

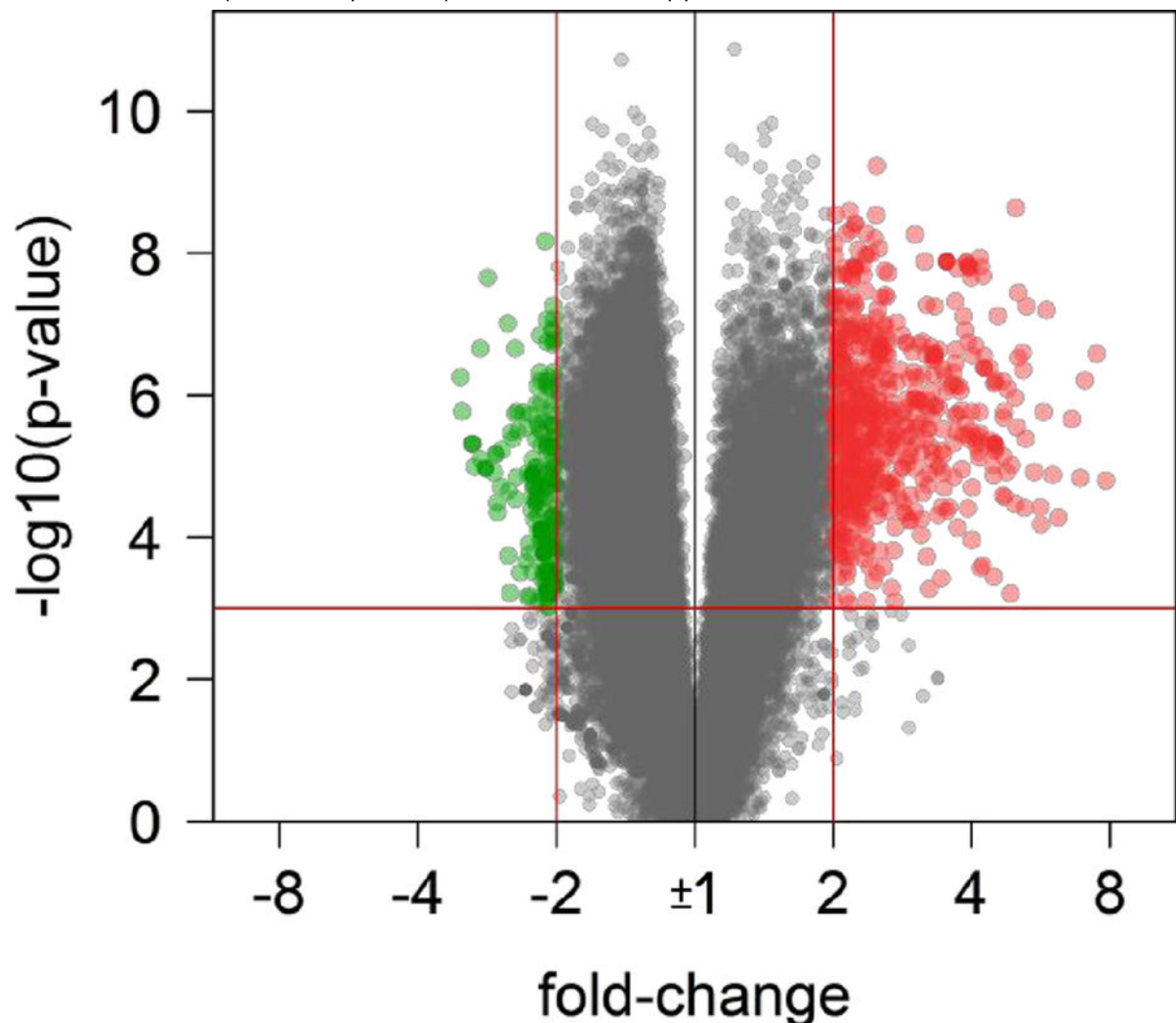
Lab 17

Microarray analysis of squamous cell lung cancer

BACKGROUND: Lung cancer is a heterogeneous disease with several well-defined subtypes. One of these, squamous cell cancer (SCC), generally arises from the bronchial airway epithelium, and has features reminiscent of squamous epithelium, which includes flat, layered cells that line cavities such as the mouth or throat, or epidermal skin cells (aka keratinocytes).

In today's lab you will use three web resources to see which changes in gene expression occur in squamous lung tumors. The first of these, the Gene Expression Omnibus (GEO), is the largest repository of public gene expression data in the world. The second of these, DAVID, is a free set of tools from the National Institutes of Health (NIH).

Some Jupyter. Embed images into Jupyter notebook from anywhere! Change this cell from Raw NBConvert (a.k.a. simple text) and see what happens.



PROCEDURES, Part 1 (GEO):

Today you'll be performing analysis on a GEO DataSet. GEO DataSets are sets of raw gene expression data (GEO Series) that have been manually curated so that analysis can be performed quickly between two groups of samples (e.g., normal vs. disease, drug A vs. drug B).

Start by opening the GEO start page at: <http://www.ncbi.nlm.nih.gov/geo/>

NCBI

Resources

How To

Sign in to NCBI

GEO Home

Documentation

Query & Browse

Email 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.

Keyword or GEO Accession

Search

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: 3848

Series: 55892

Platforms: 14065

Samples: 1362090

Information for Submitters

Login to Submit

Submission Guidelines

Update Guidelines

MIAME Standards

Citing and Linking to GEO

Guidelines for Reviewers

GEO Publications

The GEO Query interface is organized according to the different types of results. Click Search for Studies under Tools and then type "squamous cell lung cancer" into the DataSets text box and click SEARCH.

NCBI

Resources

How To

Sign in to NCBI

GEO DataSets

GEO DataSets

squamous cell lung cancer

Search

Save search

Advanced

Help

Show additional filters

Entry type

DataSets (14)

Series (222)

Samples (2038)

Platforms (9)

Organism

Select ...

Study type

Expression profiling by array

Methylation profiling by array

More ...

Author

Select ...

Attribute name

tissue

strain

More ...

Publication dates

30 days

1 year

Custom range...

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

Results: 1 to 20 of 2283

1. [Small cell lung cancers](#)

Analysis of 23 clinical small cell lung cancer (SCLC) samples from patients undergoing pulmonary resection and 42 normal tissue samples including the lung. SCLC is a lung cancer subtype with poor prognosis. Results provide insight into the molecular mechanisms underlying SCLC.

Organism: Homo sapiens

Type: Expression profiling by array, count, 2 disease state, 43 tissue sets

Platform: GPL570 Series: GSE43346 65 Samples

Download data: GEO (CEL)

DataSet Accession: GDS4794 ID: 4794

[PubMed](#) [Full text in PMC](#) [Similar studies](#) [GEO Profiles](#) [Analyze DataSet](#)

2. [Non-small lung cancer subtypes: adenocarcinoma and squamous cell carcinoma](#)

Comparison of two non-small cell lung cancer histological subtypes: adenocarcinomas (AC) and squamous cell carcinomas (SCC). Results provide insight into the molecular differences between AC and SCC.

Organism: Homo sapiens

Type: Expression profiling by array, transformed count, 2 disease state sets

Platform: GPL570 Series: GSE10245 58 Samples

Download data: GEO (CEL)

DataSet Accession: GDS3627 ID: 3627

[PubMed](#) [Similar studies](#) [GEO Profiles](#) [Analyze DataSet](#)

3. [Cigarette smoking effect on the nasal epithelium](#)

Analysis of nasal epithelia from cigarette smokers. Cigarette smoke creates a field of injury in

Send to:

Filters: Manage Filters

Top Organisms [Tree]

Homo sapiens (2259)

Mus musculus (29)

Rattus norvegicus (5)

Lymphocryptovirus (3)

Murid herpesvirus 4 (2)

More...

Find related data

Database: Select

Find items

Search details

("epithelial cells"[MeSH Terms] OR squamous cell[All Fields]) AND ("lung neoplasms"[MeSH Terms] OR lung cancer[All Fields])

Search

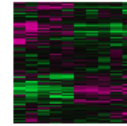
See more...

Recent activity

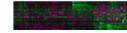
Several kinds of search results (DataSets, Platforms, Samples, and Series) are returned. Click on DataSets at the left to restrict the search to curated DataSets.

☐ [Squamous lung cancer](#)

13. Expression profiling of squamous lung cancer biopsy specimens and paired normal specimens from 5 patients. Differentially expressed genes integrated with protein interaction maps. Results suggest that differentially expressed genes are highly connected through protein interactions.
- Organism: Homo sapiens
- Type: Expression profiling by array, transformed count, 2 disease state, 5 individual sets
- Platform: GPL96 Series: GSE3268 10 Samples
- Download data: [GEO](#)
- DataSet Accession: GDS1312 ID: 1312
- [PubMed](#)
[Similar studies](#)
[GEO Profiles](#)
[Analyze DataSet](#)



☐ [Smoking-induced changes in airway transcriptome](#)



Scroll down and click on DataSet GDS1312. This is a set of lung tumor biopsies and paired normal samples from 5 individuals, and is a good example of a typical microarray experiment.

Search for

DataSet Record GDS1312: [Expression Profiles](#) [Data Analysis Tools](#) [Sample Subsets](#)

Title:	Squamous lung cancer		
Summary:	Expression profiling of squamous lung cancer biopsy specimens and paired normal specimens from 5 patients. Differentially expressed genes integrated with protein interaction maps. Results suggest that differentially expressed genes are highly connected through protein interactions.		
Organism:	<i>Homo sapiens</i>		
Platform:	GPL96: [HG-U133A] Affymetrix Human Genome U133A Array		
Citation:	Wachi S, Yoneda K, Wu R. Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. <i>Bioinformatics</i> 2005 Dec 1;21(23):4205-8. PMID: 16188928		
Reference Series:	GSE3268	Sample count:	10
Value type:	transformed count	Series published:	2005/09/09

Cluster Analysis

Download

- DataSet SOFT file
- Series family SOFT file
- Series family MINML file
- Annotation SOFT file

Data Analysis Tools

Find genes ☒

Compare 2 sets of samples

Cluster heatmaps

Experiment design and value distribution

Find gene name or symbol:

Find genes that are up/down for this condition(s): ☒ disease state ☒ individual

NLM NIH GEO Help Disclaimer Section 508

Find: **fire** ☐ Match case

In this lab, your goal is to find groups of genes that have significantly higher expression in squamous cell tumors than in normal lung tissue. One way to do this is with a t-test.

Click on the **Compare 2 sets of samples** tab at the bottom left.

Data Analysis Tools

Find genes

Compare 2 sets of samples ☒

Cluster heatmaps

Experiment design and value distribution

Step 1: Select test and significance level

One-tailed t-test (A < B) Significance level:

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

Step 3: Query Group A vs. B

Step 1: Choose the **One-tailed t-test (A < B)** and a **Significance level (p-value) of 0.001**.

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
	disease state	individual	
GSM73386	normal	patient 31	GSM73386
GSM73388		patient 33	GSM73388
GSM73390		patient 35	GSM73390
GSM73392		patient 36	GSM73392
GSM73394		patient 42	GSM73394
GSM73387	cancer	patient 31	GSM73387
GSM73389		patient 33	GSM73389
GSM73391		patient 35	GSM73391
GSM73393		patient 36	GSM73393
GSM73395		patient 42	GSM73395

Reference Series: GSE3268
Value type: transformed count
Sample count: 10
Series published: 2005/09/09

Find genes
Compare 2 sets of samples
Cluster heatmaps
Experiment design and value distribution

Step 1: Select test and significance level
One-tailed t-test (A < B) Significance level: 0.001

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

Step 3: Query Group A vs. B

NLM NIH GEO Help Disclaimer Section 508

Step 2: Click **Select which Samples to put in Group A and Group B**. Click "normal" and click the left arrows to assign that group to Group A. In the same way, assign "cancer" to Group B. Click **Ok**.

Step 3: Click **Query Group A vs. B** to perform the t-test. A new window will open with the results. *Question: How many genes were significantly increased in tumor vs normal at this significance level? There are 22,283 genes on this chip. About how many genes would you expect to find differentially expressed by chance alone at $p < 0.001$? How many would you expect to change in each direction (higher in tumor vs. higher in normal)?*

There are 327 genes upregulated in tumor vs normal at this significant level. 45 genes can be expected to differentially expressed by chance. Approximate 22 genes can be expected to be upregulated in tumor, and about 22 genes can be expected to be higher in normal cells.

Because this data set is small, and we want to find genes with strong changes in expression, we might benefit from using a more robust measure instead: ranked fold change. To run this test, first close the current window and return to the DataSet record for GDS1312.

Data Analysis Tools

Find genes
Compare 2 sets of samples
Cluster heatmaps
Experiment design and value distribution

Step 1: Select test and significance level
Rank means difference \bar{A} vs \bar{B} : 3+ fold lower

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

Step 3: Query Group A vs. B

Step 1: Choose **Rank means difference** and a fold-change cutoff of 3 (A < B).

Step 2: You already did this.

Display Settings: ☒ Summary, 500 per page, Sorted by Subgroup effect

Format	Items per page	Sort by
<input checked="" type="radio"/> Summary	<input type="radio"/> 5	<input type="radio"/> Default order
<input type="radio"/> Summary (text)	<input type="radio"/> 10	<input checked="" type="radio"/> Subgroup effect
<input type="radio"/> Unique Identifier	<input type="radio"/> 20	<input type="radio"/> Deviation
List	<input type="radio"/> 50	<input type="radio"/> Mean Value
	<input type="radio"/> 100	<input type="radio"/> Outliers
	<input type="radio"/> 200	
	<input checked="" type="radio"/> 500	

Apply

Step 3: Click **Query Group A vs. B**. A new window will open with the results. Change the **Sort By** dropdown box to read **Subgroup effect**. This will sort genes in order from greatest to least fold change.

Question: What is the symbol of the gene at the top of the list? Click on the link in the Annotation line to go to its Entrez Gene page. What's the gene's common name?

KRT6C keratin 6C

Go back to the page with the results of the ranked fold change analysis.

Display Settings: ☒ Summary, 500 per page, Sorted by Subgroup effect

Format	Items per page	Sort by
<input checked="" type="radio"/> Summary	<input type="radio"/> 5	<input type="radio"/> Default order
<input type="radio"/> Summary (text)	<input type="radio"/> 10	<input checked="" type="radio"/> Subgroup effect
<input type="radio"/> Unique Identifier	<input type="radio"/> 20	<input type="radio"/> Deviation
List	<input type="radio"/> 50	<input type="radio"/> Mean Value
	<input type="radio"/> 100	<input type="radio"/> Outliers
	<input type="radio"/> 200	
	<input checked="" type="radio"/> 500	

Apply

Under **Display Settings** Change the **Show** button to **500**, to show the first 500 of the genes that are at least 3-fold higher in squamous cell cancer than normal tissue.

Click on the **Download profile data** button to retrieve a data matrix for the entire list of genes.

Open the file with Microsoft Excel or OpenOffice Calc. Select the Affymetrix probeset identifiers (they should look something like 2xxxxx_at) in the first column (starting in row 5) and copy them to the clipboard (Ctrl-C or Edit->Copy).

Alternatively, try selecting the Affy IDs using UNIX!

Hint: you should use 'less', 'cut', and 'grep' commands, connected through pipes.

PROCEDURES, Part 2 (DAVID):

To analyze the functions of the genes in the list we just generated, we need some better tools. DAVID is a free public resource that allows rapid functional annotation of lists of genes. Open a new browser window to: <http://david.abcc.ncifcrf.gov/>

Click **Start Analysis** to begin.

Analysis Wizard

[Upload](#) **List** [Background](#)

Upload Gene List

[Demolist 1](#) [Demolist 2](#)

[Upload Help](#)

Step 1: Enter Gene List

A: Paste a list

Or

B: Choose From a File

Step 2: Select Identifier

Step 3: List Type

Gene List ☐

Background ☐

Step 4: Submit List

← Step 1. Submit your gene list through left panel.

new! Note: Affy Exon IDs and Affy Gene Array IDs are now supported in DAVID, as "affy_id" type

An example:

Copy/paste IDs to "box A" -> Select Identifier as "Affy_ID" -> List Type as "Gene List" -> Click "Submit" button

1007_s_at
1053_at
117_at
121_at
1255_g_at
1294_at
1316_at
1320_at
1405_i_at
1431_at
1438_at
1487_at
1494_f_at
1598_g_at

[Tell us how you like it](#)
[Contact us for more information](#)

To upload the list of Affymetrix probeset IDs you just copied from the file you downloaded from GEO, click in the **Paste a list** box under **Step 1** on the left side of the screen, and paste them in (Ctrl-V or right click-> Paste). Leave Step 2 as **AFFY_ID**. Select **Gene List** under **Step 3**, and click **Submit List**.

List Manager [Help](#)

Uploaded List_1

Select List to:

new!

The next window of the Wizard will appear. Rename the list to something more useful with the **Rename** button.

The next step is to run a clustering analysis that will group genes into sublists based on their shared functional annotation (GO terms, KEGG pathway terms, etc.)

Click on **Gene Functional Classification Tool**. A results page will appear with a set of clusters of genes.



To see which terms are enriched in each cluster, click on the red **Terms** symbol at the top right of each cluster.

Question: What do the functions of the top 5 or 6 clusters seem to be? Describe each in a few words using the Terms feature. Why do you think some of these clusters might be upregulated in tumor cells?

The functions of top 1,3, and 4 clusters seem to be related to activities of cell division, mitosis, the formation of cytoskeleton,etc. These functions are upregulated in tumor cells may because the tumor cells often reproduce very quickly. The top 2 cluster is related to tumor antigen, which might be a common feature in tumor cells. The top 5 cluster is related to protease inhibitor function, which may be related to inflammatory responses in tumors.