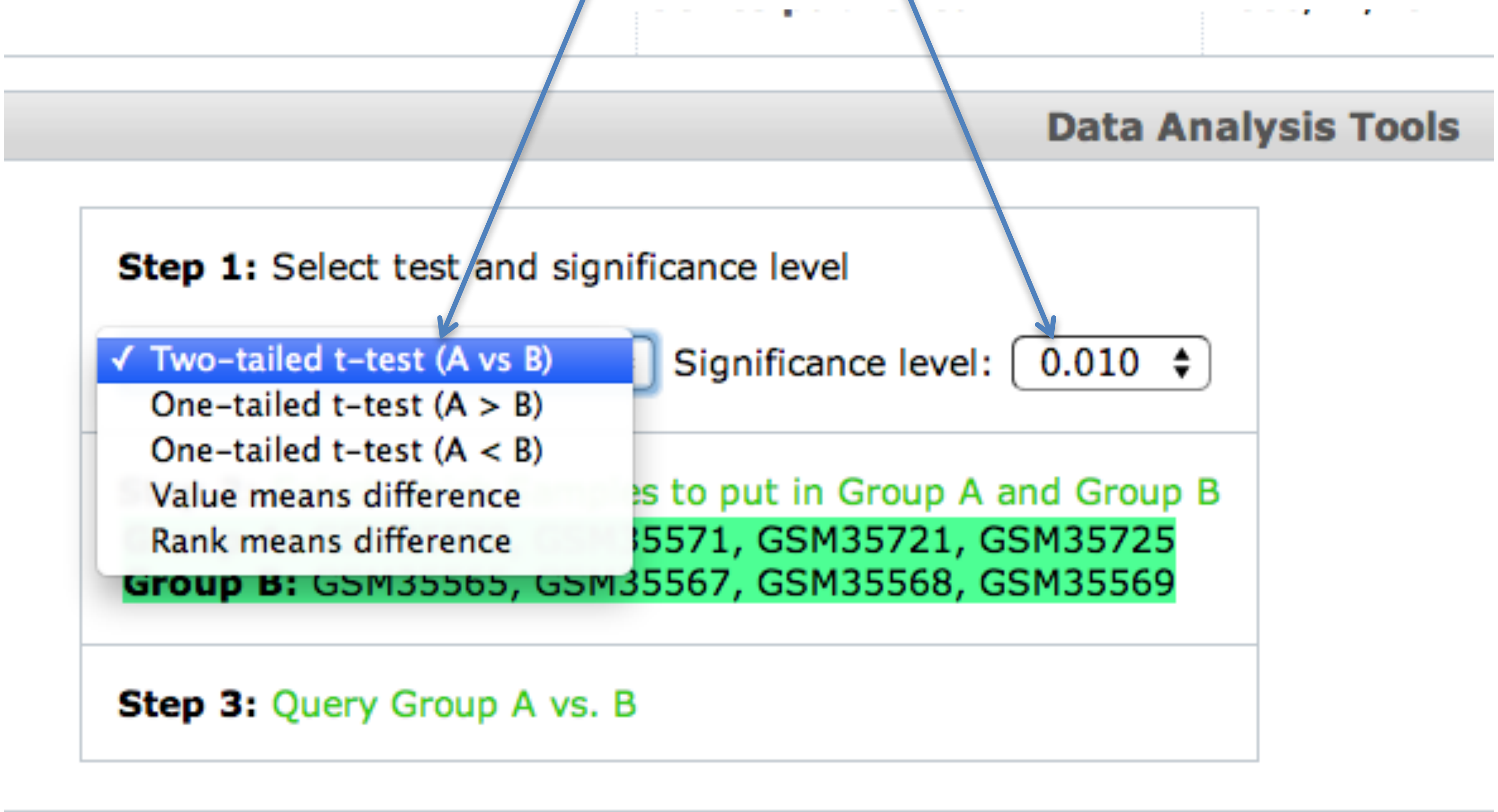
Cluster heatmaps

Experiment design and value distribution

Step 2: Select which Samples to put in Group A and Group B
Group A: GSM35570, GSM35571, GSM35721, GSM35725
Group B: GSM35565, GSM35567, GSM35568, GSM35569

Step 3: Query Group A vs. B

Step 1: you can choose different statistical methods for analysis



Step 1: Select test and significance level

Significance level: 0.010

es to put in Group A and Group B

5571, GSM35721, GSM35725

Group B: GSM35565, GSM35567, GSM35568, GSM35569

Step 3: Query Group A vs. B

Step 3 to perform analysis

Result page is a list of genes with significantly different expression between two groups of samples

Group 2

Group 1

Results: 1 to 20 of 570

<< First < Prev Page 1 of 29 Next > Last >>

☐ bZIP42 - Stem development

1. Annotation: bZIP42, basic leucine-zipper 42
Organism: Arabidopsis thaliana
Reporter: GPL1713, 389 (ID_REF), GDS2895, AB016878, At3g30530 (ORF)
DataSet type: Expression profiling by array, log2 ratio, 20 samples
ID: 43500489
GEO DataSets Gene UniGene Profile neighbors Chromosome neighbors Sequence neighbors

☐ AT1G17200 - Stem development

2. Annotation: AT1G17200, hypothetical protein
Organism: Arabidopsis thaliana
Reporter: GPL1713, 645 (ID_REF), GDS2895, At1g17200 (ORF)
DataSet type: Expression profiling by array, log2 ratio, 20 samples
ID: 43500745
GEO DataSets Gene UniGene Profile neighbors Chromosome neighbors Homologene neighbors

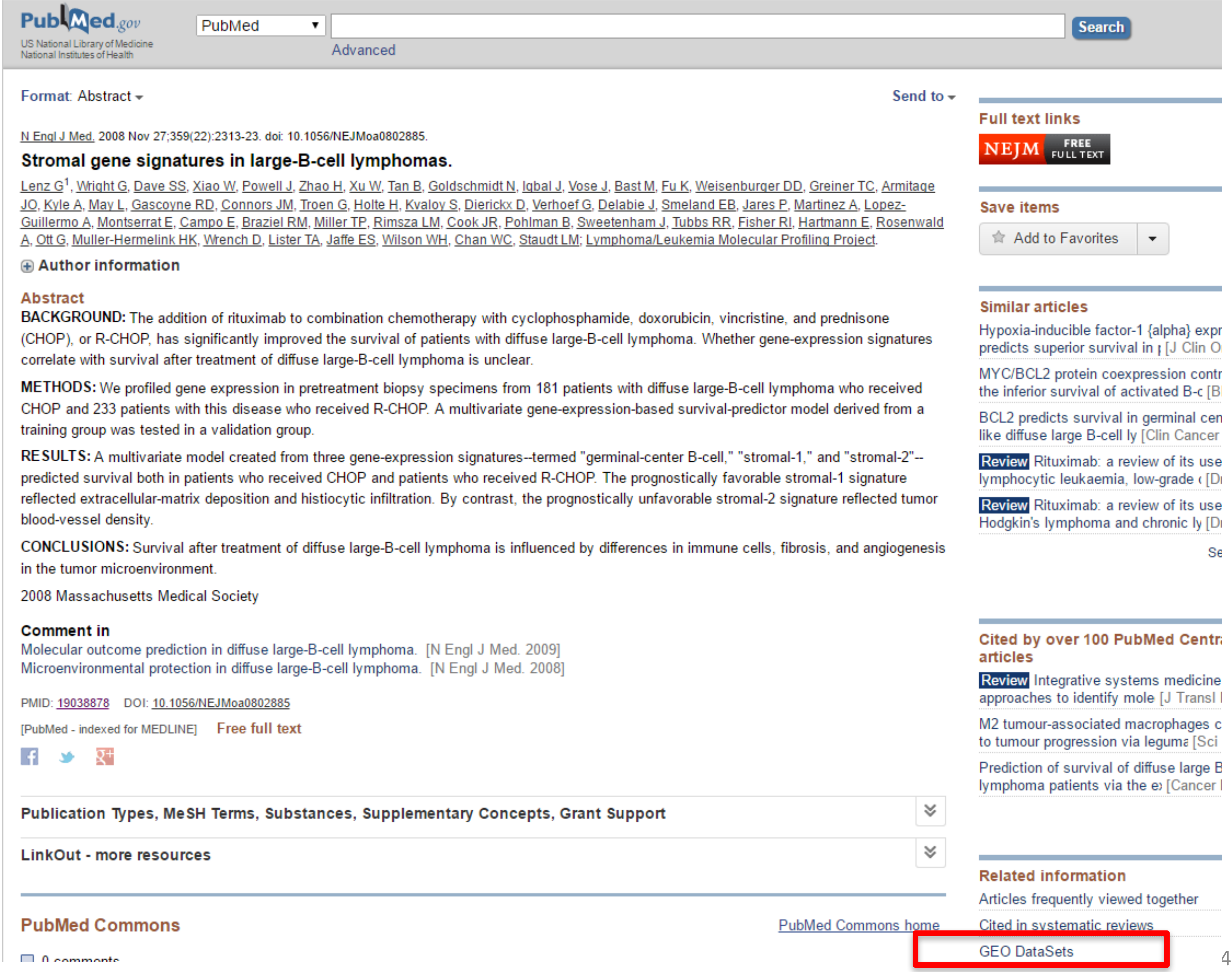
☐ AGL8 - Stem development

3. Annotation: AGL8, agamous-like MADS-box protein AGL8
Organism: Arabidopsis thaliana
Reporter: GPL1713, 773 (ID_REF), GDS2895, U33473, At5g60910 (ORF)
DataSet type: Expression profiling by array, log2 ratio, 20 samples
ID: 43500873

“Analyze DataSet” is for GEO DataSets
“GEO2R” is for GEO Series

GEO2R: differentially expressed genes

http://www.youtube.com/watch?v=EUPmGWS8ik0



N Engl J Med. 2008 Nov 27;359(22):2313-23. doi: 10.1056/NEJMoa0802885.

Stromal gene signatures in large-B-cell lymphomas.

Lenz G[1], Wright G, Dave SS, Xiao W, Powell J, Zhao H, Xu W, Tan B, Goldschmidt N, Iqbal J, Vose J, Bast M, Fu K, Weisenburger DD, Greiner TC, Armitage JO, Kyle A, May L, Gascoyne RD, Connors JM, Troen G, Holte H, Kvaloy S, Dierickx D, Verhoef G, Delabie J, Smeland EB, Jares P, Martinez A, Lopez-Guillermo A, Montserrat E, Campo E, Braziel RM, Miller TP, Rimsza LM, Cook JR, Pohlman B, Sweetenham J, Tubbs RR, Fisher RI, Hartmann E, Rosenwald A, Ott G, Muller-Hermelink HK, Wrench D, Lister TA, Jaffe ES, Wilson WH, Chan WC, Staudt LM; Lymphoma/Leukemia Molecular Profiling Project.

⊕ Author information

Abstract

BACKGROUND: The addition of rituximab to combination chemotherapy with cyclophosphamide, doxorubicin, vincristine, and prednisone (CHOP), or R-CHOP, has significantly improved the survival of patients with diffuse large-B-cell lymphoma. Whether gene-expression signatures correlate with survival after treatment of diffuse large-B-cell lymphoma is unclear.

METHODS: We profiled gene expression in pretreatment biopsy specimens from 181 patients with diffuse large-B-cell lymphoma who received CHOP and 233 patients with this disease who received R-CHOP. A multivariate gene-expression-based survival-predictor model derived from a training group was tested in a validation group.

RESULTS: A multivariate model created from three gene-expression signatures--termed "germinal-center B-cell," "stromal-1," and "stromal-2"--predicted survival both in patients who received CHOP and patients who received R-CHOP. The prognostically favorable stromal-1 signature reflected extracellular-matrix deposition and histiocytic infiltration. By contrast, the prognostically unfavorable stromal-2 signature reflected tumor blood-vessel density.

CONCLUSIONS: Survival after treatment of diffuse large-B-cell lymphoma is influenced by differences in immune cells, fibrosis, and angiogenesis in the tumor microenvironment.

2008 Massachusetts Medical Society

Comment in

Molecular outcome prediction in diffuse large-B-cell lymphoma. [N Engl J Med. 2009]
Microenvironmental protection in diffuse large-B-cell lymphoma. [N Engl J Med. 2008]

PMID: 19038878 DOI: 10.1056/NEJMoa0802885

[PubMed - indexed for MEDLINE] Free full text

Publication Types, MeSH Terms, Substances, Supplementary Concepts, Grant Support

LinkOut - more resources

PubMed Commons

PubMed Commons home

☐ 0 comments

Full text links

NEJM FREE FULL TEXT

Save items

☆ Add to Favorites

Similar articles

Hypoxia-inducible factor-1 {alpha} expr predicts superior survival in ɪ [J Clin O

MYC/BCL2 protein coexpression contr the inferior survival of activated B-c [Bl

BCL2 predicts survival in germinal cen like diffuse large B-cell ly [Clin Cancer

Review Rituximab: a review of its use lymphocytic leukaemia, low-grade ([Dr

Review Rituximab: a review of its use Hodgkin's lymphoma and chronic ly [Dr

Se

Cited by over 100 PubMed Centr articles

Review Integrative systems medicine approaches to identify mole [J Transl

M2 tumour-associated macrophages c to tumour progression via legumε [Sci

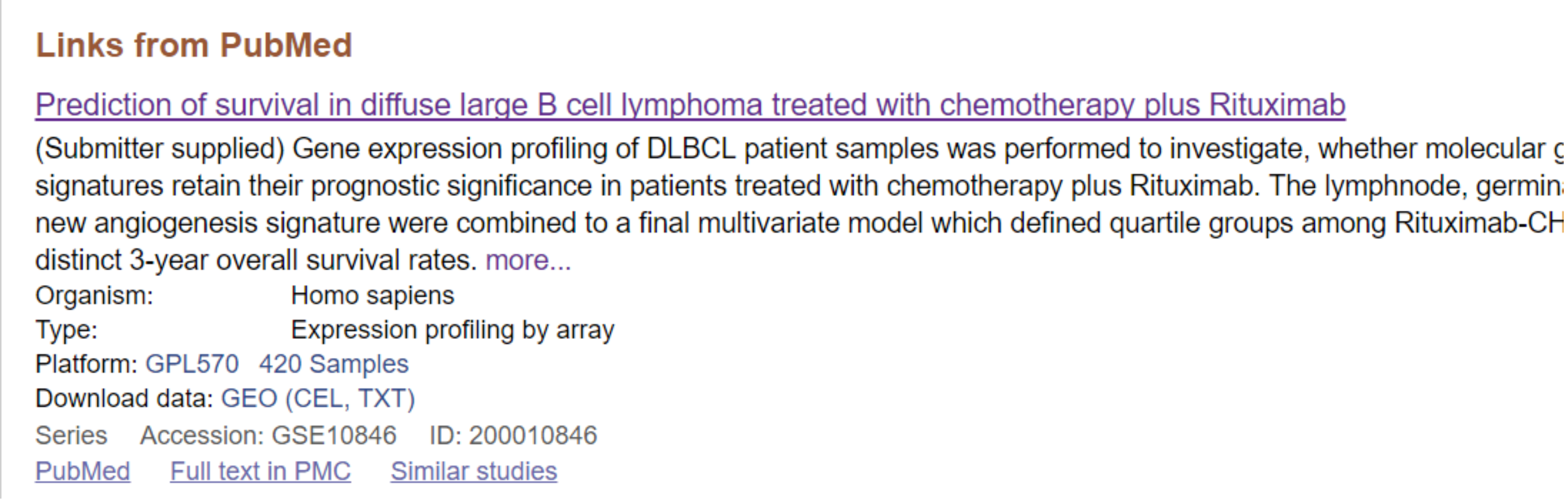
Prediction of survival of diffuse large B lymphoma patients via the e› [Cancer

Related information

Articles frequently viewed together

Cited in systematic reviews

GEO DataSets



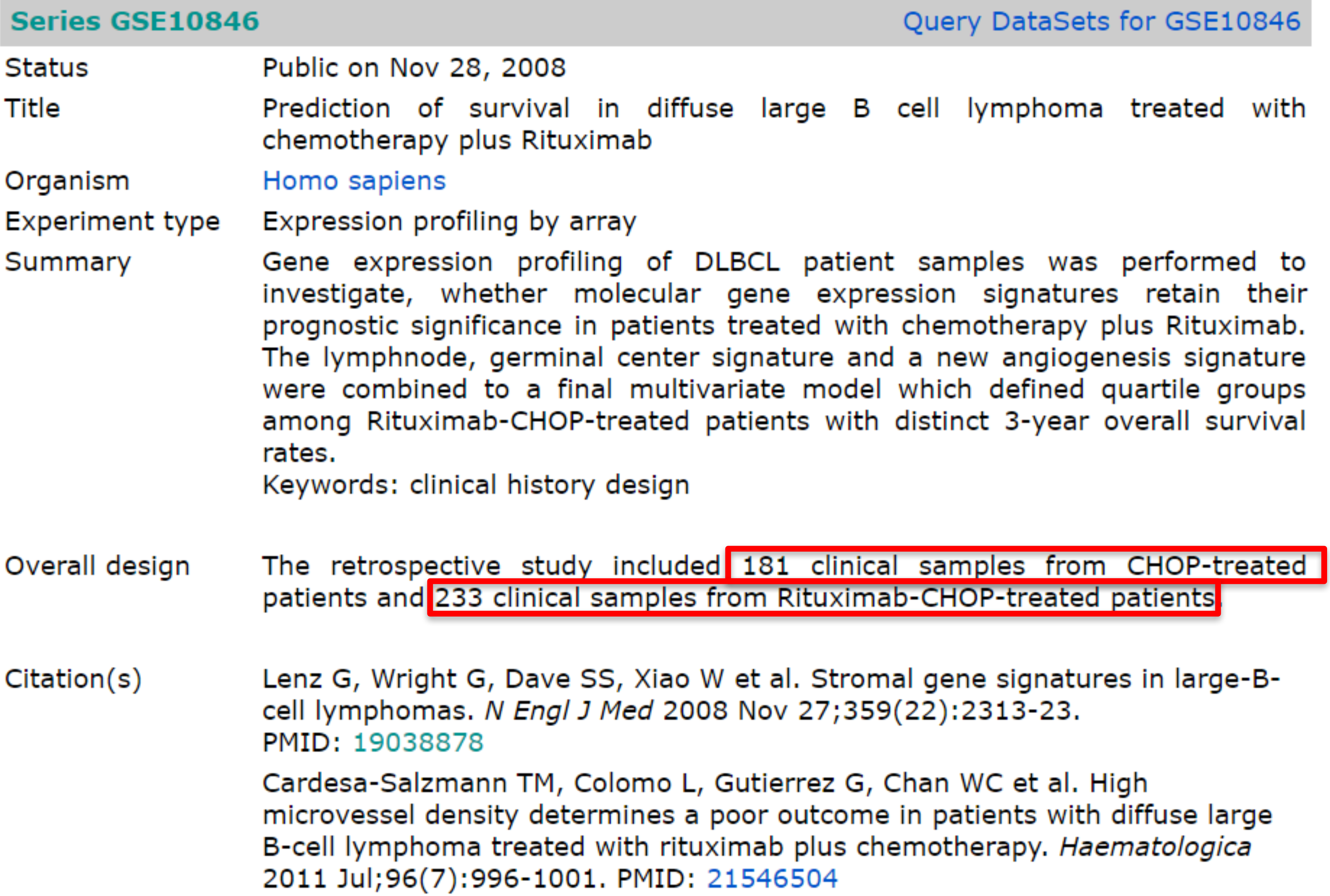
Links from PubMed

Prediction of survival in diffuse large B cell lymphoma treated with chemotherapy plus Rituximab

(Submitter supplied) Gene expression profiling of DLBCL patient samples was performed to investigate, whether molecular g
signatures retain their prognostic significance in patients treated with chemotherapy plus Rituximab. The lymphnode, germin
new angiogenesis signature were combined to a final multivariate model which defined quartile groups among Rituximab-CH
distinct 3-year overall survival rates. more...

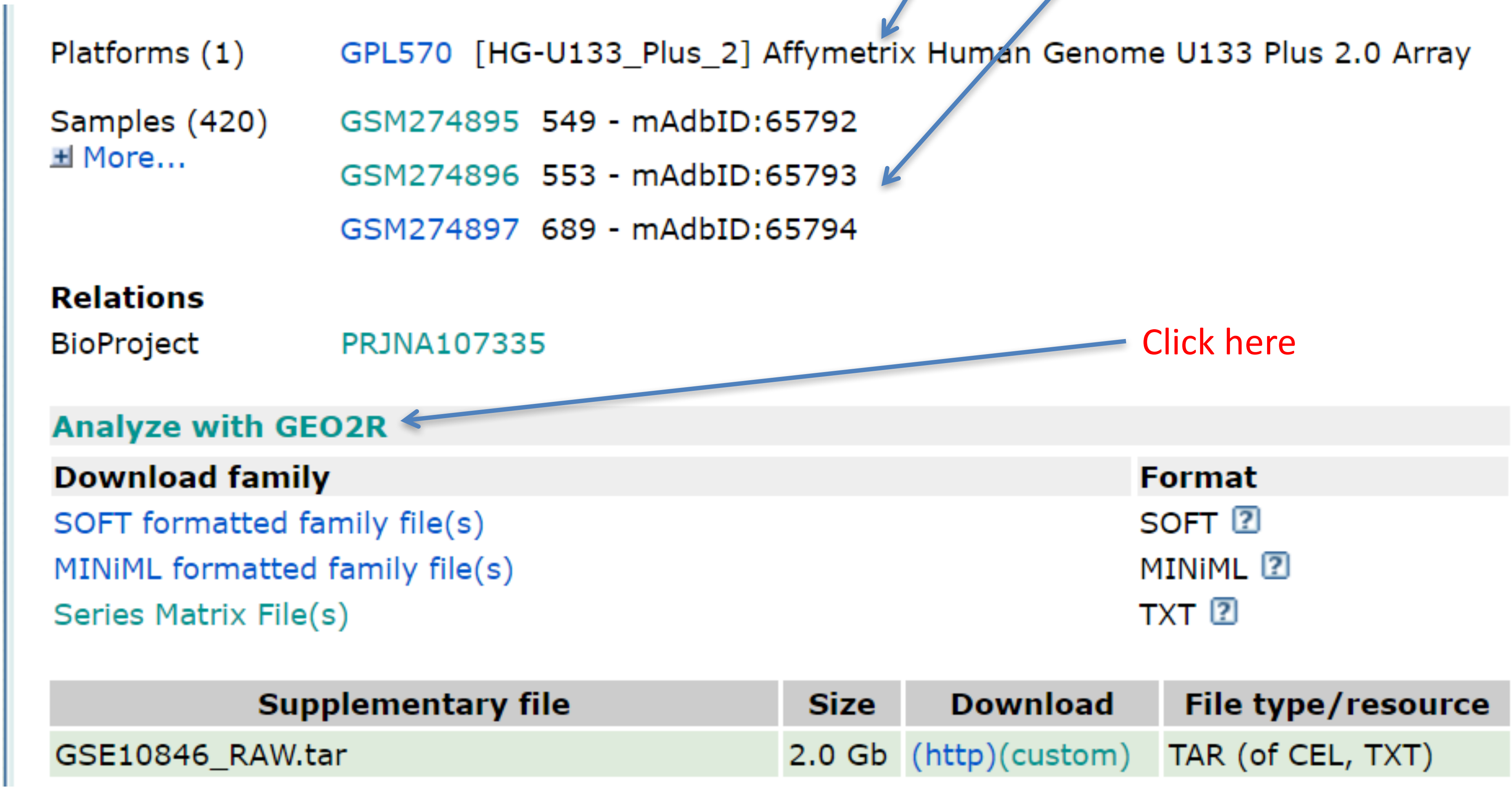
Organism:	Homo sapiens
Type:	Expression profiling by array

Platform: GPL570 420 Samples
Download data: GEO (CEL, TXT)
Series Accession: GSE10846 ID: 200010846



Series GSE10846	Query DataSets for GSE10846
Status	Public on Nov 28, 2008
Title	Prediction of survival in diffuse large B cell lymphoma treated with chemotherapy plus Rituximab
Organism	Homo sapiens
Experiment type	Expression profiling by array
Summary	Gene expression profiling of DLBCL patient samples was performed to investigate, whether molecular gene expression signatures retain their prognostic significance in patients treated with chemotherapy plus Rituximab. The lymphnode, germinal center signature and a new angiogenesis signature were combined to a final multivariate model which defined quartile groups among Rituximab-CHOP-treated patients with distinct 3-year overall survival rates. Keywords: clinical history design
Overall design	The retrospective study included 181 clinical samples from CHOP-treated patients and 233 clinical samples from Rituximab-CHOP-treated patients.
Citation(s)	Lenz G, Wright G, Dave SS, Xiao W et al. Stromal gene signatures in large-B-cell lymphomas. *N Engl J Med* 2008 Nov 27;359(22):2313-23. PMID: 19038878 Cardesa-Salzmann TM, Colomo L, Gutierrez G, Chan WC et al. High microvessel density determines a poor outcome in patients with diffuse large B-cell lymphoma treated with rituximab plus chemotherapy. *Haematologica* 2011 Jul;96(7):996-1001. PMID: 21546504

Platform and sample data



Platforms (1) GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array

Samples (420) GSM274895 549 - mAdbID:65792
GSM274896 553 - mAdbID:65793
GSM274897 689 - mAdbID:65794

Relations

BioProject PRJNA107335

Click here

Analyze with GEO2R

Download family	**Format**
SOFT formatted family file(s)	SOFT ?
MINiML formatted family file(s)	MINiML ?
Series Matrix File(s)	TXT ?

Supplementary file	**Size**	**Download**	**File type/resource**
GSE10846_RAW.tar	2.0 Gb	(http)(custom)	TAR (of CEL, TXT)

Samples (420)
Less...

GSM274895 549 - mAdbID:65792
GSM274896 553 - mAdbID:65793
GSM274897 689 - mAdbID:65794
GSM274898 704 - mAdbID:65795
GSM274899 694 - mAdbID:65796
GSM274900 756 - mAdbID:65797
GSM274901 547 - mAdbID:65798
GSM274902 841 - mAdbID:65799
GSM274903 745 - mAdbID:65801

……

GSM275072 1007 - mAdbID:74487
GSM275073 1018 - mAdbID:74488
GSM275074 1054 - mAdbID:74489
GSM275075 1041 - mAdbID:74490

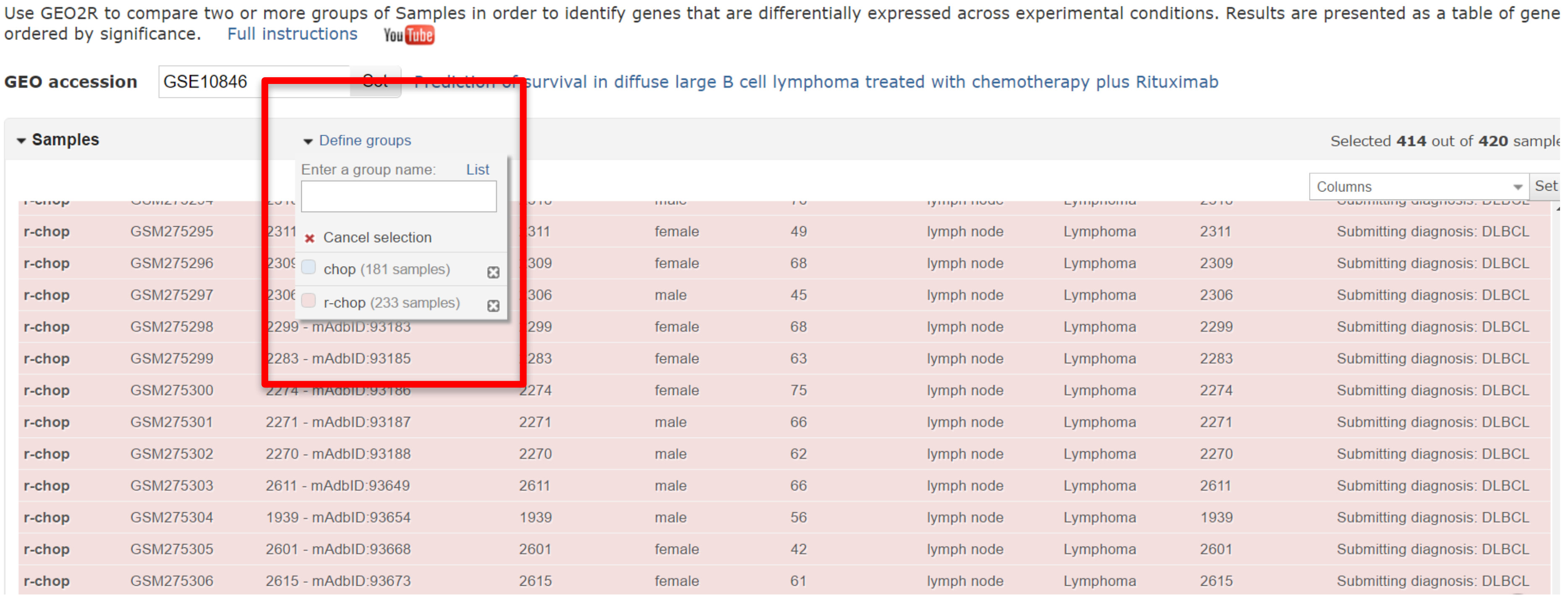
GSM275076 1179 - mAdbID:87725
GSM275077 1181 - mAdbID:87727
GSM275078 1184 - mAdbID:87730
GSM275079 1185 - mAdbID:87731

181 CHOP

4895+181-1=5075

233 R-CHOP

Click on Define groups and type in group names
Select samples from the table and click on the defined group to assign to the group
Click on Top 250 in the bottom of the page to run the job



Use GEO2R to compare two or more groups of Samples in order to identify genes that are differentially expressed across experimental conditions. Results are presented as a table of gene ordered by significance. Full instructions

GEO accession GSE10846 Prediction of survival in diffuse large B cell lymphoma treated with chemotherapy plus Rituximab

Samples

Define groups

Enter a group name: List

Cancel selection

☐ chop (181 samples)

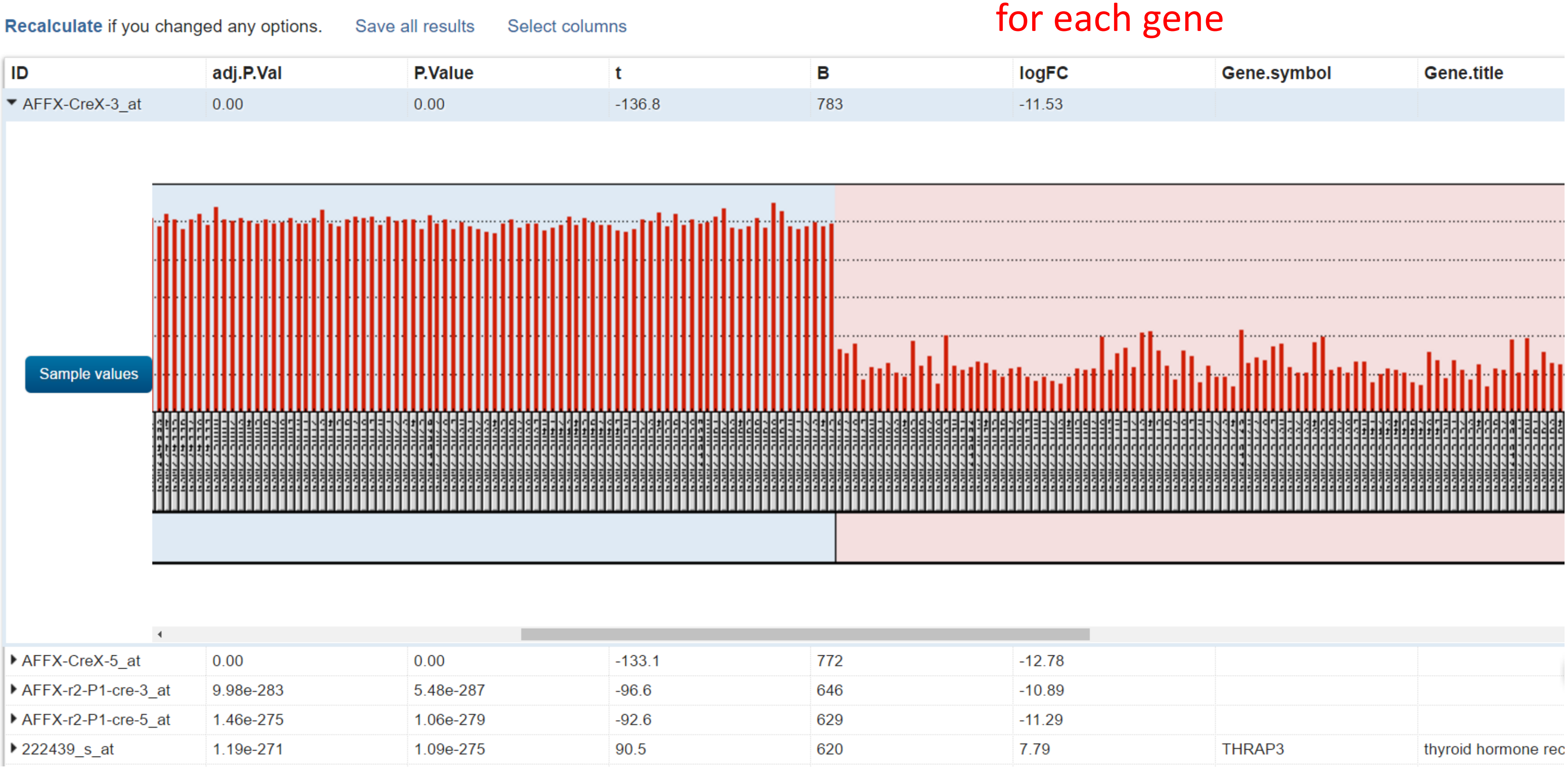
☐ r-chop (233 samples)

Selected **414** out of **420** sample

r-chop	GSM275295	2311	311	female	49	lymph node	Lymphoma	2311	Submitting diagnosis: DLBCL
r-chop	GSM275296	2309	309	female	68	lymph node	Lymphoma	2309	Submitting diagnosis: DLBCL
r-chop	GSM275297	2306	306	male	45	lymph node	Lymphoma	2306	Submitting diagnosis: DLBCL
r-chop	GSM275298	2299 - mAdbID:93183	299	female	68	lymph node	Lymphoma	2299	Submitting diagnosis: DLBCL
r-chop	GSM275299	2283 - mAdbID:93185	283	female	63	lymph node	Lymphoma	2283	Submitting diagnosis: DLBCL
r-chop	GSM275300	2274 - mAdbID:93186	2274	female	75	lymph node	Lymphoma	2274	Submitting diagnosis: DLBCL
r-chop	GSM275301	2271 - mAdbID:93187	2271	male	66	lymph node	Lymphoma	2271	Submitting diagnosis: DLBCL
r-chop	GSM275302	2270 - mAdbID:93188	2270	male	62	lymph node	Lymphoma	2270	Submitting diagnosis: DLBCL
r-chop	GSM275303	2611 - mAdbID:93649	2611	male	66	lymph node	Lymphoma	2611	Submitting diagnosis: DLBCL
r-chop	GSM275304	1939 - mAdbID:93654	1939	male	56	lymph node	Lymphoma	1939	Submitting diagnosis: DLBCL
r-chop	GSM275305	2601 - mAdbID:93668	2601	female	42	lymph node	Lymphoma	2601	Submitting diagnosis: DLBCL
r-chop	GSM275306	2615 - mAdbID:93673	2615	female	61	lymph node	Lymphoma	2615	Submitting diagnosis: DLBCL

The result page, click on the ID will give the graph

The 2 groups have different profiles for each gene



Recalculate if you changed any options. Save all results Select columns

ID	adj.P.Val	P.Value	t	B	logFC	Gene.symbol	Gene.title
AFFX-CreX-3_at	0.00	0.00	-136.8	783	-11.53		
AFFX-CreX-5_at	0.00	0.00	-133.1	772	-12.78		
AFFX-r2-P1-cre-3_at	9.98e-283	5.48e-287	-96.6	646	-10.89		
AFFX-r2-P1-cre-5_at	1.46e-275	1.06e-279	-92.6	629	-11.29		
222439_s_at	1.19e-271	1.09e-275	90.5	620	7.79	THRAP3	thyroid hormone rec

You may check the detailed info of the gene in GPL table

Next lecture: `EBI resources I`