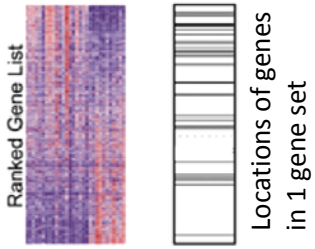
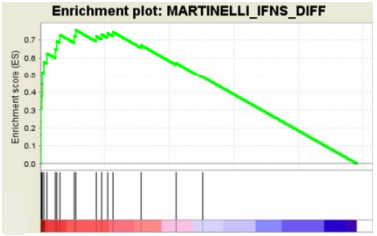
Center for Cancer Research

BTEP

A Bioinformatics Training and Education Program *Lecture*

http://bioinformatics.nci.nih.gov/training/





Introduction to the Broad Institute's Gene Set Enrichment Analysis (GSEA) software

Presented by Alan E. Berger, PhD

Lowe Family Genomics Core, School of Medicine, Johns Hopkins University

Tuesday April 16, 2013 2:00 – 3:30 pm Bldg. 37 Room 4041 - 4047

- Using gene sets, e.g., pathways, GO categories, to interpret microarray (and other) biology data
- Using a measure of differential expression for all the genes, rather than a list of distinguished genes
- The general approach of the Broad Institute's **GSEA** software // comparison with **DAVID** (NIAID)
- The statistics behind GSEA // The data files required to use GSEA
- Understanding the output files produced by GSEA (April 23: hands on running the GSEA software)

For more information contact: Dr. David Goldstein at 301-496-4357, goldsted@mail.nih.gov,
Dr. Peter FitzGerald at 301-402-3044, pcf@helix.nih.gov,
Dr. Maggie Cam at 301-443-2695, maggiec@mail.nih.gov

The content of this set of slides is derived from several NIH-CIT tutorials on GSEA given by Aiguo Li, Ph.D., NIH-NCI and Alan Berger, Ph.D., in 2007 & 2008

Outline

- **Functional Analysis of Microarray Data – Analysis at the Level of *Gene Sets***
  - **DAVID (NIAID) which is based on DEG lists**
  - ***Gene Set Enrichment Analysis* (Broad Institute)**

- **Hands-on (April 23)**
  - **Running GSEA: required input data files and formats & Parameter selection; Broad Institute Utilities**
  - **Understanding the GSEA outputs**

Background

- **Genome-wide expression profiling with microarrays has become an effective frequently used technique in molecular biology**

- **Interpreting the results to gain insights into biological mechanisms remains a major challenge**

- **For a typical two group comparison, e.g., tumor vs. normal, a standard approach has been to produce a list of differentially expressed genes (DEGs)**

Criteria for Differential Expression of a Gene

- **Statistically significant differential expression**
  - **by t-test, multi-way ANOVA, etc.**
  - **P-value cut-off: require, e.g., $p \leq 0.01$, but see FDR**

- **Satisfactory false discovery rate (FDR)**
  - **What fraction of the DEG list is false positives?**
  - **Benjamini-Hochberg procedure for estimating the FDR is a common choice (e.g., require FDR ≤ 0.1 or 0.2).**

- **Sufficient level of fold change (FC)**
  - **require $|FC| \geq 1.5$ or 2**
    (common convention: groups A, B, gene g with ave. expression levels μ_A, μ_B; $FC \equiv \mu_A / \mu_B$ when $\mu_A \geq \mu_B$; $FC \equiv -\mu_B / \mu_A$ when $\mu_B \geq \mu_A$)

DEG lists II

- **Large fraction of "Present" calls for the expression values in the group with the higher average expression level for that probe**
  - 80% but require 3 out of 3 when group size = 3
  - If this is not satisfied for a given probe, do not do any statistical testing on it.
  - This avoids false positives based on noise, and also reduces the number of comparisons N used in calculating the FDR.

- **Specific criteria and cutoffs depend on user preference and the biological situation (e.g., would like "reasonable amount" of mRNA and $|FC| \geq 2$ for qRT-PCR verification)**

Challenges in Interpreting Gene Microarray Data

- **Even with DEG list(s) of up and/or down-regulated genes, still need to accurately extract valid biological inferences. Cutoff for inclusion in DEG lists is somewhat arbitrary.**
- **May obtain a long list of statistically significant genes without any obvious unifying biological theme**
- **May have few individual genes meeting the threshold for statistical significance**
- **Lists of statistically significant genes from two studies of the same biological system may show limited overlap depending on the analysis methods and the criteria for significance**

Enrichment of Gene Categories in a Gene List

- **Statistical procedures such as Fisher's exact test based on the hypergeometric distribution are used to test if members of a list of differentially expressed genes are overrepresented in given GO categories or in predefined gene sets compared with the distribution of the whole set of genes represented on the chip.**

- **Tools developed along this line include:**
  - **DAVID**
  - **GoMinor**
  - **GenMAPP**
  - **Onto-Express**
  - **GOstat**

Fisher Exact Viewpoint:
2 X 2 Contingency Table

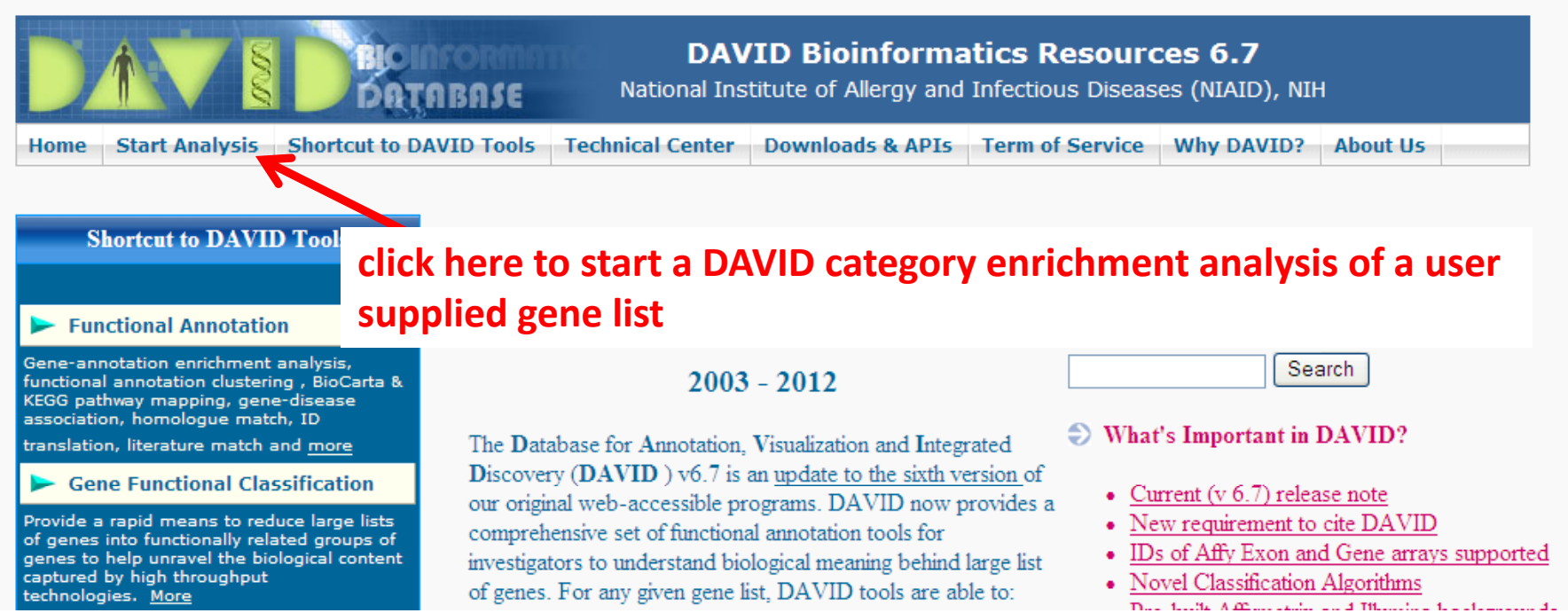
	in pathway S	not in pathway S	
in DEG list	*5*	45	50
not in DEG list	95	9855	9950
Totals	100	9900	10000

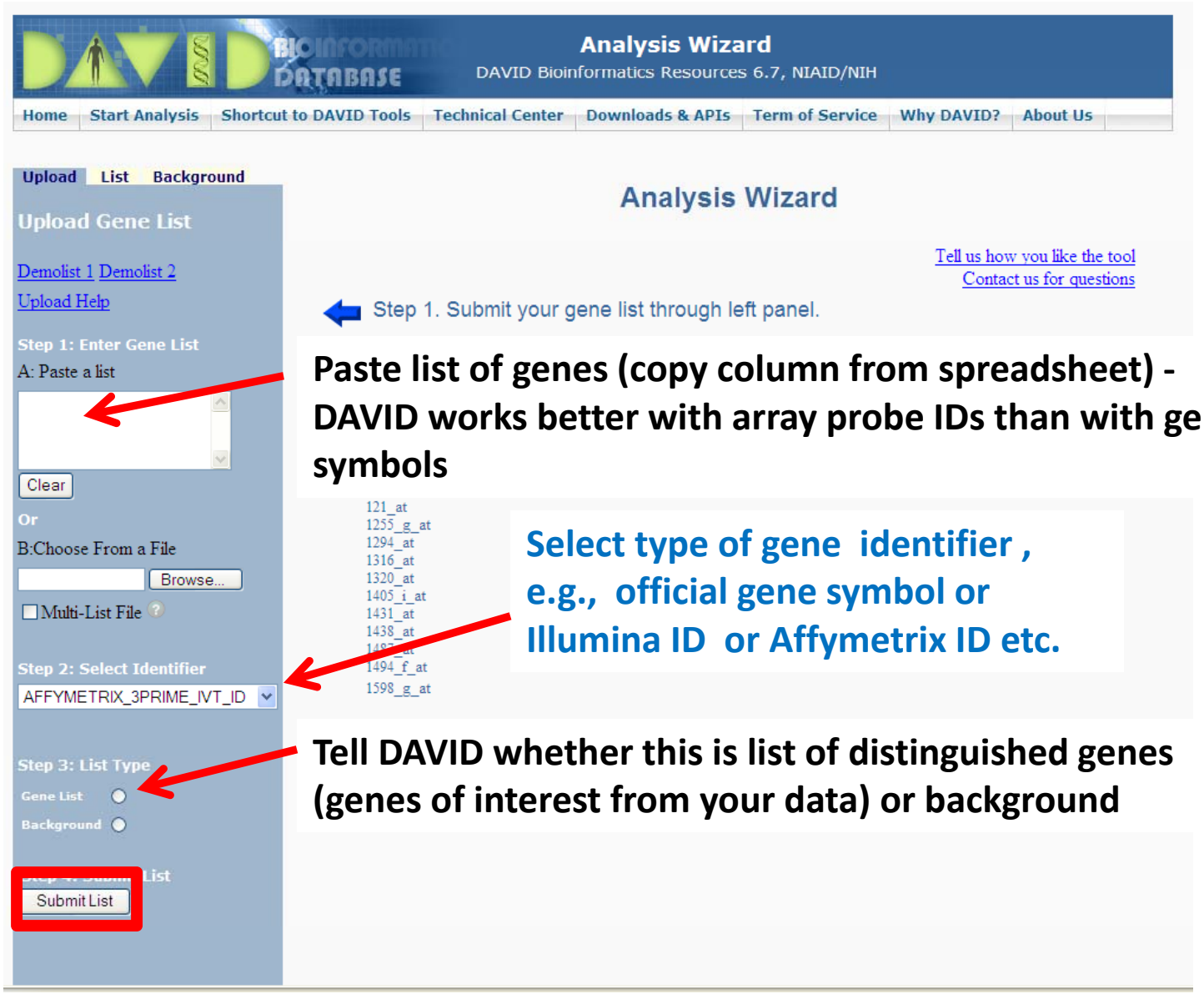
One way to view this is think of there being 10,000 candies (genes) in a bin (array), 100 of which are Ghirardelli chocolates (in the pathway S), and being given a random batch of 50 candies from the bin (a random DEG list). If you got 5 or more of the chocolates, were you unusually lucky? ***Indeed yes!***

Answer: P = 0.000134

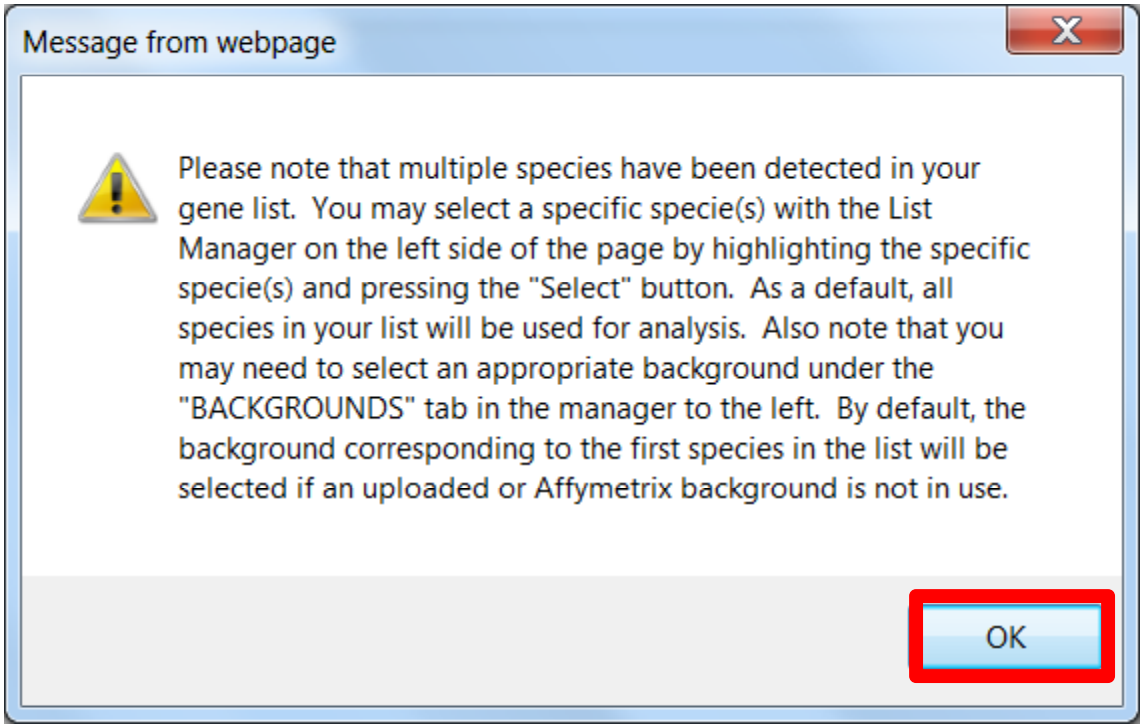
DAVID: Database for Annotation, Visualization and Integrated Discovery

http://david.abcc.ncifcrf.gov/

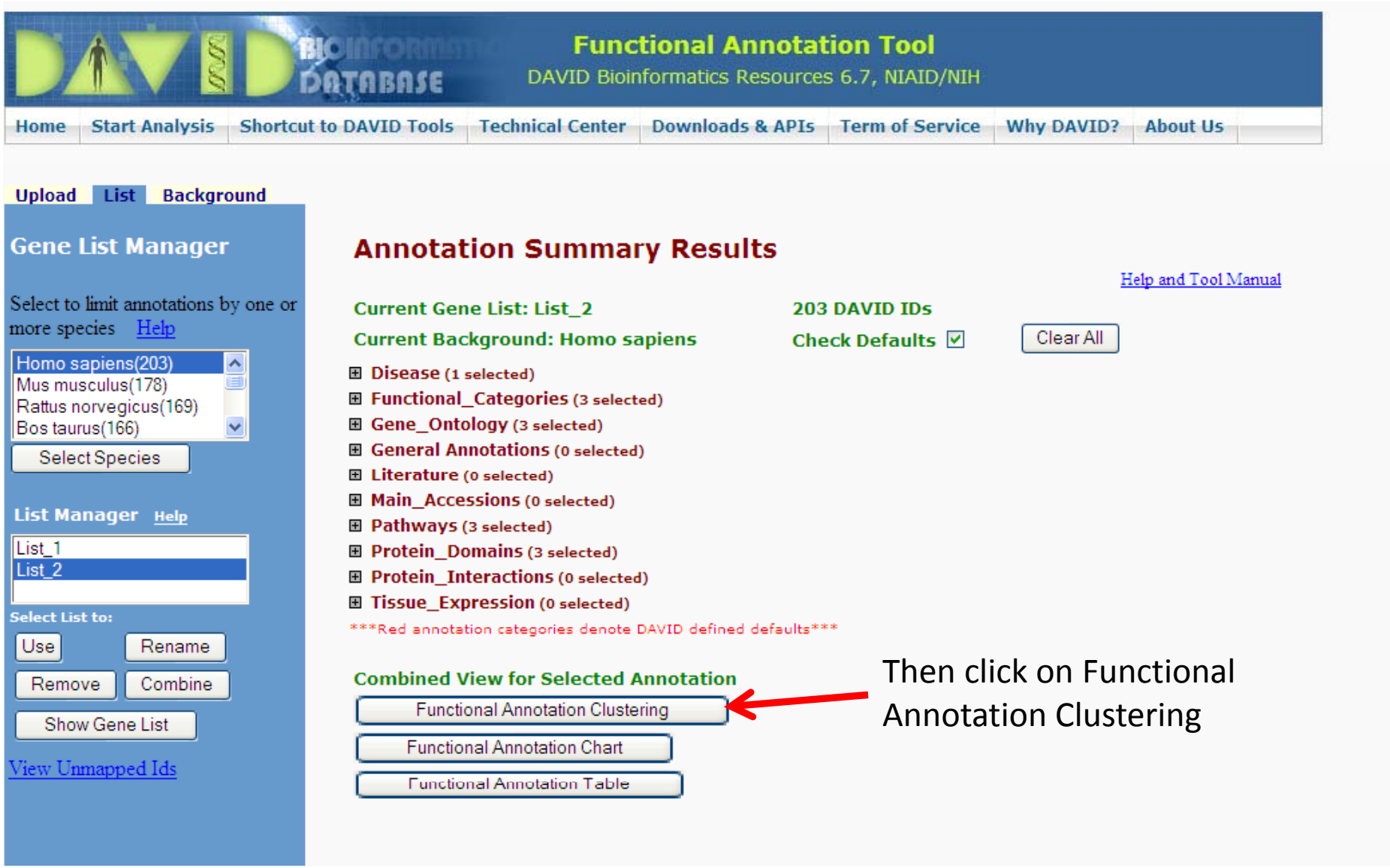



Analysis Wizard
DAVID Bioinformatics Resources 6.7, NIAID/NIH
Home
Start Analysis
Shortcut to DAVID Tools
Technical Center
Downloads & APIs
Term of Service
Why DAVID?
About Us
Upload
List
Background
Upload Gene List
Demolist 1 Demolist 2
Upload Help
Step 1: Enter Gene List
A: Paste a list
Clear
Or
B:Choose From a File
Browse...
Multi-List File
Step 2: Select Identifier
AFFYMETRIX_3PRIME_IVT_ID
Step 3: List Type
Gene List
Background
Submit List
Analysis Wizard
Tell us how you like the tool
Contact us for questions
Step 1. Submit your gene list through left panel.
121_at
1255_g_at
1294_at
1316_at
1320_at
1405_i_at
1431_at
1438_at
1494_f_at
1598_g_at
Paste list of genes (copy column from spreadsheet) - DAVID works better with array probe IDs than with gene symbols
Select type of gene identifier , e.g., official gene symbol or Illumina ID or Affymetrix ID etc.
Tell DAVID whether this is list of distinguished genes (genes of interest from your data) or background

If you gave DAVID gene symbols rather than array Probe IDs then will usually see a box indicating multiple corresponding species, click on OK




Analysis Wizard
DAVID Bioinformatics Resources 6.7, NIAID/NIH
Home
Start Analysis
Shortcut to DAVID Tools
Technical Center
Downloads & APIs
Term of Service
Why DAVID?
About Us
Upload
List
Background
Gene List Manager
Select to limit annotations by one or more species Help
Homo sapiens(203)
Mus musculus(178)
Rattus norvegicus(169)
Bos taurus(166)
Select Species
List Manager Help
List_1
List_2
Select List to:
Use
Rename
Remove
Combine
Show Gene List
View Unmapped Ids
Analysis Wizard
Click on organism
Tell us how you like the tool
Contact us for questions
Step 1. Successfully submitted gene list
Current Gene List: List_2
Current Background: Homo sapiens
Step 2. Analyze above gene list with one of DAVID tools
Which DAVID tools to use?
Functional Annotation Tool
Functional Annotation Clustering
Functional Annotation Chart
Functional Annotation Table
Gene Functional Classification Tool
Gene ID Conversion Tool
Gene Name Batch Viewer
Then click on Functional Annotation Tool



Then click on Functional Annotation Clustering

DAVID Bioinformatics Resources 6.7
National Institute of Allergy and Infectious Diseases (NIAID), NIH

Functional Annotation Clustering

Help and Manual

Current Gene List: List_1
Current Background: Homo sapiens
352 DAVID IDs

Options Classification Stringency Medium

Rerun using options | Create Sublist

140 Cluster(s)

Download File

Annotation Cluster 1		Enrichment Score: 3.35	G		Count	P_Value	Benjamini
	SP_PIR_KEYWORDS	oxidoreductase	RT		26	2.8E-5	5.3E-3
	GOTERM_BP_FAT	oxidation reduction	RT		26	4.5E-4	2.3E-1
	SP_PIR_KEYWORDS	nadp	RT		9	7.0E-3	2.1E-1
Annotation Cluster 2		Enrichment Score: 2.6	G		Count	P_Value	Benjamini
	SP_PIR_KEYWORDS	lipid synthesis	RT		12	1.4E-6	5.2E-4
	SP_PIR_KEYWORDS	Steroid biosynthesis	RT		7	7.1E-5	8.8E-3
	GOTERM_BP_FAT	lipid biosynthetic process	RT		18	1.4E-4	2.2E-1
	GOTERM_BP_FAT	sterol biosynthetic process	RT		6	4.7E-4	1.9E-1
	SP_PIR_KEYWORDS	sterol biosynthesis	RT		5	9.3E-4	6.8E-2
	SP_PIR_KEYWORDS	Cholesterol biosynthesis	RT		4	4.4E-3	1.7E-1
	GOTERM_BP_FAT	steroid biosynthetic process	RT		7	5.3E-3	5.5E-1
	GOTERM_BP_FAT	cholesterol biosynthetic process	RT		4	1.2E-2	7.6E-1
	GOTERM_BP_FAT	sterol metabolic process	RT		6	4.2E-2	8.8E-1
	KEGG_PATHWAY	Steroid biosynthesis	RT		3	5.4E-2	6.0E-1
	GOTERM_BP_FAT	cholesterol metabolic process	RT		5	9.6E-2	9.2E-1
	GOTERM_BP_FAT	steroid metabolic process	RT		7	1.8E-1	9.6E-1
Annotation Cluster 3		Enrichment Score: 2.4	G		Count	P_Value	Benjamini
	SP_PIR_KEYWORDS	lipid synthesis	RT		12	1.4E-6	5.2E-4
	SP_PIR_KEYWORDS	Fatty acid biosynthesis	RT		6	9.7E-4	5.9E-2
	SP_PIR_KEYWORDS	multifunctional enzyme	RT		7	1.5E-3	7.0E-2
	KEGG_PATHWAY	Fatty acid biosynthesis	RT		3	6.9E-3	2.4E-1

False discovery rate (FDR) column

For DAVID results, want FDR < 0.1 before consider category as possibly significant (and preferably below 0.01)

Sample Gene List for DAVID

Experiment:

Wegener's granulomatosis (WG) vs. normal controls (C)
n = 41 patients, 23 controls

***Genelist* (84 distinct genes):**

FC ≥ 1.5 (up in WG)
p-value ≤ 0.01
FDR ≤ 0.1

DAVID Output

DAVID Bioinformatics Resources 6.7
National Institute of Allergy and Infectious Diseases (NIAID), NIH

Functional Annotation Clustering

Help and Manua

Current Gene List: List_1
Current Background: Homo sapiens
71 DAVID IDs

Options Classification Stringency Medium

Rerun using options Create Sublist

30 Cluster(s) Download File

Annotation Cluster 1	Enrichment Score: 10.42		Count	P_Value	Benjamin
SP_PIR_KEYWORDS	Antimicrobial	RT	11	8.1E-14	8.4E-12
SP_PIR_KEYWORDS	antibiotic	RT	10	2.7E-12	1.9E-10
GOTERM_BP_FAT	defense response	RT	20	5.5E-11	4.0E-8
GOTERM_BP_FAT	response to bacterium	RT	12	2.0E-9	7.4E-7
GOTERM_BP_FAT	defense response to bacterium	RT	10	3.3E-9	8.0E-7
Annotation Cluster 2	**Enrichment Score: 9.21**		**Count**	**P_Value**	**Benjamin**
SP_PIR_KEYWORDS	disulfide bond	RT	41	7.4E-16	1.6E-13
UP_SEQ_FEATURE	disulfide bond	RT	40	1.9E-15	4.8E-13
SP_PIR_KEYWORDS	signal	RT	38	6.5E-12	3.4E-10
UP_SEQ_FEATURE	signal peptide	RT	38	7.8E-12	1.0E-9
SP_PIR_KEYWORDS	glycoprotein	RT	39	6.9E-9	2.9E-7
SP_PIR_KEYWORDS	Secreted	RT	24	1.3E-8	4.7E-7

Limitations with Category Enrichment Methods[1]

- **No further use made of information contained in expression values for the genes not in the submitted list**
- **The level of differential expression of the genes in the "distinguished" gene list is not taken into consideration: only counts the number of the "distinguished" genes that are contained in each category being considered**
- **The correlation structure of the expression data is not accounted for at all**

[1] ***Discovering statistically significant pathways in expression profiling studies*, Tian et al., PNAS 2005, 102:13544**

Gene Set Enrichment Methods

- These methods formulate a statistic reflecting the difference in expression level between the two phenotypes under consideration for the ensemble of genes in each gene set being considered
- The levels of differential expression for all the genes in the chip are utilized
- Can be applied to gene sets from, e.g., pathways in BioCarta & KEGG; genes co-located in cytobands; genes having common transcription factor motifs; genes changed in response to some disease state or experimental condition
- But note: results depend on the collection of gene sets examined, and still must address multiple testing error control (though much less severe than for all probes on a large array)

Some References for Gene Set Methods

1. Gene Set Enrichment Analysis (GSEA): Subramanian et al., *A knowledge-based approach for interpreting genome-wide expression profiles*, PNAS 2005, 102:15545; note Lamb et al., The Connectivity Map…, Science 2006, 313:1929. (see Broad Institute web page for this and other software)

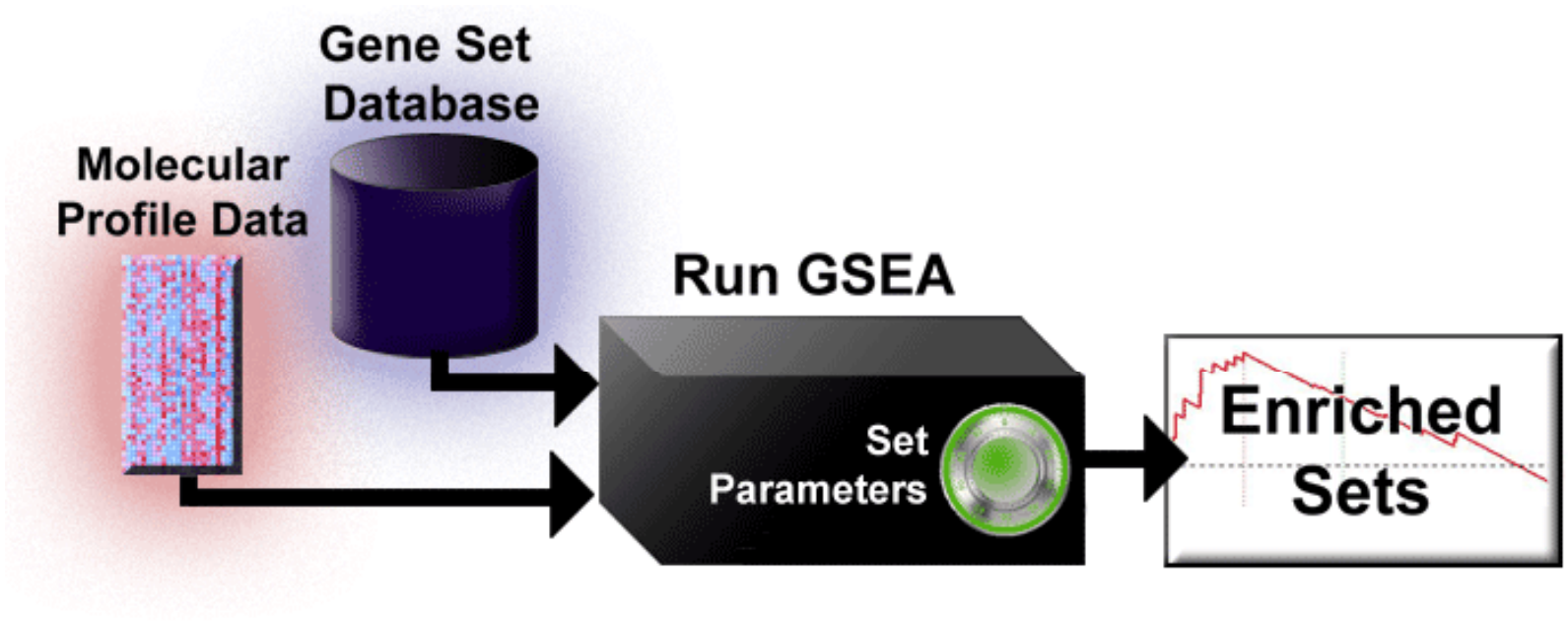
2. PAGE: Parametric Analysis of Gene-Set Enrichment: Kim and Volsky, BMC Bioinformatics 2005, 6:144 (uses the average of the measure of differential expression (DE) of genes in a gene set, and values of DE over the chip to get a *gene set z-score*).

3. Systematic consideration of the issues in formulating and evaluating gene set methods: Ackermann and Strimmer, *A general modular framework for gene set enrichment analysis*, BMC Bioinformatics 2009, 10:47

4. Systematic consideration of the issues in formulating and evaluating gene set methods: Varemo, Nielsen and Nookaew, *Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods*, Nucleic Acids Research 2013 Mar 22 [Epub ahead of print]

GSEA Overview -- Workflow

GSEA is a computational method that determines whether an *a priori* defined set of genes shows statistically significant, concordant differences between two biological states (e.g. phenotypes).



text and figure from the Broad Institute web pages for GSEA : http://www.broad.mit.edu/gsea/index.html
the current version of the figure at the Broad site is slightly different from the one above

Broad Institute GSEA Documentation Main Page

Main Page - GeneSetEnrichmentAnalysisWiki - Windows Internet Explorer

File Edit View Favorites Tools Help

http://www.broadinstitute.org/cancer/software/gsea/wiki/index.php/Main_Page

Favorites | Interactive Statistical Calc... | Johns Hopkins Departmen... | http--www.weathertap | Windows | Windows Media | Windows Marketplace | Free Hotmail | Customize Links

Main Page - Gen... | WeatherTAP.com - ...

Page | Safety | Tools

Log in

GSEA
Gene Set Enrichment Analysis

navigation
- Documentation Home
- Tutorial
- User Guide
- Data Formats
- FAQ
- Known Issues
- R-GSEA Readme
- Algorithm
- PNAS 2005 Examples
- Papers that use GSEA

msigdb
- Release Notes
- Summary Stats
- Guide to a GeneSetCard
- MSigDB XML description
- MSigDB License
- MSigDB Acknowledgements

software
- Release Notes
- Software License
- Software Acknowledgements
- For Broad Users

internal only
- Wiki How-to-Use

Main Page

GSEA Home | Downloads | Molecular Signatures Database | Documentation | Contact

Use the navigation bar on the left to display documentation on GSEA software, MSigDB database or GSEA/MSigDB web site. If you have comments or questions not answered by the FAQ or the User Guide, contact us: gsea@broadinstitute.org.

When contacting our team with questions about java GSEA programs, please send the following information:
- your computer's operation system
- version of java which you used to run GSEA
- detailed log transcript from the GSEA session in question

to view the log, click [+] at the bottom of main screen of GSEA java desktop application, copy the text to a separate file and attach it to your request

Where to start [edit]

If you are new to GSEA, see the Tutorial for a brief overview of the software. If you have a question, see the FAQ or the User Guide. The User Guide describes how to prepare data files, load data files, run the gene set enrichment analysis, and interpret the results. It also includes instructions for running GSEA from the command line and a Quick Reference section, which describes each window of the GSEA desktop application.

MSigDB gene sets [edit]

Current release of the Molecular Signatures Database (v3.1 MSigDB) contains 8,513 gene sets for use with GSEA. The best source of information about the gene sets is the MSigDB web site.

Please note that gene sets can change their names or become deprecated in subsequent releases of MSigDB. It is thus important to indicate version of MSigDB to fully reference gene sets used in your study.

For further details about gene set name or status changes, please check here:
- between **v3.1 (current)** and v3.0
- between v3.0 and v2.5

Software [edit]

We provide the following software implementations of the GSEA method:
- Java desktop application -- Easy-to-use graphical interface that can be run from the Downloads page. The User Guide fully describes this application (referred to as GSEA or GSEA-P).

Internet | Protected Mode: On | 125%

Three Main Components in GSEA

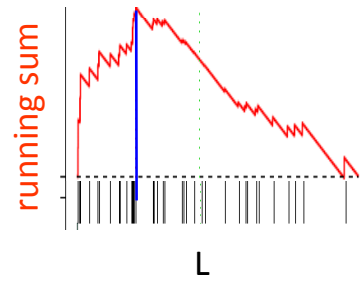
- Algorithm
- Software implementation (Broad Institute)
- Database of gene sets:
  - Molecular signature database (MSigDB at Broad Institute) containing gene sets of interest
  - Utilities mapping chip features to genes (e.g., Illumina or Affymetrix probe set IDs to HUGO gene symbols)

GSEA: Gene Set Enrichment Analysis

Get ranked list L of all the genes on the chip based on a chosen measure of the difference of their expression levels between the phenotypes A & B under study, e.g., tumor vs. normal

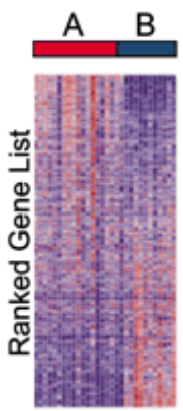
For each gene set S: find the location of each gene s in S within L

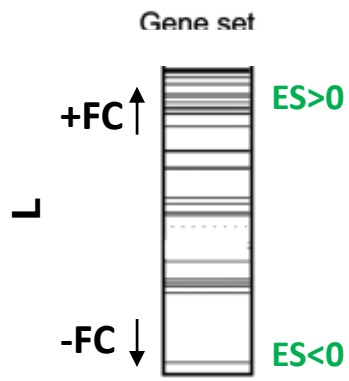
Generate enrichment score ES for S based on running-sum statistic: "reward" presence of s toward top or bottom of L



Analyze significance of this Kolmogorov-Smirnov type statistic by permutation testing

Multiple hypothesis testing (MHT) error control for multiple S's using the false discovery rate (FDR)





bands are locations in L of genes from S

Enrichment Score (ES) Calculation

Start with ranked list (L) of genes that are in (Hit) or not in (Miss) a gene set (S)

Ranked List (L)	FC		Contribution to running sum for ES	Hits $+\lvert FC\rvert / \Sigma$	Misses $-1/(N-N_H)$	Running sum for ES
	15	Hit	+0.15	+0.15		0.15
	12	Hit	+0.12	+0.12		0.27
	10	Miss	-0.001		-0.001	0.269
	9	Hit	+0.09	+0.09		0.359
	8	Hit	+0.08	+0.08		0.439
	6	Miss	-0.001		-0.001	0.438
…	…	…				…

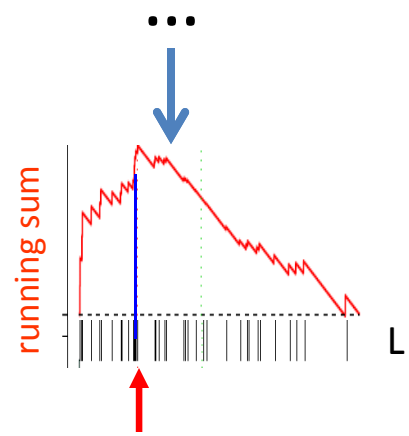
Hits: Genes $\in$ S $\quad +\lvert FC\rvert / \Sigma$

Misses: Genes $\notin$ S $\quad -1/(N-N_H)$

Σ = sum of fold changes for genes in gene set (S) (e.g., 100)

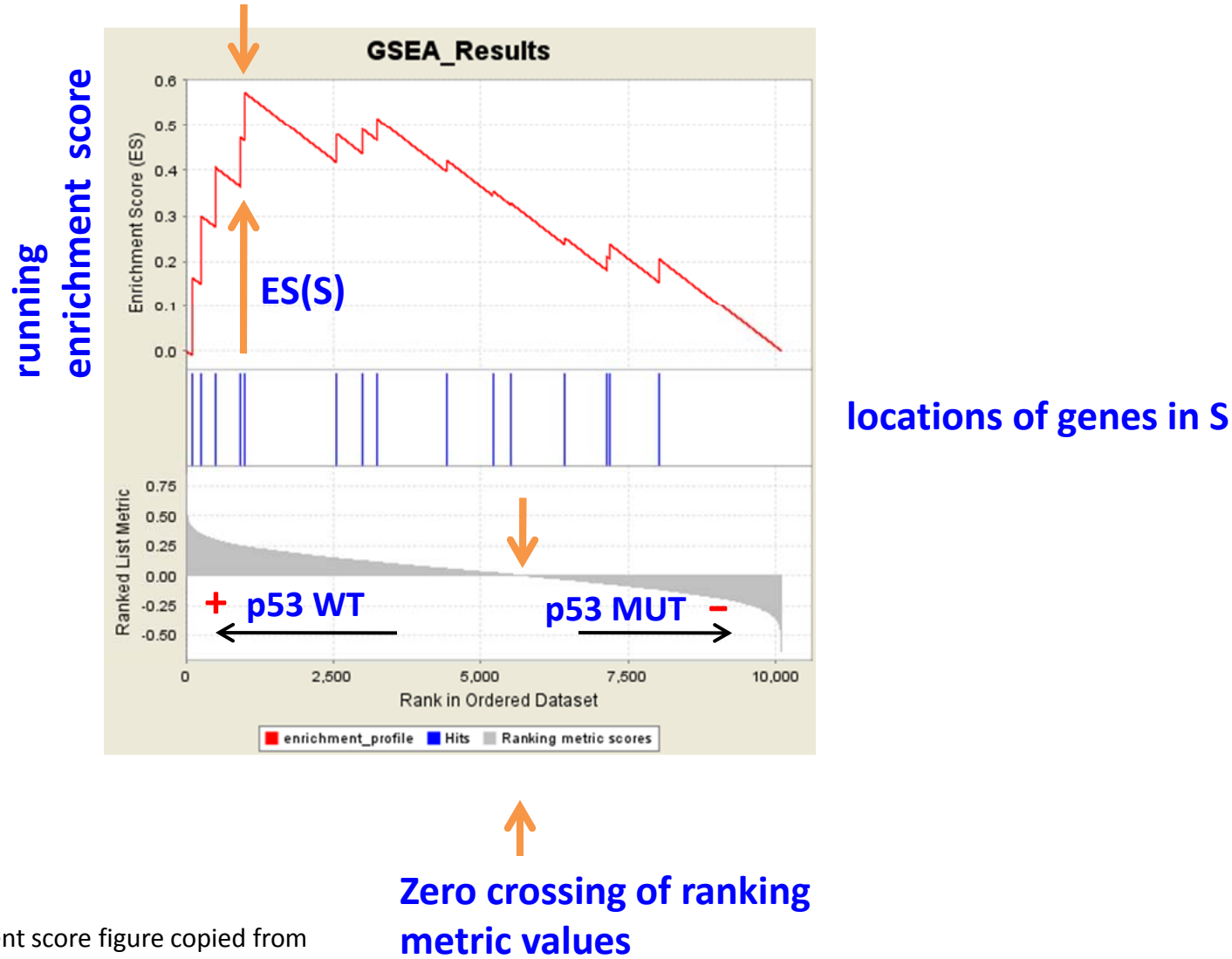
N = no. of genes in the array (e.g., 1020)

NH = no. of genes in the gene set (S) (e.g., 20)



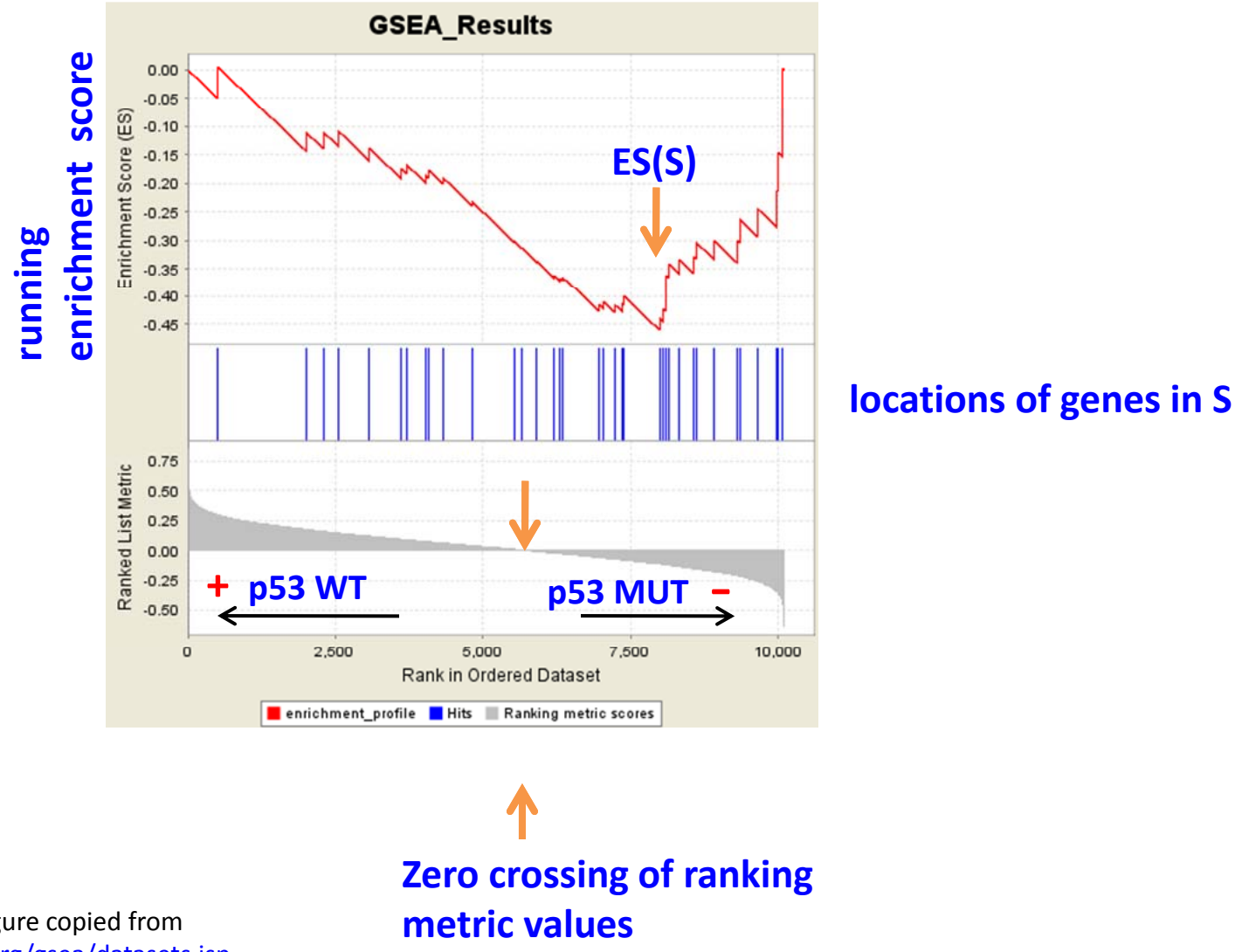
ES(S) ≡ value of max deviation from 0 (*extr*) of the running sum

The running enrichment score for a positive ES gene set from the P53 GSEA example data set



underlying running enrichment score figure copied from
http://www.broadinstitute.org/gsea/datasets.jsp
p53 dataset (gene set is lairPathway)

The running enrichment score for a negative ES gene set from the P53 GSEA example data set



running enrichment score figure copied from
http://www.broadinstitute.org/gsea/datasets.jsp
p53 dataset (gene set is BRCA_UP)

GSEA Algorithm: Definition of Enrichment Scores

the equations

w_j = **measure of differential expression of gene *j* between group A and group B**

1. **Order the genes in a ranked list L so w_j decreases from the top (*j*=1) to the bottom (*j*=*N*) of the list**

2. **Account for the locations of the genes in *S* ("hits") weighted by w_j and the locations of genes not in *S* ("misses") from the top of the list down to a given position *i* in *L***

$$K_{hit}(S,i) = \sum_{\substack{g_j \in S \\ j \le i}} \frac{|w_j|^t}{N_R} \quad \text{where} \quad N_R = \sum_{g_j \in S} |w_j|^t$$

for GSEA the default is *t* = 1, for Kolmogorov-Smirnov *t* = 0

$$K_{miss}(S,i) = \sum_{\substack{g_j \notin S \\ j \le i}} \frac{1}{(N - N_H)}$$

N_H = # genes in S

N = # genes in chip

3. **Calculate maximum deviation from zero of K_{hit} - K_{miss}**

$$ES(S,i) = K_{hit}(S,i) - K_{miss}(S,i)$$

$ES(S) = \text{extr}_i(ES(S,i))$ (greatest excursion of the ES(S,i) from 0)

Testing the Significance of ES: permute π times

gene expression matrix, sample labels indicate phenotype

gene \ sample	T1	T2	T3	T4	N1	N2	N3	N4
CASP4	7.82	7.87	8.15	7.81	7.96	7.92	7.90	7.96
BAX	8.01	7.85	7.82	7.95	8.05	7.91	7.78	7.96
CASP8	7.73	7.82	7.92	8.13	8.18	8.01	7.90	7.86
CD40	8.12	8.15	8.32	8.21	8.06	8.02	8.00	8.08
BIRC3	7.87	8.01	7.99	7.84	7.99	7.89	8.01	7.96
GADD45A	7.84	7.77	7.99	7.94	7.93	7.99	7.75	7.69
BIRC2	8.07	8.01	7.88	8.01	7.94	7.86	8.06	7.92
ATM	9.40	9.54	9.32	9.60	9.11	9.45	9.42	9.34
...	...	...	...	...	...	...	...	...

compute the differential expression value for each gene (DE(g)), and then the ES(S) values for all the gene sets

do ≈ 1000 sample label permutations* - for each permutation **leave the rest of the expression matrix fixed, and recalculate {DE(g)} and then the enrichment score for each S**

permutation number

1	T4	N3	N4	T3	T1	T2	N1	N2
2	T3	N2	T1	N3	T4	N4	T2	N1
3	N4	T4	N1	N3	T3	T2	N2	T1
4	N2	T4	N3	T1	T2	N1	T3	N4
...	...	...	...	...	...	...	...	...

$\{ES(S,\pi_1)\}$
$\{ES(S,\pi_2)\}$
$\{ES(S,\pi_3)\}$
$\{ES(S,\pi_4)\}$

*actually want at least 7 samples in each group for sample label permutation, else do gene permutation

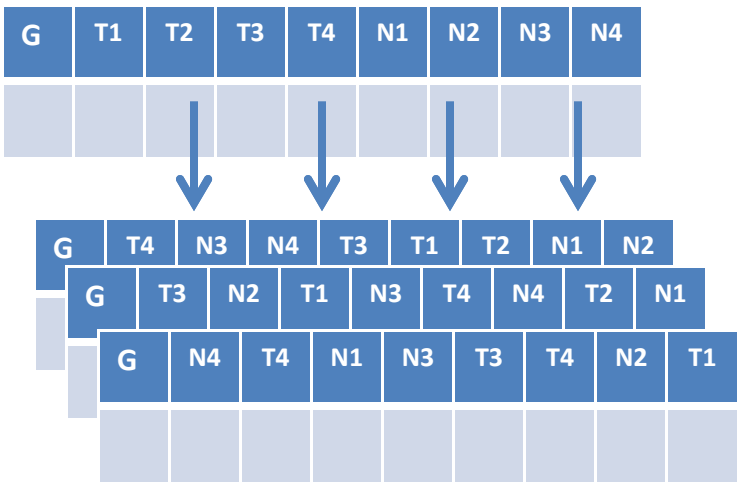
If insufficient number of samples for sample label permutation, do gene permutation

For a given gene set S (count of genes in S = s), each permutation π is now the random selection of s genes from the array forming a random gene set S_π

the corresponding permuted enrichment score $ES(S, \pi) = ES(S_\pi)$ calculated from the original expression matrix

Testing the Significance of ES

Significance of the observed ES(S) is compared with the set of scores $ES_{NULL}(S)$ computed with the *randomly assigned phenotypes*.



:

1000 x

ES(S)

$ES(S,\pi_1)$

$ES(S,\pi_2)$

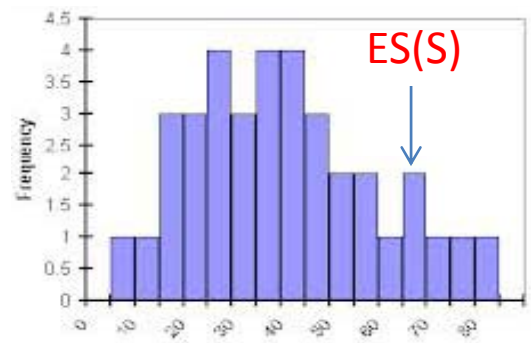
$ES(S,\pi_3)$

:

$ES(S,\pi_{1000})$

$ES_{NULL}(S)$:null distribution for ES(S)

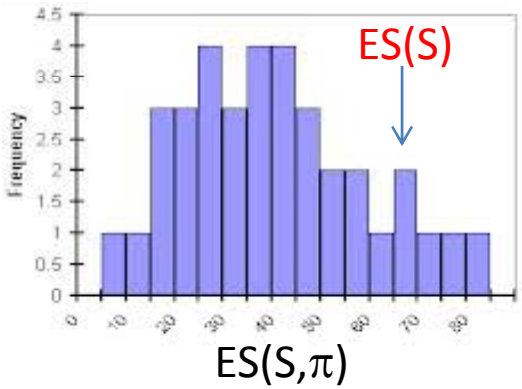
Histogram of 1000 $ES_{NULL}(S)$ Scores



The empirical, nominal p value for each ES(S) is then calculated relative to the null distribution for ES(S).

How normalized enrichment scores (NES) are calculated from ES

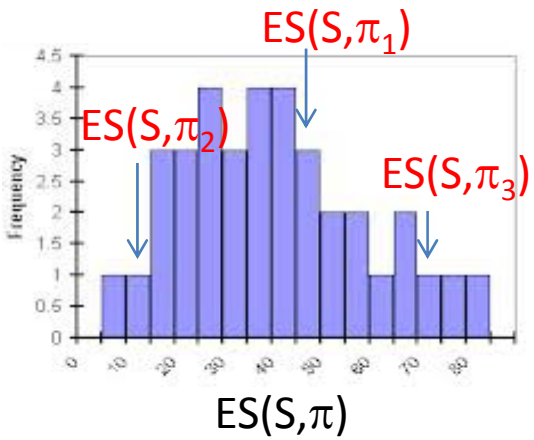
Histogram of the ES(S,π) values for a given S from the permutations



$$\mathbf{NES(S)} \equiv \frac{\textbf{original ES(S)}}{\textbf{mean}_{\pi}\{\textbf{ES(S, }\pi\textbf{) values with the same sign as ES(S)}\}}$$

ES_{NULL}: null distribution for the ES

For each permutation π and gene set S, compute NES(S, π) to use in computing the FDR:



$$\mathbf{NES(S, \pi_k)} \equiv \frac{\textbf{ES(S, }\pi_k\textbf{)}}{\textbf{mean}_{\pi}\{\textbf{ES(S, }\pi\textbf{) values with the same sign as ES(S, }\pi_k\textbf{)}\}}$$

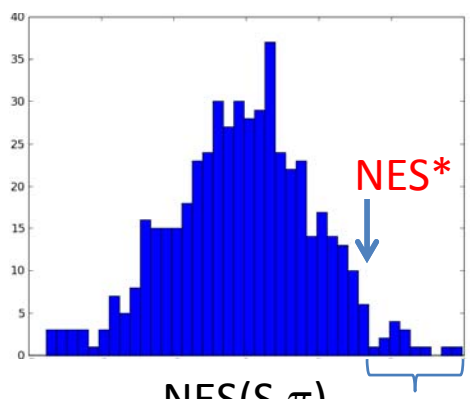
False Discovery Rate (FDR) q value

Create a histogram of all NES(S, π), over all S and π.
Use this null distribution to compute an FDR q value, for a given NES(S) = NES* > 0.

FDR value for S: D(S) = {gene sets with NES ≥ NES* }

$$\frac{\text{estimate of \# of false positives in D(S)} = \mathbf{F} * \mathbf{N_S^+}}{\text{size of D(S) = \# of S with NES(S) ≥ NES*}}$$

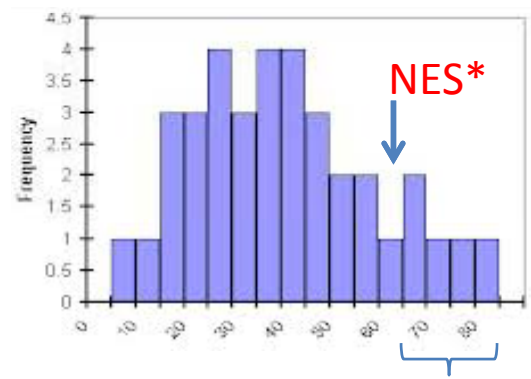
Histogram of **NES(S, π)** Scores



*NES(S, π) ≥ NES**

F ≡ fraction of the positive *NES(S, π) ≥ NES**

Histogram of NES(S) Scores



*NES(S) ≥ NES**

$\mathbf{N_S^+}$ ≡ # of gene sets with NES(S) > 0

similarly for NES(S) < 0

Outline of the Hands-on GSEA Class (April 23)

- **Running GSEA:**
  - **required input data files and formats & Parameter selection**
  - **Broad Institute Utilities**
- **Understanding the GSEA outputs**
- **Live demonstration / hands on running the Desktop GSEA software**

Example GSEA output

Dataset:

Wegener's granulomatosis (WG) vs. normal controls (C)
41 patients, 23 controls

GSEA output: gene sets upregulated in WG

	GS follow link to MSigDB	GS DETAILS	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX
1	HUMAN_TISSUE_LIVER	Details ...	33	0.86	2.19	0.000	0.000	0.000	1547
2	VERHAAK_AML_NPM1_MUT_VS_WT_UP	Details ...	172	0.67	2.13	0.000	0.000	0.000	3599
3	HSIAO_LIVER_SPECIFIC_GENES	Details ...	244	0.64	2.08	0.000	0.000	0.000	3509
4	HSA01032_GLYCAN_STRUCTURES_DEGRADATION	Details ...	29	0.84	2.04	0.000	0.002	0.005	2311
5	HSA04610_COMPLEMENT_AND_COAGULATION_CASCADES	Details ...	68	0.74	2.02	0.000	0.001	0.005	1772
6	APPEL_IMATINIB_UP	Details ...	31	0.84	2.01	0.000	0.002	0.010	1894
7	INTRINSICPATHWAY	Details ...	22	0.84	1.94	0.000	0.011	0.055	1772
8	TSA_HEPATOMA_CANCER_UP	Details ...	38	0.76	1.94	0.000	0.011	0.060	2861
9	MATRIX_METALLOPROTEINASES	Details ...	30	0.80	1.93	0.000	0.010	0.060	2220
10	MARTINELLI_IFNS_DIFF	Details ...	22	0.85	1.93	0.000	0.009	0.060	2656
11	LIAN_MYELOID_DIFF_GRANULE	Details ...	23	0.82	1.91	0.000	0.009	0.065	623
12	UVC_TTD_ALL_UP	Details ...	76	0.64	1.90	0.000	0.010	0.085	3351
13	LEE_TCELLS3_UP	Details ...	95	0.63	1.89	0.000	0.011	0.100	3056

DAVID Output for WG dataset

DAVID Bioinformatics Resources 6.7
National Institute of Allergy and Infectious Diseases (NIAID), NIH

Functional Annotation Clustering

Help and Manua

Current Gene List: List_1
Current Background: Homo sapiens
71 DAVID IDs

Options **Classification Stringency** Medium

Rerun using options | Create Sublist

30 Cluster(s) Download File

Annotation Cluster 1	Enrichment Score: 10.42			Count	P_Value	Benjamin
SP_PIR_KEYWORDS	Antimicrobial	RT		11	8.1E-14	8.4E-12
SP_PIR_KEYWORDS	antibiotic	RT		10	2.7E-12	1.9E-10
GOTERM_BP_FAT	defense response	RT		20	5.5E-11	4.0E-8
GOTERM_BP_FAT	response to bacterium	RT		12	2.0E-9	7.4E-7
GOTERM_BP_FAT	defense response to bacterium	RT		10	3.3E-9	8.0E-7
Annotation Cluster 2	**Enrichment Score: 9.21**			**Count**	**P_Value**	**Benjamin**
SP_PIR_KEYWORDS	disulfide bond	RT		41	7.4E-16	1.6E-13
UP_SEQ_FEATURE	disulfide bond	RT		40	1.9E-15	4.8E-13
SP_PIR_KEYWORDS	signal	RT		38	6.5E-12	3.4E-10
UP_SEQ_FEATURE	signal peptide	RT		38	7.8E-12	1.0E-9
SP_PIR_KEYWORDS	glycoprotein	RT		39	6.9E-9	2.9E-7
SP_PIR_KEYWORDS	Secreted	RT		24	1.3E-8	4.7E-7

Gene set associated with immature neutrophils

THE JOURNAL OF BIOLOGICAL CHEMISTRY
Vol. 279, No. 42, Issue of October 15, pp. 44123–44132, 2004
Printed in U.S.A.

Induction of Genes Mediating Interferon-dependent Extracellular Trap Formation during Neutrophil Differentiation*

Received for publication, May 26, 2004, and in revised form, July 16, 2004
Published, JBC Papers in Press, August 9, 2004, DOI 10.1074/jbc.M405883200

Sibylla Martinelli‡§; Mirjana Urosevic¶§, Arezoo Daryadel‡, Patrick Antony Oberholzer¶, Christa Baumann‖, Martin F. Fey‖, Reinhard Dummer¶, Hans-Uwe Simon‡, and Shida Yousefi‡**

From the ‡Department of Pharmacology, University of Bern, CH-3010 Bern, Switzerland, ¶Department of Dermatology, University of Zurich, CH-8091 Zurich, Switzerland, and ‖Institute of Medical Oncology, Inselspital, University of Bern, CH-3010 Bern, Switzerland

Interferons (IFNs) are cytokines that possess potent anti-viral and immunoregulatory activities. In contrast, their potential role(s) in anti-bacterial defense and neutrophil activation mechanisms is less well explored. By comparing gene expression patterns between immature and mature human neutrophils, we obtained evidence that intracellular proteases and other anti-bacterial proteins are produced at earlier stages of maturation, whereas the genes for receptors and signaling molecules required for the release of these effector molecules are preferentially induced during terminal differentiation. For instance, mature neutrophils strongly expressed genes that increase their responses to type I and type II IFNs. Interestingly, granulocyte/macrophage colony-stimulating factor was identified as a repressor of IFN signaling components and consequently of IFN-responsive genes. Both IFN-α and IFN-γ induced strong tyrosine phosphorylation of STAT1 in mature but not in immature neutrophils. Functional *in vitro* studies suggested that IFNs act as priming factors on mature neutrophils, allowing the formation of extracellular traps upon subsequent stimulation with complement factor 5a (C5a). In contrast, both IFN-α and IFN-γ had only little capacity to prime immature cells in this system. Moreover, both IFNs did not have significant anti-proliferative effects on immature neutrophils. These data contribute to our understanding regarding changes of gene expression during neutrophil differentiation and IFN-mediated anti-bacterial defense mechanisms.

Neutrophils are a critical component of the innate immune system with several effector and immunoregulatory functions (1). They are generated in the bone marrow under the influence of cytokines, such as granulocyte colony-stimulating factor (G-CSF)[1] and granulocyte/macrophage colony-stimulating factor (GM-CSF), from hematopoietic stem cells. Interestingly, G-CSF is not expressed in normal bone marrow cells under physiologic conditions (2), suggesting that it drives myeloid differentiation in a hormonal manner. Multiple cell types such as endothelial cells, epithelial cells, fibroblasts, and macrophages are able to produce G-CSF and GM-CSF (3, 4). All these cells make early contact with invading microorganisms and/or their products, resulting in increased cytokine production after infection. For instance, blood G-CSF levels have been described to rise from 25 to up to 10,000 pg/ml under pathologic conditions (5). Moreover, systemic injection of G-CSF (6) or GM-CSF (7) results in a dramatic increase of neutrophil production. Taken together, G-CSF and GM-CSF have been demonstrated to be major regulators of neutrophil production. Under conditions of stress, such as infections, neutrophil numbers in blood can increase as a consequence of cytokine-forced neutrophil differentiation.

Although immature neutrophils can be classified by morphology as well as by the expression of more or less specific surface proteins (8), it is difficult to obtain pure cell populations characterized by a certain maturation stage. Therefore, most of the studies trying to understand neutrophil differentiation at the molecular level were performed by using cell lines derived from leukemias. Previously published work resulted in the identification of genes that may play critical roles in the differentiation of neutrophils (9). Moreover, a gene expression profile of neutrophils has been established (10). Despite these previous studies, the underlying molecular events of normal neutrophil differentiation are not well understood, and many of the genes that are expressed by mature neutrophils have not been related to function.

The objective of this study was to compare the transcriptional repertoire of immature and mature human neutrophils by using oligonucleotide microarrays. In addition we investigated whether certain differences in gene expression are reversible by *in vitro* re-exposure of mature neutrophils with GM-CSF. Although multiple genes were more expressed in mature compared with immature cells, it was interesting to see that mature neutrophils also demonstrated higher expression of genes, which transduce signals of type I and type II interferons. Consequently, several known IFN-responsive genes had elevated expression levels in mature compared with immature cells. The subsequently obtained functional data demonstrate

* This work was supported by Swiss National Science Foundation Grants 31-68449.02 and 31-58916.99), the Bernische Krebsliga (Bern), and the Gottfried and Julia Bangerter-Rhyner Foundation (Zurich). The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Table: GSEA Results Summary

Dataset	WG-vs-C-for-GSEA_u_syms5_GaussianZ
Phenotype	NoPhenotypeAvailable
Upregulated in class	na_pos
GeneSet	MARTINELLI_IFNS_DIFF
Enrichment Score (ES)	0.8470181
Normalized Enrichment Score (NES)	1.9260814
Nominal p-value	0.0
FDR q-value	0.008648531
FWER p-Value	0.06

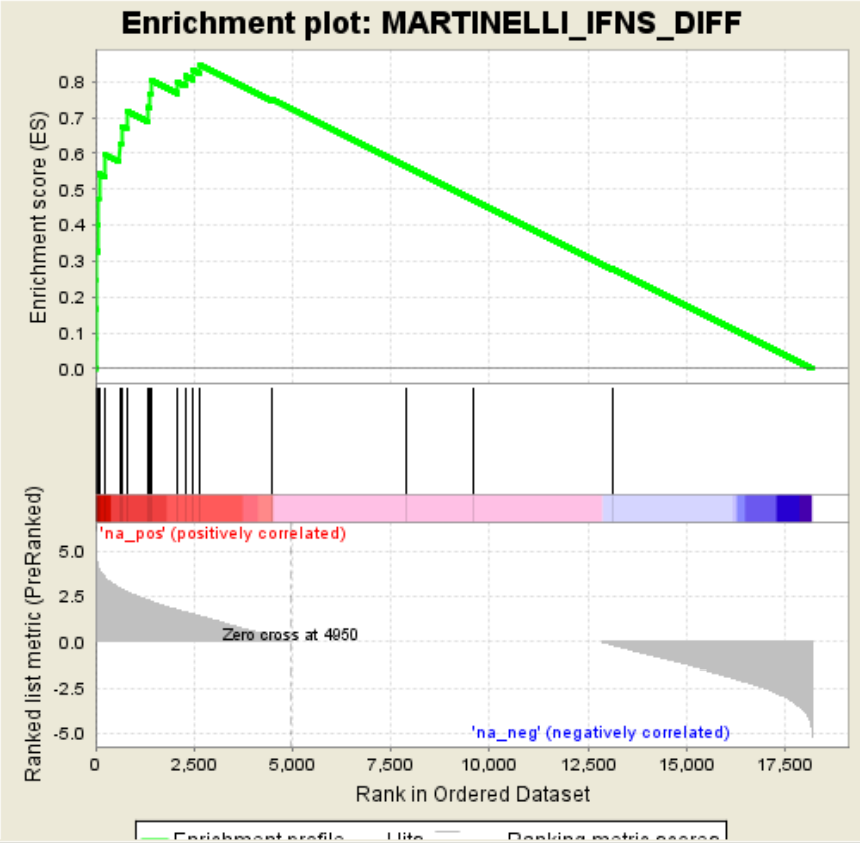


Fig 1: Enrichment plot: MARTINELLI_IFNS_DIFF
Profile of the Running ES Score & Positions of GeneSet Members on the Rank Ordered List

Table: GSEA details [plain text format]

	PROBE	GENE SYMBOL	GENE_TITLE	RANK IN GENE LIST	RANK METRIC SCORE	RUNNING ES	CORE ENRICHMENT
1	MPO			9	4.799	0.0838	Yes
2	LCN2			10	4.775	0.1677	Yes
3	AZU1			22	4.587	0.2477	Yes
4	BPI			27	4.540	0.3272	Yes
5	ELA2			57	4.233	0.4000	Yes
6	DEFA4			70	4.165	0.4725	Yes
7	CAMP			86	4.043	0.5426	Yes
8	CTSG			250	3.514	0.5954	Yes
9	CYBB			608	2.964	0.6278	Yes
10	CSF3R			657	2.910	0.6763	Yes
11	AOAH			791	2.773	0.7177	Yes
12	ALPL			1321	2.321	0.7294	Yes
13	NCF2			1364	2.277	0.7670	Yes
14	SOD2			1401	2.240	0.8044	Yes
15	FCGR3A			2086	1.761	0.7977	Yes
16	IL8RA			2277	1.642	0.8161	Yes
17	IL8RB			2468	1.528	0.8325	Yes
18	ST6GALNAC2			2656	1.413	0.8470	Yes
19	SEMA3C			4494	0.292	0.7510	No
20	GPR109B			7894	0.000	0.5639	No
21	MMP8			9597	0.000	0.4703	No
22	IL8			13116	-0.148	0.2792	No

Note large number of genes in the gene set at the top of the complete ranked list (relative to gene set size)

ARTHRITIS & RHEUMATISM
Vol. 62, No. 6, June 2010, pp 1744–1754
DOI 10.1002/art.27398

Transcription of Proteinase 3 and Related Myelopoiesis Genes in Peripheral Blood Mononuclear Cells of Patients With Active Wegener's Granulomatosis

Chris Cheadle, Alan E. Berger, Felipe Andrade, Regina James, Kristen Johnson, Tonya Watkins, Jin Kyun Park, Yu-Chi Chen, Eva Ehrlich, Marissa Mullins, Francis Chrest, Kathleen C. Barnes, and Stuart M. Levine

Objective. **Wegener's granulomatosis (WG) is a systemic inflammatory disease that is associated with substantial morbidity. The aim of this study was to understand the biology underlying WG and to discover markers of disease activity that would be useful for prognosis and treatment guidance.**

Methods. **Gene expression profiling was performed using total RNA from peripheral blood mononuclear cells (PBMCs) and granulocyte fractions from 41 patients with WG and 23 healthy control subjects. Gene set enrichment analysis (GSEA) was performed to search for candidate WG-associated molecular pathways and disease activity biomarkers. Principal compo-**

cluding remission status and disease activity, were determined using the Birmingham Vasculitis Activity Score for WG (BVAS-WG).

Results. **Eighty-six genes in WG PBMCs and 40 in WG polymorphonuclear neutrophils (PMNs) were significantly up-regulated relative to controls. Genes up-regulated in WG PBMCs were involved in myeloid differentiation, and these included the WG autoantigen PR3. The coordinated regulation of myeloid differentiation genes was confirmed by GSEA. The median expression values of the 86 up-regulated genes in WG PBMCs were associated with disease activity ($P = 1.3 \times$**