Downstream Analysis of Microarray Data: Beyond the genelist

Suraj Menon
Cancer Research UK Cambridge Institute

G3bp2
Rab8B
Col4a1
D830014E11Rik
Cxcl1
Adap1
Hspg2
Pxmp4
Marcks
Robo4
AK054271
Sdpr
Ahdc1
Oaf
Zfp143
Inpp5k
Npr2
Fas
Sult5a1
Sult1a1
Ndufa12
Lmo2
Abcb1b
Usp30
Gabra2
Cyp17a1
Saps3
Aldh16a1
Nrbp2
Fhl1
Cml2
Crtap
Cd93
Prodh
Rps8
Rdh9
1110033J19Rik
Tbcel
Phlda2
Rcan3
Tspan7
6430548M08Rik
Mfsd7b
A830073O21Rik
Darc
Tnfrsf11b
Hist1h2bk
Ndufc1
Tmprss2

So you have a genelist ...

TYPICAL RESULT OF PRIMARY ANALYSIS

- Differential gene expression analysis e.g. Limma + selection of genes at FDR and/or FC cutoff
- Classification/ clustering

WHAT NEXT?

- Annotation
- Exploratory analyses
- Focussed biological questions

So you have a genelist ...

G3bp2
Rab8B
Col4a1
D830014E11Rik
Cxcl1
Adap1
Hspg2
Pxmp4
Marcks
Robo4
AK054271
Sdpr
Ahdc1
Oaf
Zfp143
Inpp5k
Npr2
Fas
Sult5a1
Sult1a1
Ndufa12
Lmo2
Abcb1b
Usp30
Gabra2
Cyp17a1
Saps3
Aldh16a1
Nrbp2
Fhl1
Cml2
Crtap
Cd93
Prodh
Rps8
Rdh9
1110033J19Rik
Tbcel
Phlda2
Rcan3
Tspan7
6430548M08Rik
Mfsd7b
A830073O21Rik
Darc
Tnfrsf11b
Hist1h2bk
Ndufc1
Tmprss2

HOWEVER ...

– Manual annotation of genes a HIGHLY resource intensive process!

" ... biomedical research literature accumulates at a rate far surpassing that at which anyone can read it, let alone assimilate it."

"... at the current rate, one would need to scan in excess of 130 journals and read in excess of 27 papers a day to keep up with the field of Breast Cancer Genes"

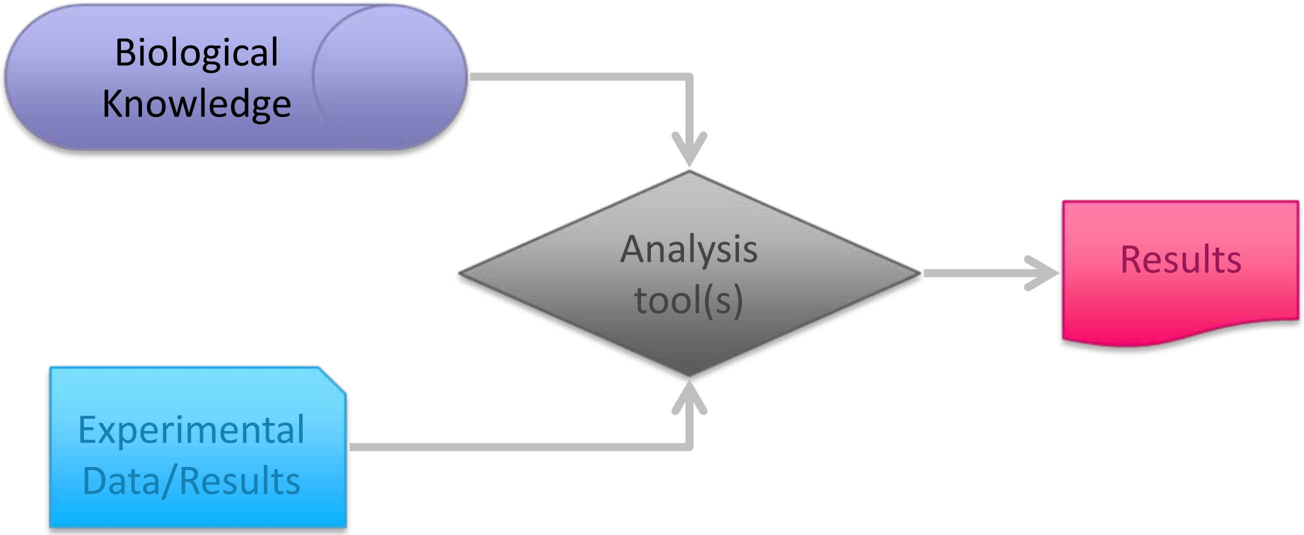
- Baasiri et al. *Oncogene* (1999)

UNIVERSITY OF
CAMBRIDGE

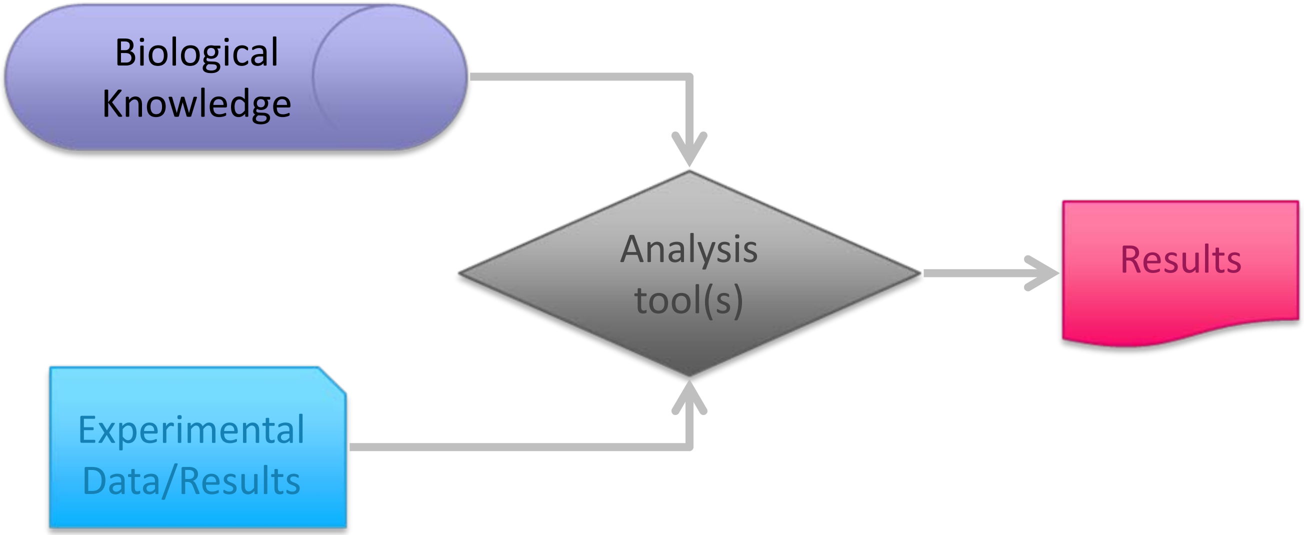
CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

Downstream Analysis of Microarray Data



Downstream Analysis of Microarray Data



DEEPER BIOLOGICAL UNDERSTANDING OF DATA

– Quickly
– Quantitative and structured results
– Reproducibility

UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Databases of Biological Knowledge

- Biomedical literature
- Biochemical pathways
- Functional annotation
- Ontologies
- Sequence information
- Interaction data
- TF/regulatory information
- Experimental data

the Gene Ontology
BIOCARTA
KEGG
Kyoto Encyclopedia of Genes and Genomes
PANTHER
Classification System
MINT
MSigDB
Molecular Signatures Database
ONCOMINE
IntAct
biomart
HomoloGene
Discover Homologs
BIOBASE
BIOLOGICAL DATABASES
REACTOME
PubMed Central
e!
GenMAPP
GeneDB
OMIM
Online Mendelian Inheritance in Man
GEO
Gene Expression Omnibus
RefSeq
genomes·transcripts·proteins

UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Popular downstream analysis workflows

ENRICHMENT OF BIOLOGICAL THEMES:

- What processes do my genes represent?
- What are the dominant biological pathways in my data?

Popular downstream analysis workflows

ENRICHMENT OF BIOLOGICAL THEMES:

- What processes do my genes represent?
- What are the dominant biological pathways in my data?

ANALYSIS OF GENE REGULATION (MOTIF ANALYSIS):

- Are most of my genes regulated by a particular transcription factor?

Popular downstream analysis workflows

ENRICHMENT OF BIOLOGICAL THEMES:

- What processes do my genes represent?
- What are the dominant biological pathways in my data?

ANALYSIS OF GENE REGULATION (MOTIF ANALYSIS):

- Are most of my genes regulated by a particular transcription factor?

NETWORK/INTERACTION ANALYSIS:

- Are there groups of highly interacting genes within my data?

Popular downstream analysis workflows

ENRICHMENT OF BIOLOGICAL THEMES:

- What processes do my genes represent?
- What are the dominant biological pathways in my data?

ANALYSIS OF GENE REGULATION (MOTIF ANALYSIS):

- Are most of my genes regulated by a particular transcription factor?

NETWORK/INTERACTION ANALYSIS:

- Are there groups of highly interacting genes within my data?

INTEGRATION WITH OTHER DATASETS/TECHNOLOGIES:

- E.g. Do my DE genes also exhibit differential transcription factor binding? (integrate with ChIP-Seq data)

Selecting a downstream analysis workflow

DEPENDENT ON THE BIOLOGICAL QUESTION!

- Not the other way around: This could cause confusion and difficulties in inference

CONCATENATE WORKFLOWS FOR MORE COMPLEX QUESTIONS

- E.g. Enrichment analysis of over-represented themes in a network of highly inter-connected genes

YOU ARE ONLY LIMITED BY YOUR IMAGINATION!

- ... And the availability of the right data in the right format
- ... And the application of appropriate statistics

ENRICHMENT OF BIOLOGICAL THEMES

Enrichment of Biological Themes

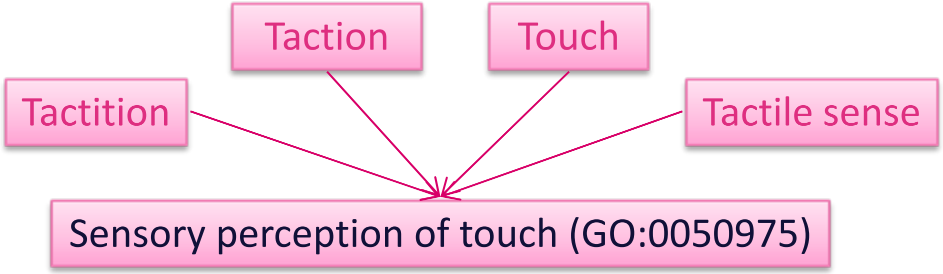
WHAT IS A THEME?

- A list of genes representing some aspect of biology
  - Biochemical pathways
  - Locations: subcellular compartments, chromosome band
  - Transcription factor targets
  - Gene interaction networks (experimental or literature based)
  - Experimental results
    - **differentially expressed genes**
    - **genes near ChIP-Seq binding sites**

Enrichment of Biological Themes

GENE ONTOLOGY

- Controlled vocabulary to describe gene function
  - One word can mean many things; many words can mean the same thing
  - Structured annotation



Taction

Touch

Tactition

Tactile sense

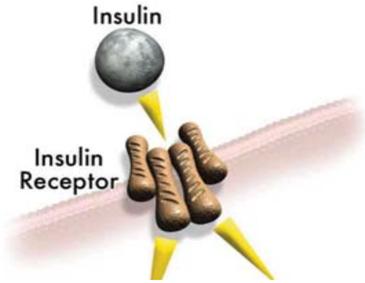
Sensory perception of touch (GO:0050975)

- Capture biological in computable form
  - Allows for quantitative analyses

Gene Ontology

1. Molecular Function

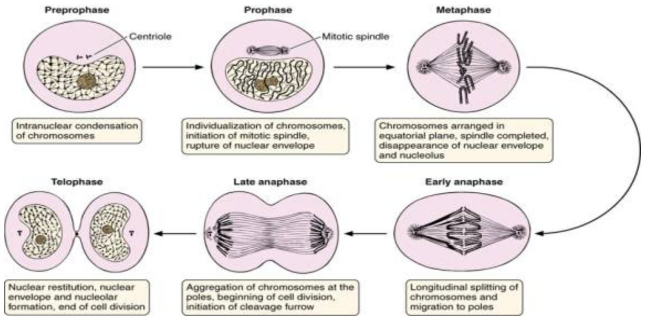
An elemental activity or task or job



- protein kinase activity
- insulin receptor activity

2. Biological Process

Commonly recognized series of events



Intranuclear condensation of chromosomes

Individualization of chromosomes, initiation of mitotic spindle, rupture of nuclear envelope

Chromosomes arranged in equatorial plane, spindle completed, disappearance of nuclear envelope and nucleolus

Nuclear restitution, nuclear envelope and nucleolar formation, end of cell division

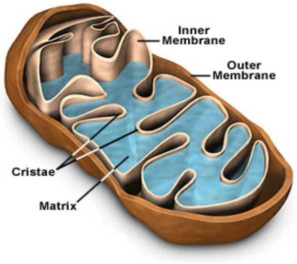
Aggregation of chromosomes at the poles, beginning of cell division, initiation of cleavage furrow

Longitudinal splitting of chromosomes and migration to poles

- cell division

3. Cellular Component

Where a gene product is located



- mitochondrion
- mitochondrial matrix

UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Enrichment Analysis Methodologies

OVER-REPRESENTATION ANALYSIS

- ‘Threshold-based’: require definition of a statistical threshold to define list of genes to test (e.g. FDR <0.01)
- Hypergeometric test, Fisher’s Exact test

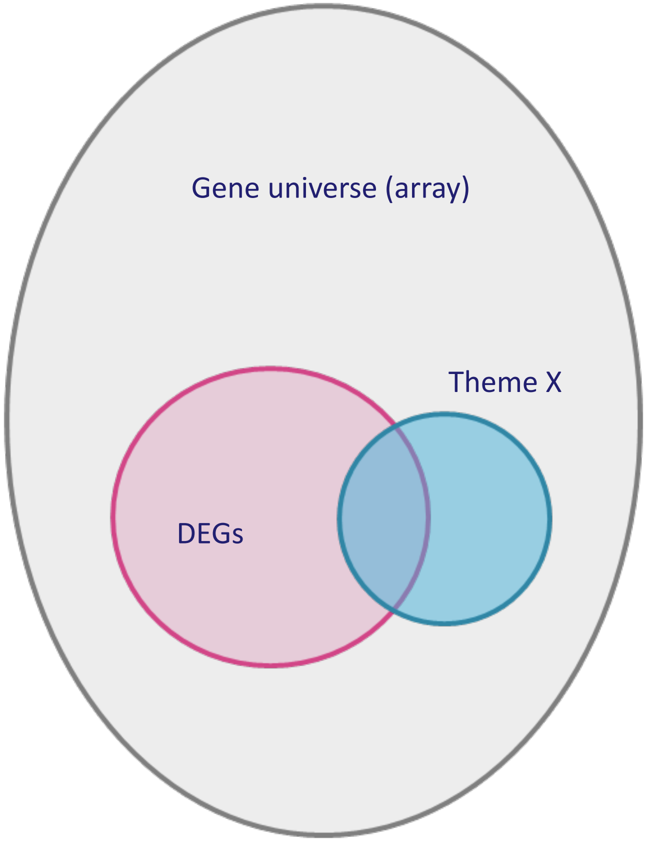
FUNCTIONAL CLASS SCORING

- ‘Threshold-free’: typically test all genes in dataset
- Gene Set Enrichment Analysis(GSEA), GlobalTest

PATHWAY TOPOLOGY BASED METHODS

- More complex analyses incorporating more data
- Signalling Pathway Impact Analysis (SPIA)

Over-Representation Analysis

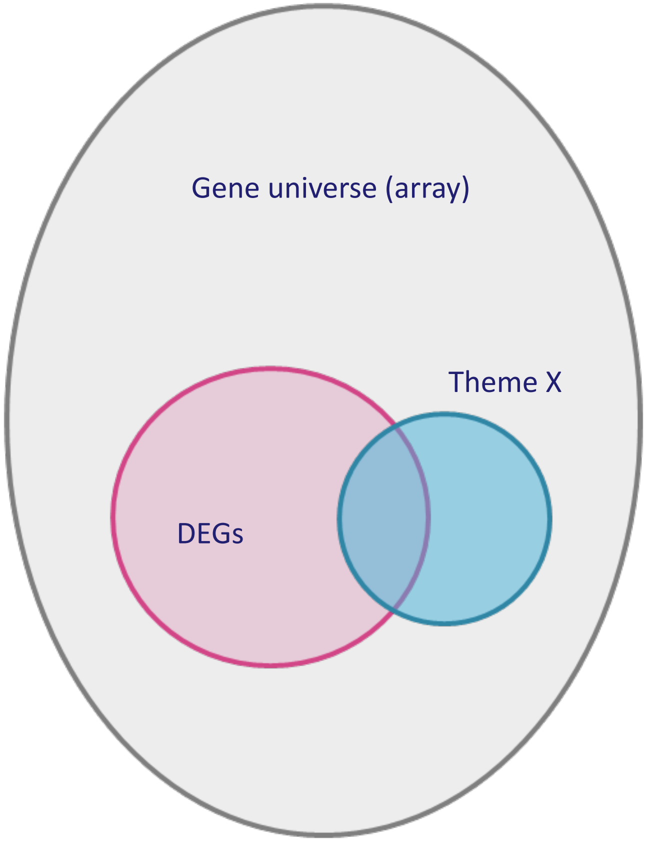


Are the number of DEGs associated with Theme X significantly greater than what might be expected by chance alone?

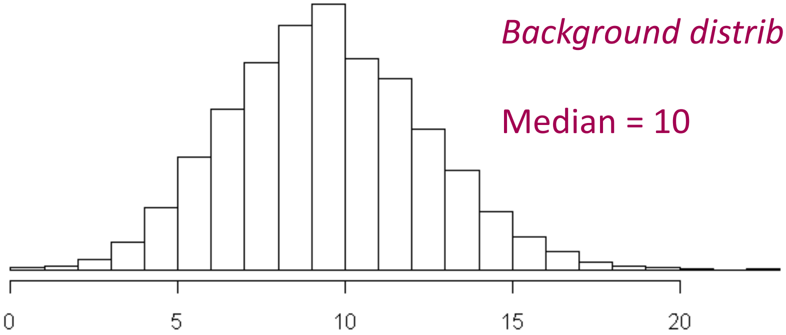
- 2000 genes on array
- 200 DEGs (10% of array)
- 100 genes associated with Theme X
- **Expected** size of overlap

  = 10% of Theme X = 10 genes

Over-Representation Analysis



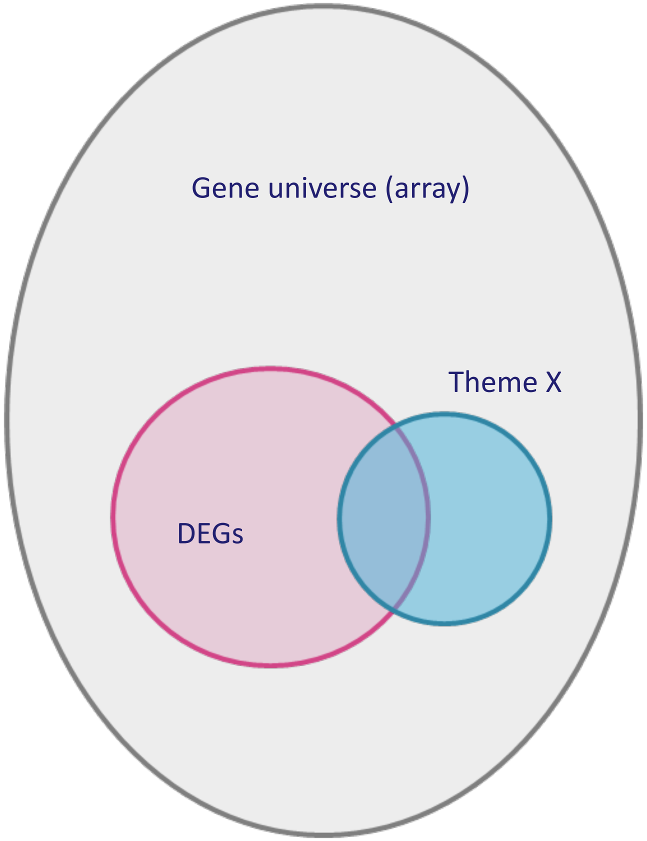
Simulation: n = 10,000



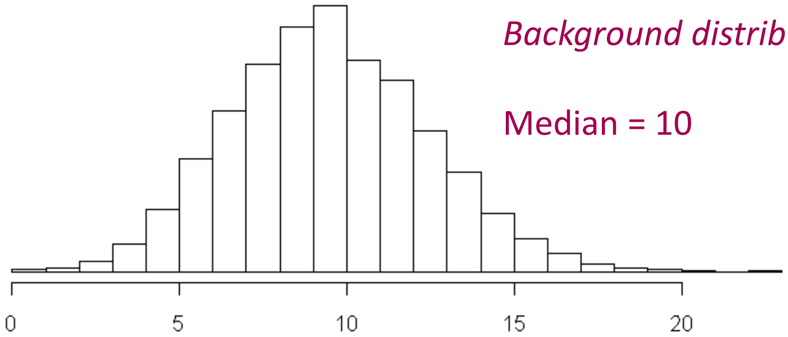
Background distribution

Median = 10

Over-Representation Analysis



Simulation: n = 10,000



Background distribution

Median = 10

Observations with an overlap size of:

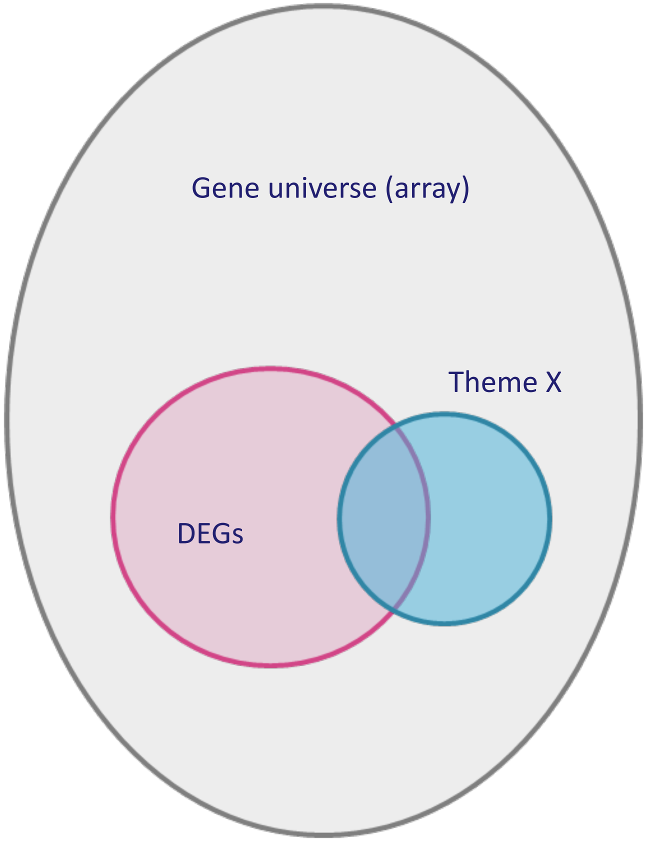
>= 5 : 9779 (/10000 = ***0.9779***)

>=10: 5544 (/10000 = ***0.5544***)

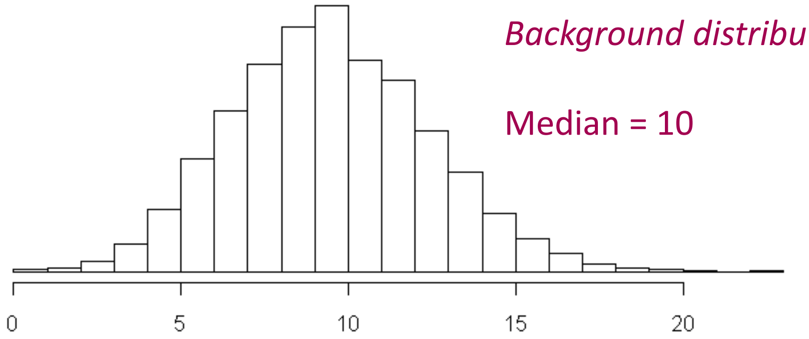
>=15: 656 (/10000 = ***0.0656***)

>= 20: 18 (/10000 = ***0.0018***)

Over-Representation Analysis



Simulation: n = 10,000



Background distribution

Median = 10

Observations with an overlap size of:

>= 5 : 9779 (/10000 = ***0.9779***)

>=10: 5544 (/10000 = ***0.5544***)

>=15: 656 (/10000 = ***0.0656***)

>= 20: 18 (/10000 = ***0.0018***)

p-values

Over Representation Analysis

HOWEVER:

- The statistical models are simple and make lots of assumptions
- E.g. All genes in the array are equally likely to be DE

  but only ~50% of genes are expressed in any tissue at any particular time

- Increased likelihood of false positives

  larger overlaps expected in a smaller universe

CONSTRAINING UNIVERSE SIZE IS IMPORTANT

- Non-specific filtering
- E.g. Using only the 50% most variable genes on the array

Over Representation Analysis

AN EXAMPLE (EXAGGERATED FOR EFFECT)

Parameter	Estimated	Reality
Gene universe	2000 genes	1000 genes
DEGs	200 (10% of universe)	200 (20% of universe)
Theme X	100	100
Expected overlap size (random)	10	20

Over Representation Analysis

AN EXAMPLE (EXAGGERATED FOR EFFECT)

Parameter	Estimated	Reality
Gene universe	2000 genes	1000 genes
DEGs	200 (10% of universe)	200 (20% of universe)
Theme X	100	100
Expected overlap size (random)	10	20

- Overlap size of 10 with universe size of 2000 : p = ~0.55
- Overlap size of 20 with universe size of 2000: p = ~0.002!

Over Representation Analysis

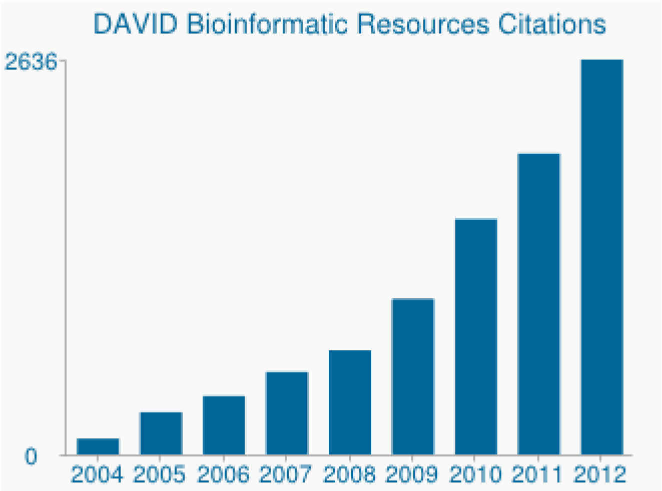
- R/Bioconductor: In today’s practical (GOstats)
- DAVID (**D**atabase for **A**nnotation, **V**isualisation and **I**ntegrated **D**iscovery)

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

Over Representation Analysis

- R/Bioconductor: In today’s practical (GOstats)
- DAVID (**D**atabase for **A**nnotation, **V**isualisation and **I**ntegrated **D**iscovery)

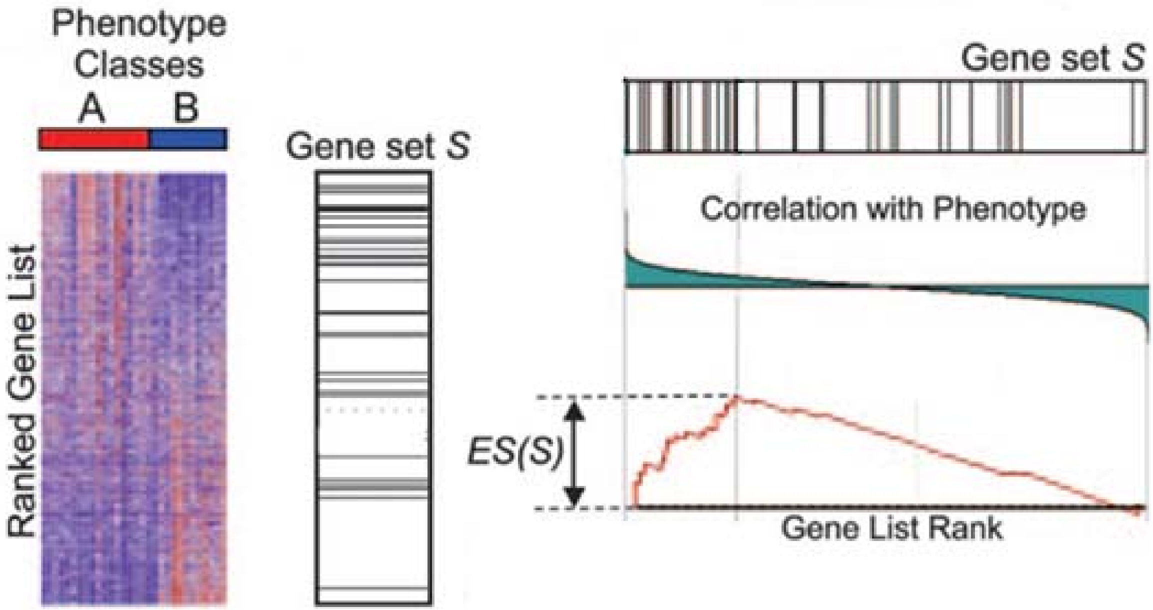
http://david.abcc.ncifcrf.gov/



- **>10,000 citations**
- **Daily Usage: ~1200 gene lists from ~400 unique researchers.**
- **Total Usage: ~800,000 gene lists from >5,000 research institutes world-wide**
- **Wide range of themes covered**
- **Clustering of redundant annotation terms**
- **Other useful tools e.g. Gene ID converter**

Gene Set Enrichment Analysis

- Avoids having to define which genes to test: uses all genes
- Useful for dirty data: theoretically more robust

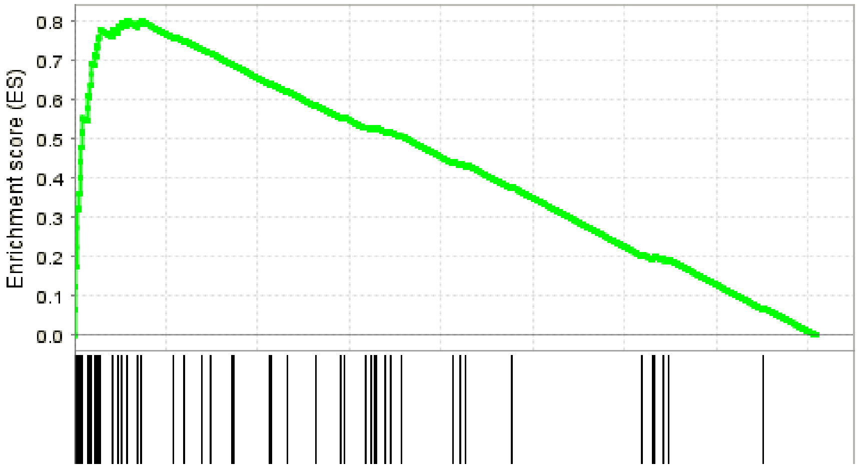


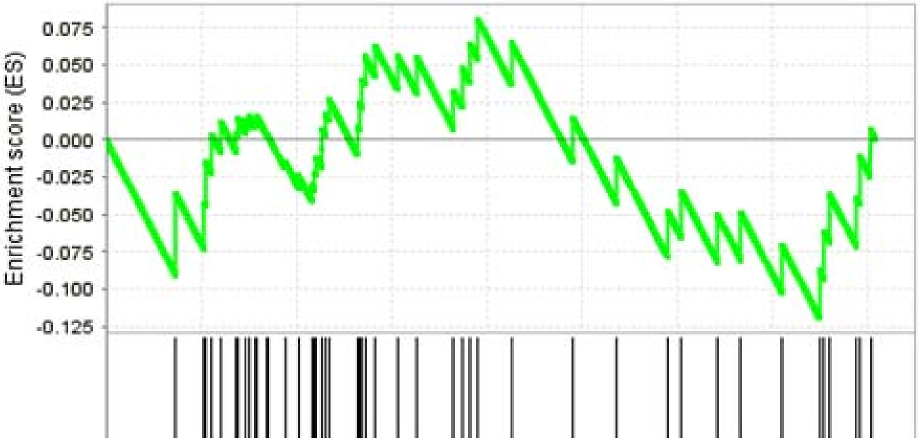
Images from Subramanian et al, *PNAS* 2005

UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Gene Set Enrichment Analysis





Gene Set Enrichment Analysis

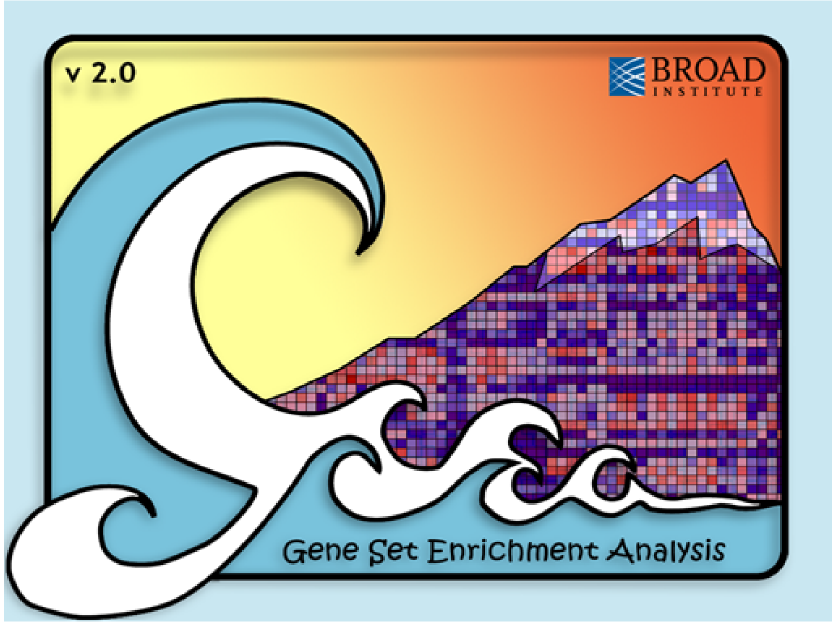
USING GSEA: http://www.broadinstitute.org/gsea/index.jsp

- Java application
- Will require formatting input data in R/Excel
- Use default gene ranking (input all data) or lists of ranked genes with weighting (GSEAPreranked Tool)

  Moderated T-statistic

  Signed –log10 p-value

  Log fold change



MOTIF ANALYSIS

UNIVERSITY OF
CAMBRIDGE

CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

What is a motif?

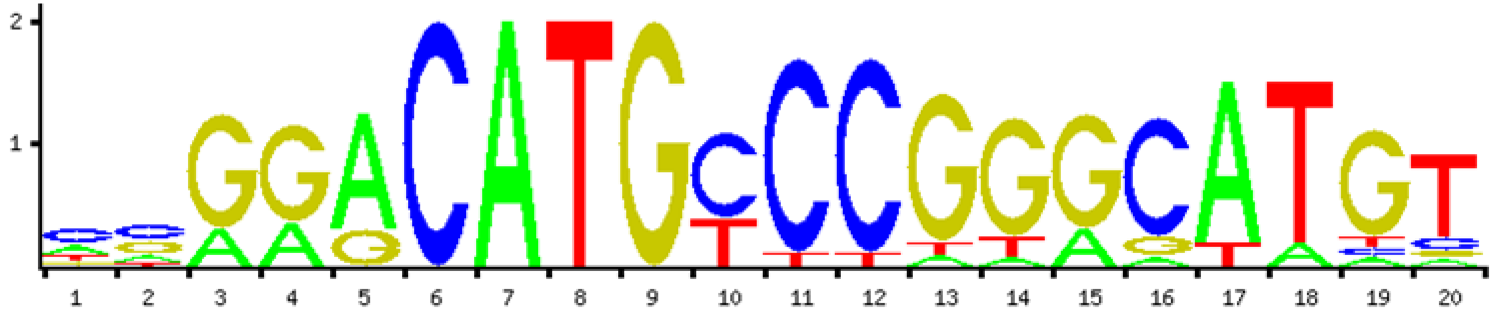
SHORT RECURRING SEQUENCE OF DNA

- Presumed to have some biological function
- Typically degenerate
- Represented by position weight matrices/ sequence logos

What is a motif?

SHORT RECURRING SEQUENCE OF DNA

- Presumed to have some biological function
- Typically degenerate
- Represented by position weight matrices/ sequence logos



A	[	5	3	4	5	13	0	17	0	0	0	0	0	1	1	4	1	15	2	1	1	]
C	[	8	7	0	0	0	17	0	0	0	11	16	16	0	0	0	14	0	0	1	2	]
G	[	2	6	13	12	4	0	0	0	17	0	0	0	15	14	13	2	0	0	14	1	]
T	[	2	1	0	0	0	0	0	17	0	6	1	1	1	2	0	0	2	15	1	13	]

UNIVERSITY OF
CAMBRIDGE

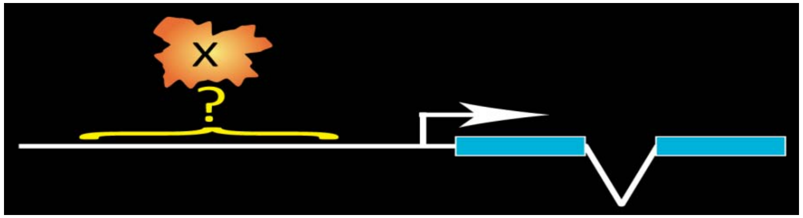
CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

Motif Analysis Methodologies

PATTERN MATCHING:

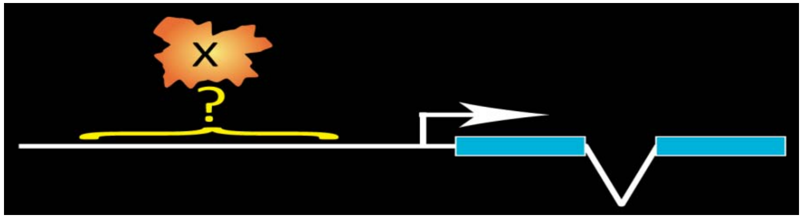
FINDING KNOWN MOTIFS



- Does protein X bind upstream of my genes?
- Does it bind more than expected by chance?

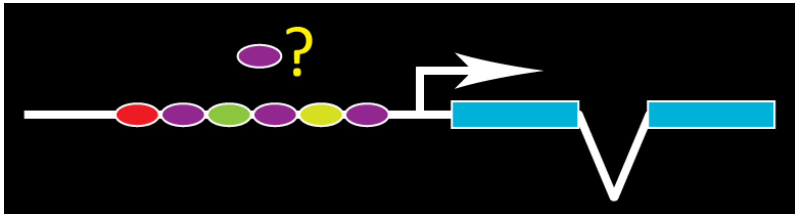
Motif Analysis Methodologies

PATTERN MATCHING:
FINDING KNOWN MOTIFS



- Does protein X bind upstream of my genes?
- Does it bind more than expected by chance?

PATTERN DISCOVERY:
FINDING UNKNOWN MOTIFS



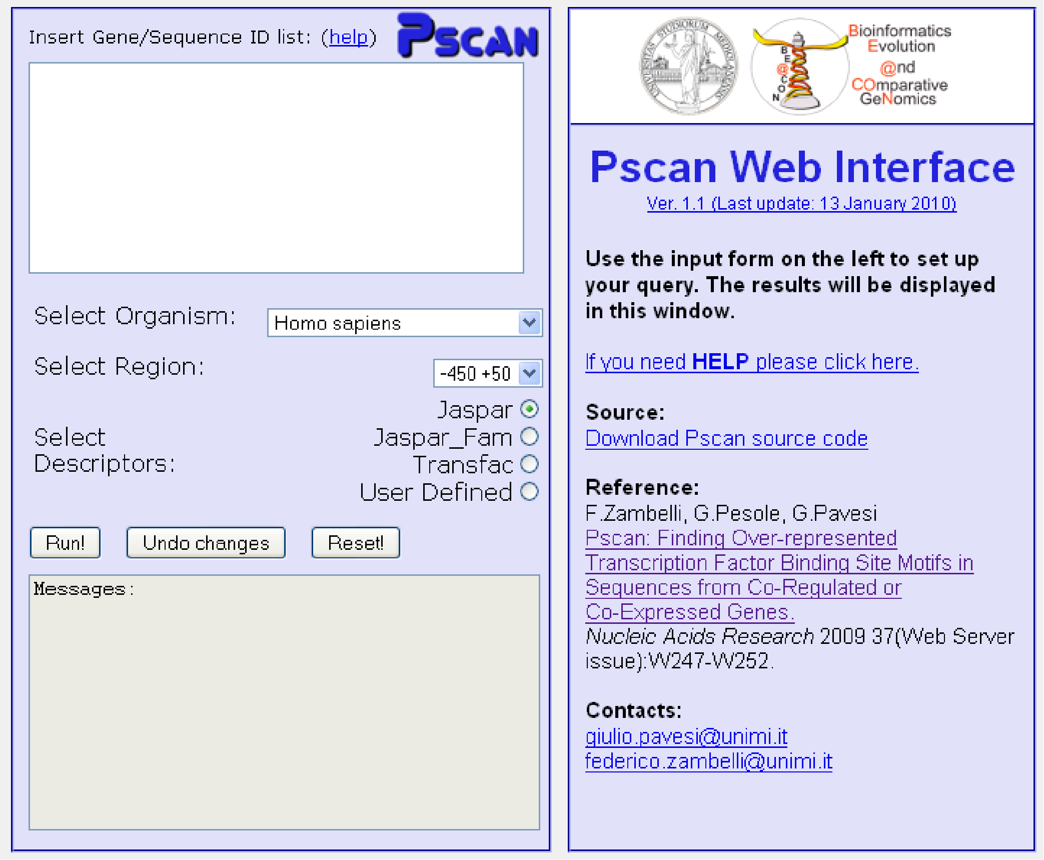
- Are there common motifs upstream of my genes?
- What are these motifs?

UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Motif Analysis Tools: PScan http://159.149.160.51/pscan/

PATTERN MATCHING


Insert Gene/Sequence ID list: (help)

Select Organism: Homo sapiens

Select Region: -450 +50

Select Descriptors: Jaspar ● Jaspar_Fam ○ Transfac ○ User Defined ○

Messages:

Pscan Web Interface

Ver. 1.1 (Last update: 13 January 2010)

Use the input form on the left to set up your query. The results will be displayed in this window.

If you need **HELP** please click here.

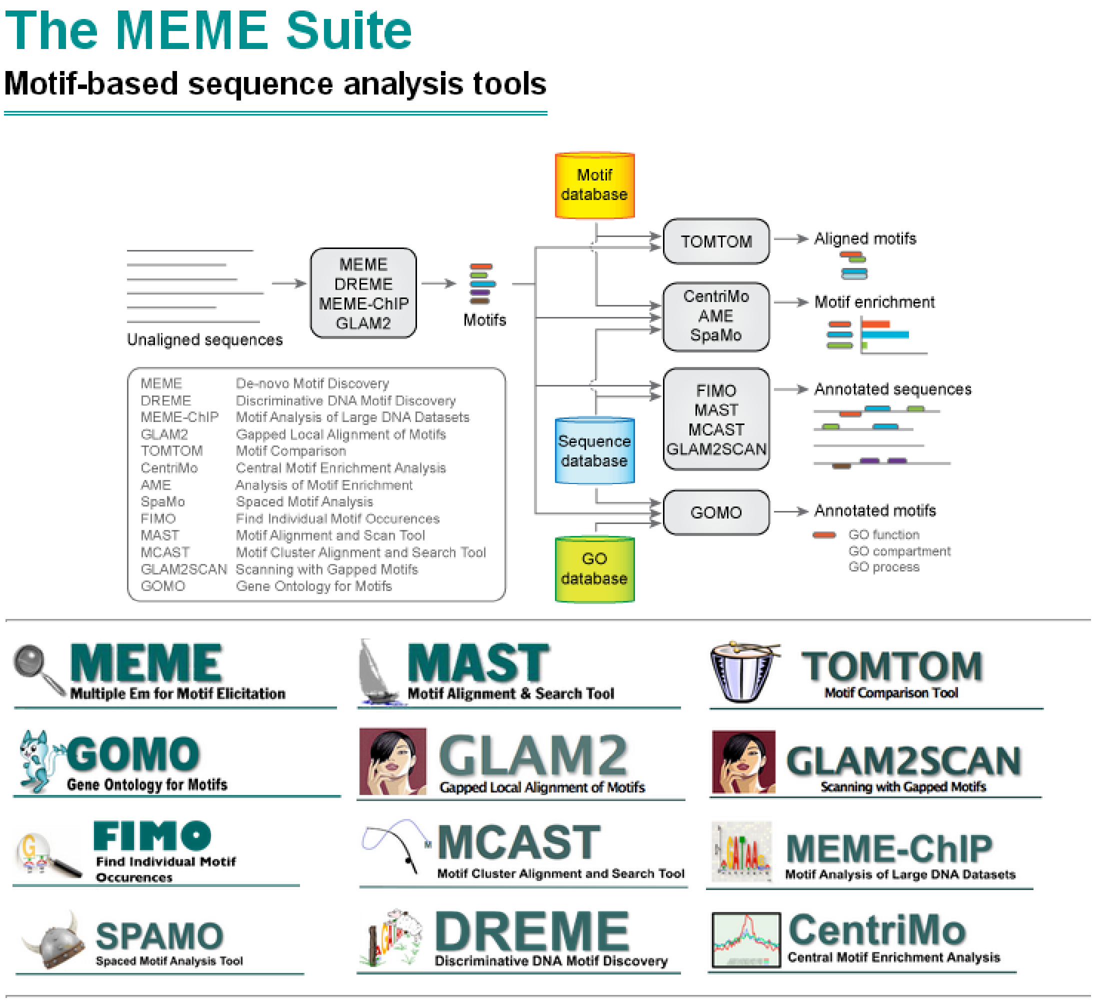
Source:
Download Pscan source code

Reference:
F.Zambelli, G.Pesole, G.Pavesi
Pscan: Finding Over-represented Transcription Factor Binding Site Motifs in Sequences from Co-Regulated or Co-Expressed Genes.
Nucleic Acids Research 2009 37(Web Server issue):W247-W252.

Contacts:
giulio.pavesi@unimi.it
federico.zambelli@unimi.it

Motif Analysis Tools: The MEME Suite
http://meme.nbcr.net/meme/intro.html

PATTERN DISCOVERY (AND VARIOUS OTHER FUNCTIONS)



The MEME Suite

Motif-based sequence analysis tools

MEME	De-novo Motif Discovery
DREME	Discriminative DNA Motif Discovery
MEME-ChIP	Motif Analysis of Large DNA Datasets
GLAM2	Gapped Local Alignment of Motifs
TOMTOM	Motif Comparison
CentriMo	Central Motif Enrichment Analysis
AME	Analysis of Motif Enrichment
SpaMo	Spaced Motif Analysis
FIMO	Find Individual Motif Occurences
MAST	Motif Alignment and Scan Tool
MCAST	Motif Cluster Alignment and Search Tool
GLAM2SCAN	Scanning with Gapped Motifs
GOMO	Gene Ontology for Motifs

MEME
Multiple Em for Motif Elicitation
MAST
Motif Alignment & Search Tool
TOMTOM
Motif Comparison Tool
GOMO
Gene Ontology for Motifs
GLAM2
Gapped Local Alignment of Motifs
GLAM2SCAN
Scanning with Gapped Motifs
FIMO
Find Individual Motif Occurences
MCAST
Motif Cluster Alignment and Search Tool
MEME-ChIP
Motif Analysis of Large DNA Datasets
SPAMO
Spaced Motif Analysis Tool
DREME
Discriminative DNA Motif Discovery
CentriMo
Central Motif Enrichment Analysis

UNIVERSITY OF CAMBRIDGE

Motif Analysis

FURTHER INFORMATION

- Stormo GD. DNA binding sites: representation and discovery. Bioinformatics. 2000 Jan;16(1):16-23. Review. PubMed PMID: 10812473.

- D’haeseleer P. How does DNA sequence motif discovery work? Nat Biotechnol. 2006 Aug;24(8):959-61. Review. PubMed PMID: 16900144.

- Das MK, Dai HK. A survey of DNA motif finding algorithms. BMC Bioinformatics. 2007 Nov 1;8 Suppl 7:S21. Review. PubMed PMID: 18047721

- Tompa M, Li N et.al. Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol. 2005 Jan;23(1):137-44. PubMed PMID: 15637633.

NETWORK ANALYSIS WITH CYTOSCAPE

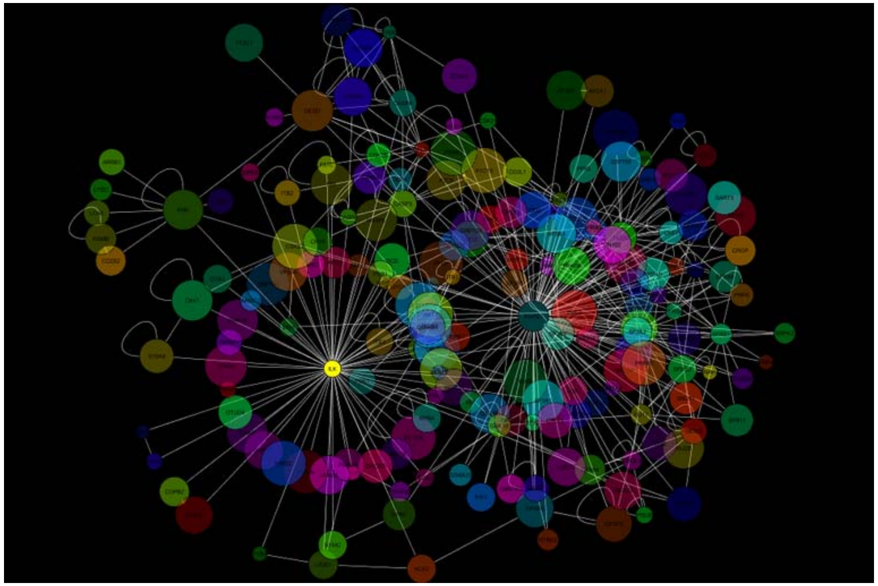
UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Network Analysis using Cytoscape

WHAT IS CYTOSCAPE?

- Interactive tool for visualisation and manipulation of network data
- Free and open source
- Java (cross platform)
- Plugins extend functionality
- Large developer community

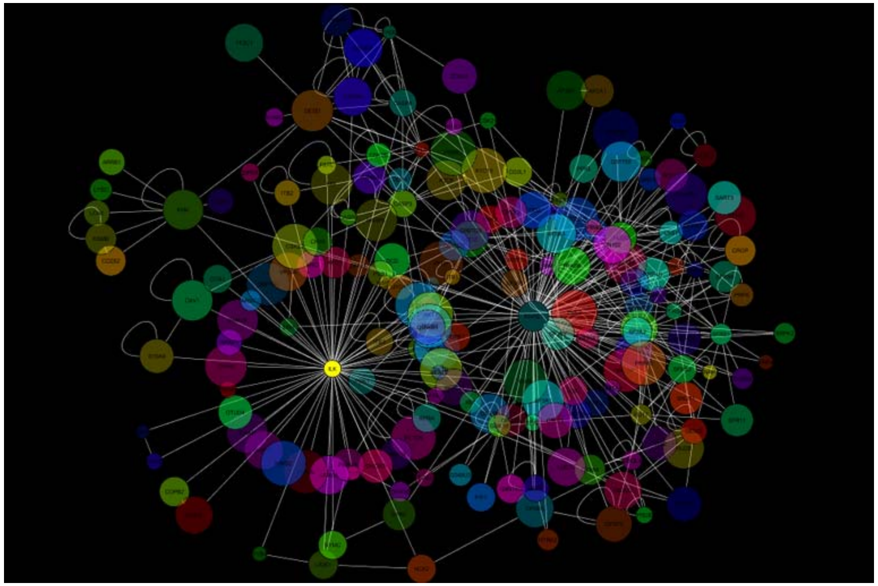


UNIVERSITY OF CAMBRIDGE

Network Analysis using Cytoscape

DATA RETRIEVAL AND INTEGRATION

Interaction data, pathways, literature searches etc.

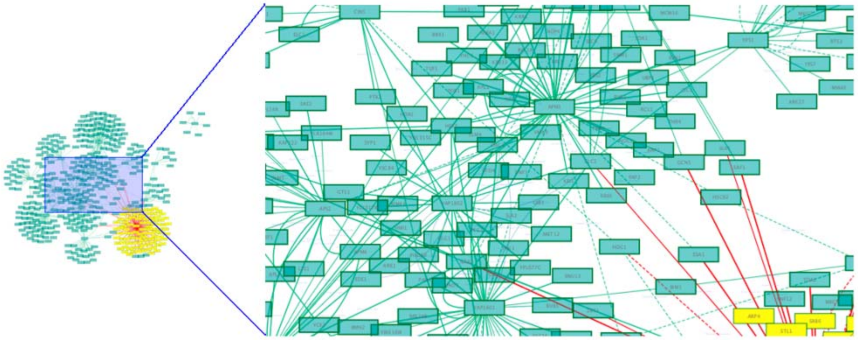


VISUALISATION, EXPLORATION AND MANIPULATION

VizMapper, VistaClara plugins

DATA ANALYSIS

MCODE, BinGO plugins



UNIVERSITY OF CAMBRIDGE

Network Analysis using Cytoscape

FURTHER INFORMATION

- Integration of biological networks and gene expression data using Cytoscape. Cline et. al. Nature Protocols 2, - 2366 - 2382 (2007)

- Cytoscape: a software environment for integrated models of biomolecular interaction networks. Shannon et. al. Genome Research 13(11):2498-504. (2003)

- Exploring biological networks with Cytoscape software. Curr Protoc Bioinformatics. 2008 Sep;Chapter 8:Unit 8.13.

http://www.cytoscape.org

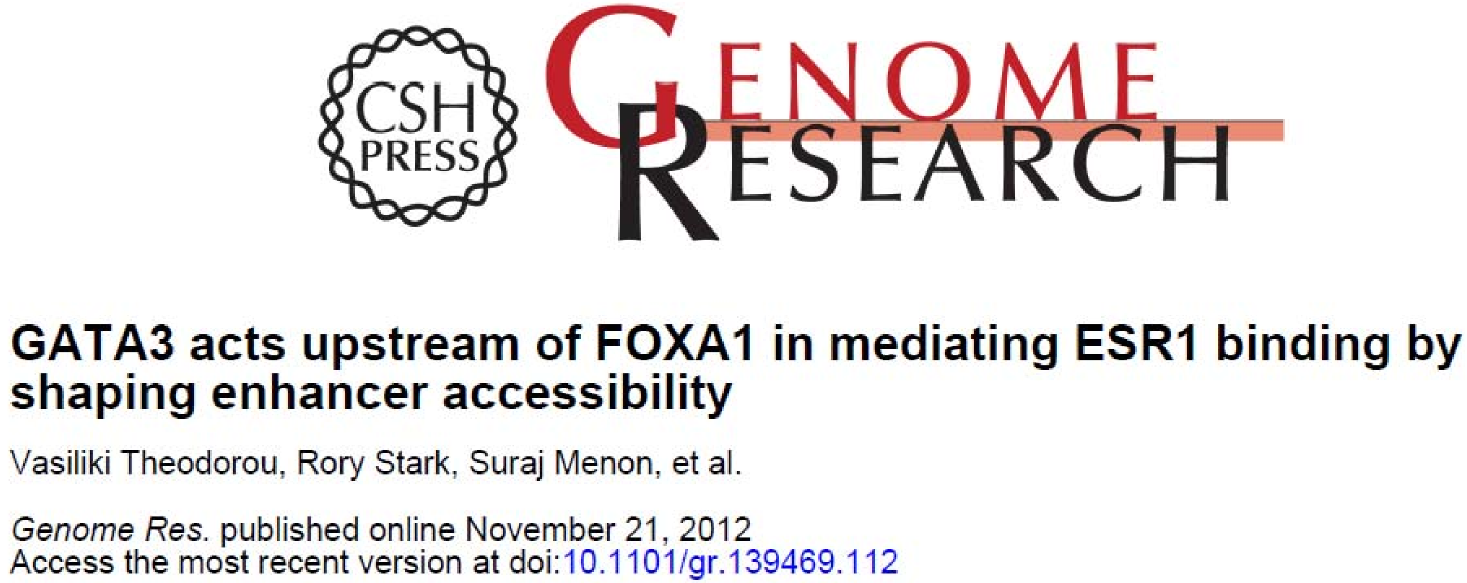
CROSS-DATASET INTEGRATIVE ANALYSES

Cross-dataset integrative analyses: Important considerations

DEFINE THE BIOLOGICAL QUESTION!!

- Which datasets to integrate?
- Integrate data at what level?
  - **Normalised data? Primary or secondary results?**
- How to translate across datasets? (e.g. Cross-platform/cross-technology analyses)
- What statistical tests/ metrics to use?

An Example: Integrating ChIP-Seq and Expression Microarray Data



CSH PRESS

GENOME RESEARCH

GATA3 acts upstream of FOXA1 in mediating ESR1 binding by shaping enhancer accessibility

Vasiliki Theodorou, Rory Stark, Suraj Menon, et al.

Genome Res. published online November 21, 2012
Access the most recent version at doi:10.1101/gr.139469.112

An Example: Integrating ChIP-Seq and Expression Microarray Data

THE DATA

- ChIP-Seq: Differentially bound sites for ESR1 in Control v GATA3 KD conditions in MCF7 cells (‘Stronger’ and ‘Weaker’)
- Array: Differentially expressed genes (DEGs) for Control v GATA3 KD in MCF7 cells (Up- and down-regulated genes)

THE CONCEPT

- Link the differentially bound sites with the DEGs
- Illustrate that ESR1 re-programming wrt GATA3 is ‘functional’

An Example: Integrating ChIP-Seq and Expression Microarray Data

DATA INTEGRATION

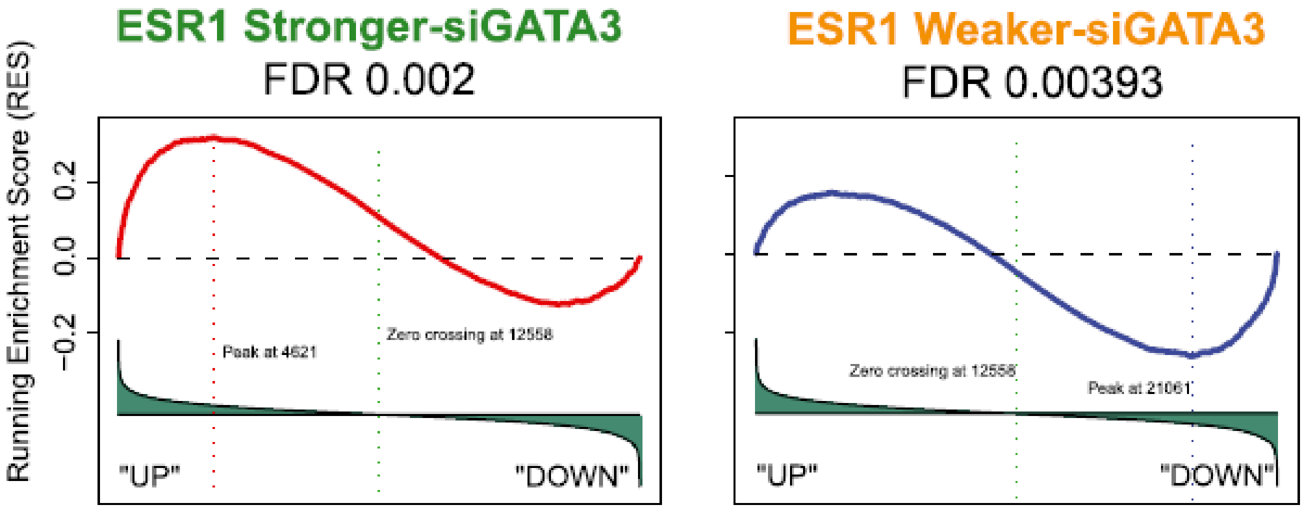
- Arrays: Probe -> Gene Symbol
  - Select most variable probe per gene symbol (IQR)
- ChIP-Seq: Differentially bound sites -> Gene Symbol
  - Overlap sites with 50KB window around gene TSS

STATISTICAL ANALYSIS

- GSEA (ChIP lists v ranked array genes)
- Hypergeometric testing (ChIP lists v DEGs)

An Example: Integrating ChIP-Seq and Expression Microarray Data

GSEA (Integrated analysis of ChIP-Seq and expression datasets)



ESR1 Stronger-siGATA3
FDR 0.002

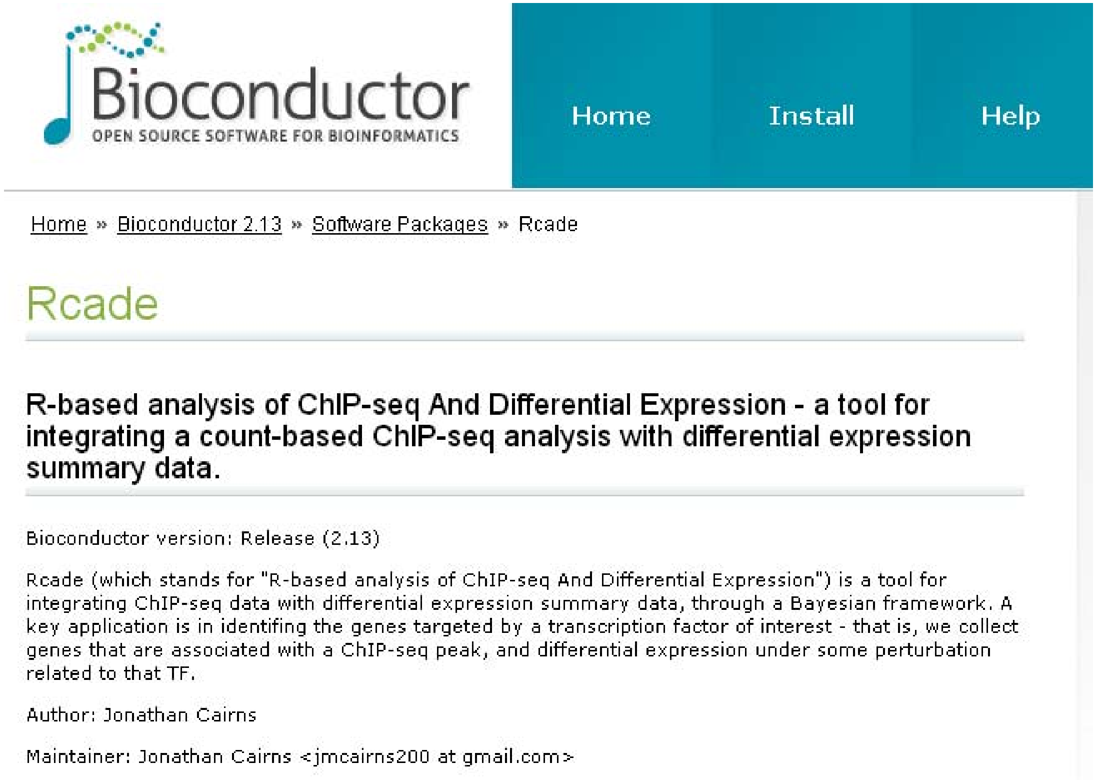
ESR1 Weaker-siGATA3
FDR 0.00393

Another Example: Integrating ChIP-Seq and Expression Microarray Data

DIFFERENT METHODOLOGY

DIFFERENT DATA

DIFFERENT QUESTION!



Home » Bioconductor 2.13 » Software Packages » Rcade

Rcade

R-based analysis of ChIP-seq And Differential Expression - a tool for integrating a count-based ChIP-seq analysis with differential expression summary data.

Bioconductor version: Release (2.13)

Rcade (which stands for "R-based analysis of ChIP-seq And Differential Expression") is a tool for integrating ChIP-seq data with differential expression summary data, through a Bayesian framework. A key application is in identifing the genes targeted by a transcription factor of interest - that is, we collect genes that are associated with a ChIP-seq peak, and differential expression under some perturbation related to that TF.

Author: Jonathan Cairns

Maintainer: Jonathan Cairns <jmcairns200 at gmail.com>

UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

COMMERCIAL SOFTWARE

UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Commercial Software

METACORE:

http://thomsonreuters.com/metacore/

INGENUITY PATHWAY ANALYSIS (IPA):

http://www.ingenuity.com/products/ipa

Functions/Tools

- Enrichment Analyses – pathways, disease/metabolic/drug target networks
- Network Analyses
- Knowledgebase search

Advantages

- High quality, manually curated (!) data
- High quality reporting and visualisation
- Highly interactive
- User friendly
- Comprehensive help and documentation

UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

ON TO THE PRACTICAL

Thanks to:

Stewart MacArthur

Contact:

Suraj.Menon@cruk.cam.ac.uk

UNIVERSITY OF
CAMBRIDGE

CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE