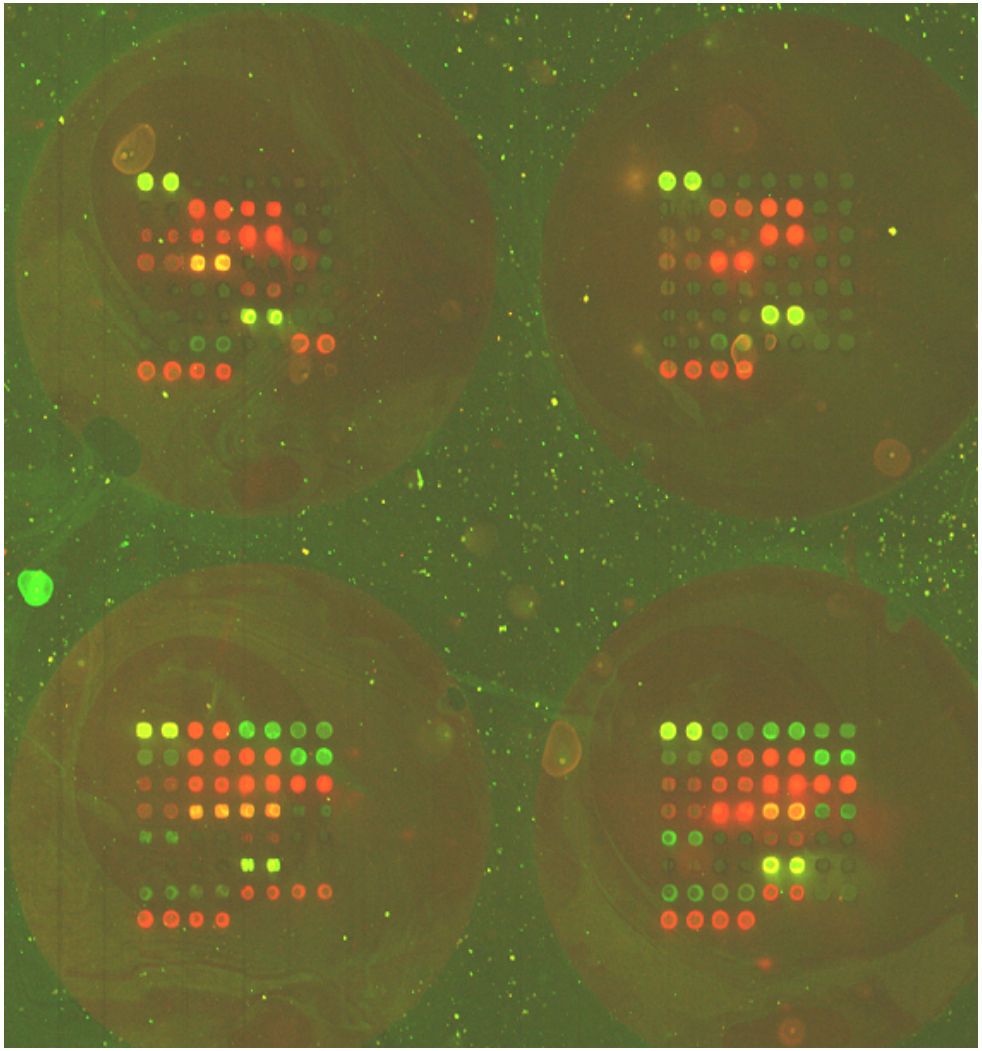
Dal dato grezzo al significato biologico

Analisi di dati prodotti mediante microarray

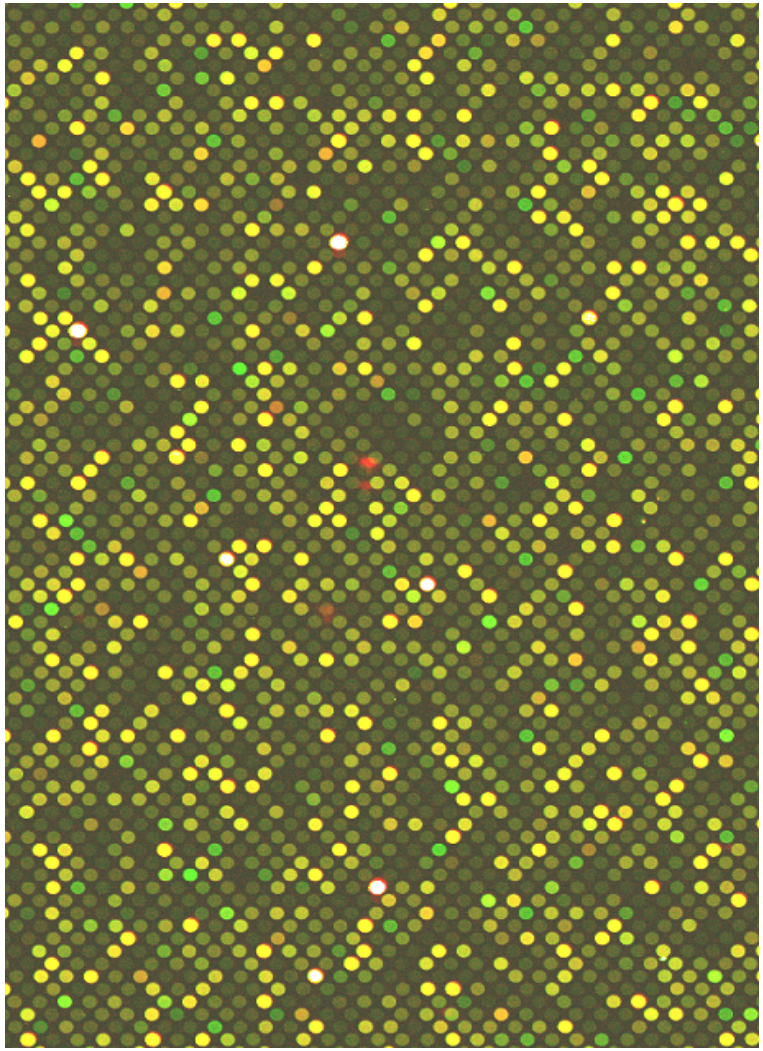
- Erika Melissari -

Microarray

GENOTIPIZZAZIONE



ESPRESSIONE GENICA



Esperimenti realizzati con Microarray

GENOTIPIZZAZIONE ED ESPRESSIONE GENICA

Somiglianze

- Parallelizzazione di molti esperimenti (~$10^{\wedge 3}$, $10^{\wedge 4}$)
- Elaborazione simultanea di molte informazioni (~$10^{\wedge 3}$, $10^{\wedge 4}$)
- Rilevazione dell'informazione biologica attraverso l'uso di molecole fluorescenti
  - scansione del vetrino per generare un'immagine dell'esperimento
  - quantizzazione numerica del segnale di colore
  - rumore di fondo dell'immagine generato da fluorescenze aspecifiche
- Normalizzazione dei dati
- …costo elevato di ogni esperimento

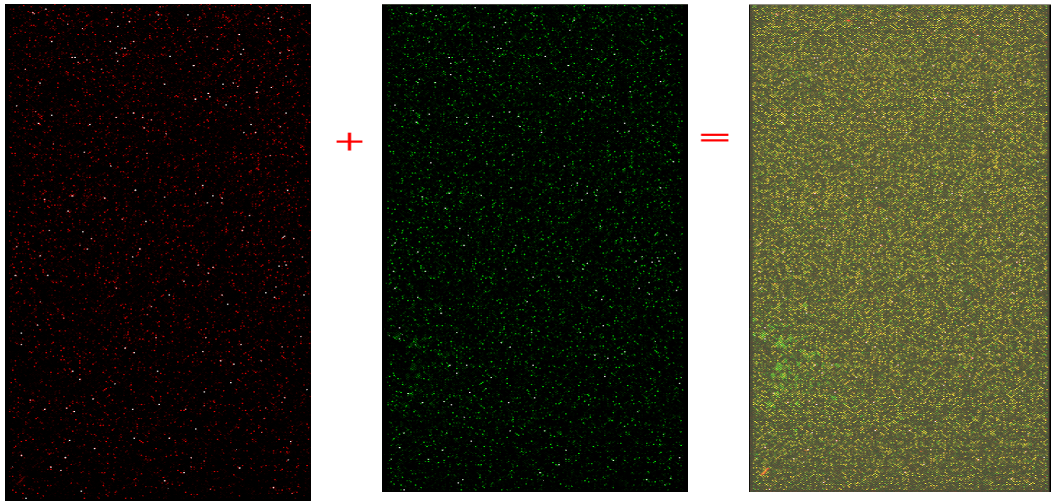
Esperimenti realizzati con Microarray

GENOTIPIZZAZIONE ED ESPRESSIONE GENICA

Differenze

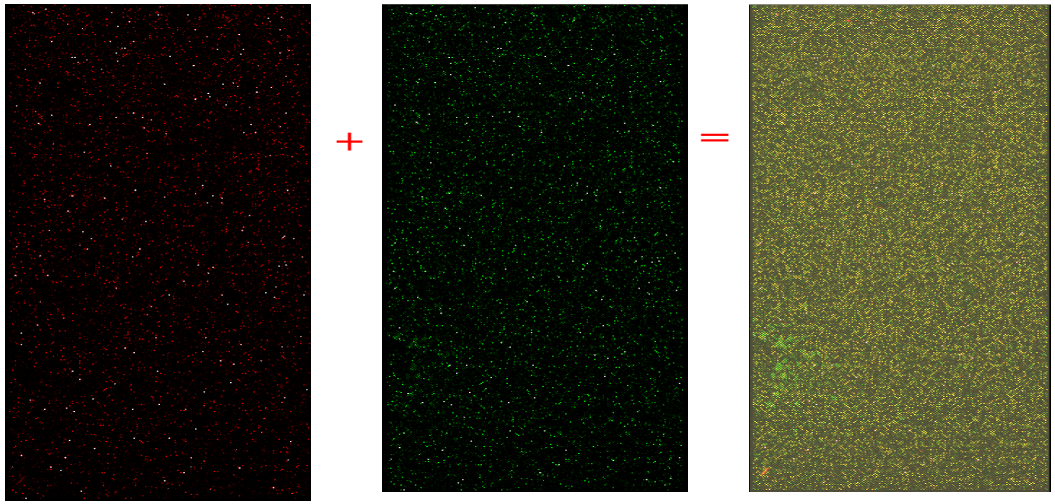
- Materiale biologico ibridizzato
- Tipo di sonde ancorate sul vetrino
- Analisi statistica
  - analisi di associazione o regressione delle frequenze degli SNP rispetto all'outcome rilevato
  - analisi di significatività statistica per ricavare la lista dei geni differenzialmente espressi
- Interpretazione biologica

Scansione del vetrino



- Scanner a due laser
  - Lunghezze d'onda di eccitazione dei fluorocromi
    - 635 nm - Red
    - 532 nm - Green
- Canali separati in acquisizione
  - formazione di due immagini
- Risoluzione di colore e spaziale
  - $2^{\wedge 16}$ = 65536 livelli di colore
  - 5 micron
- Occupazione di memoria
  - 130 MB c.a.

Acquisizione dell'immagine

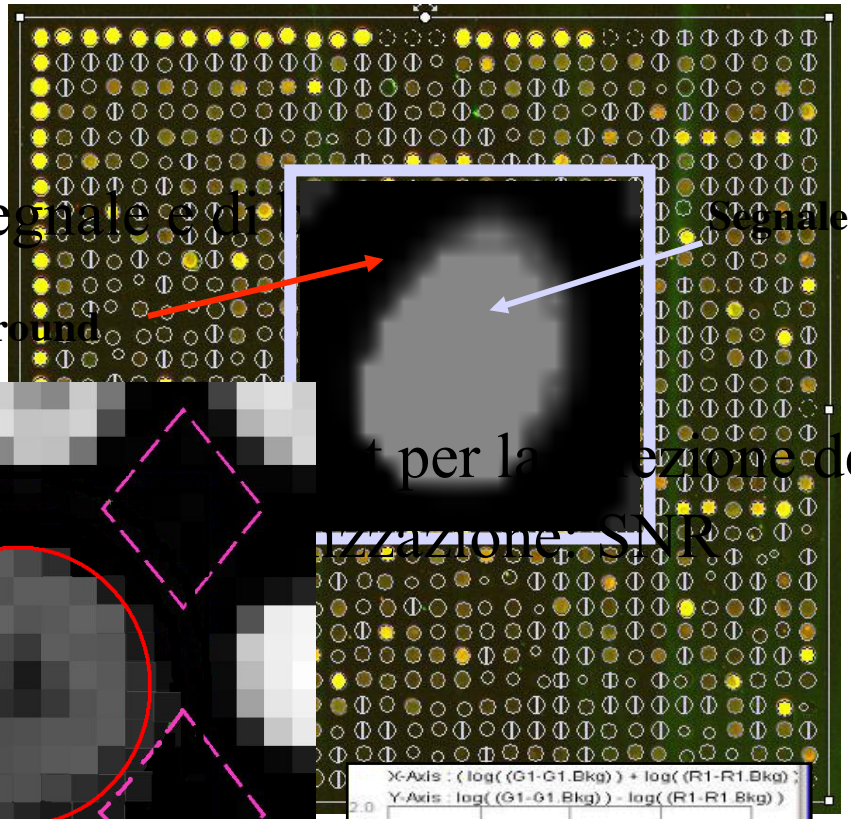


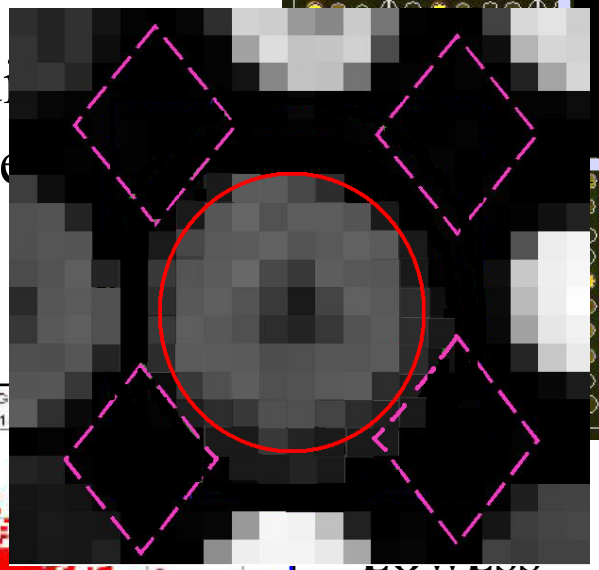
Impostazione corretta dei parametri dello scanner ottimizzata per:

- rispettare il range di emissione lineare dei fluorofori (400->800 PMT)
- bilanciare la diversa emissività dei due fluorocromi
- sfruttare al meglio l'intervallo di campionamento
- migliorare il rapporto segnale/rumore (SNR)

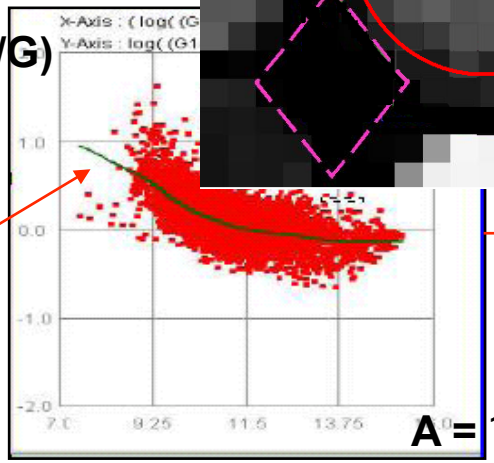
Pre-processamento dei dati grezzi

- Segmentazione
- Estrazione delle intensità di segnale e di
- Correzione del background
- Applicazione di indi t per la ezione dei geni giudicati idonei pe zzazione SNR
- Normalizzazione



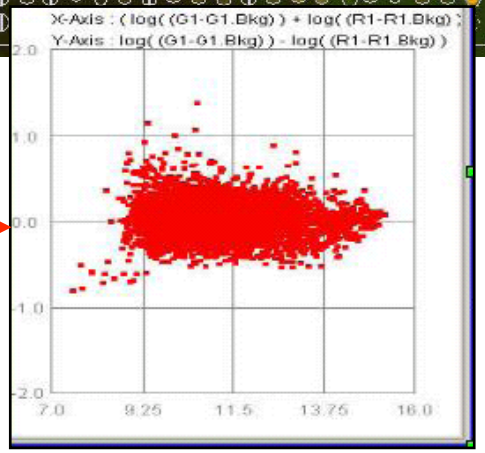


M = log (R/G)



Funzione di smoothing

A = ½ log (R*G)



Analisi statistica

SNP

- Chi-quadro
- Modelli di regressione e valutazione degli Odd Ratio

	Casi	Controlli	
Genotipo AA			Tot AA
Genotipo Aa			Tot Aa
Genotipo aa			Tot aa
	Tot Casi	Tot Controlli	

Gene expression

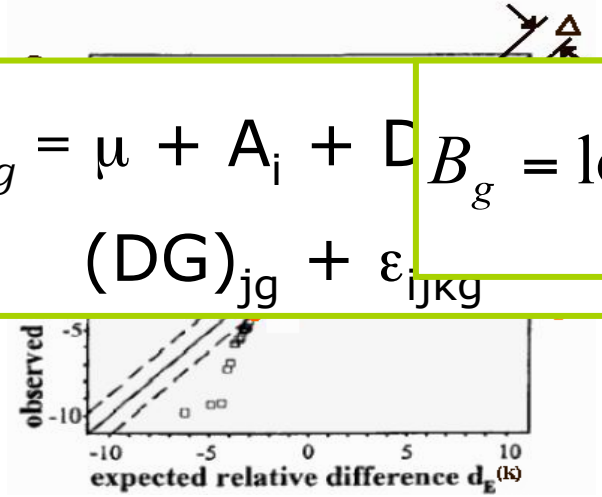
- Metodi statistici
  - t-test, ANOVA, B-statistic

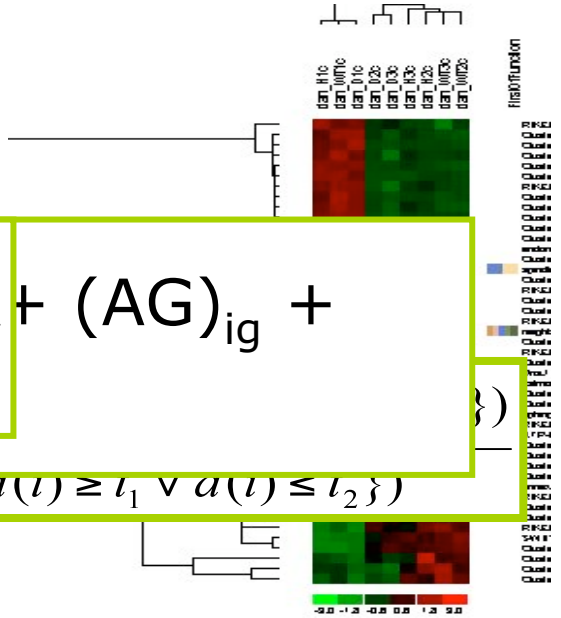
$$y_{ijkg} = \mu + A_i + D$$

$$(DG)_{jg} + \varepsilon_{ijkg}$$

$$B_g = \log \frac{\Pr(I_g = 1 \mid M_{ij})}{\Pr(I_g = 0 \mid M_{ij})}$$

$$+ (AG)_{ig} +$$





Validazione ed Interpretazione del dato

- Validazione:

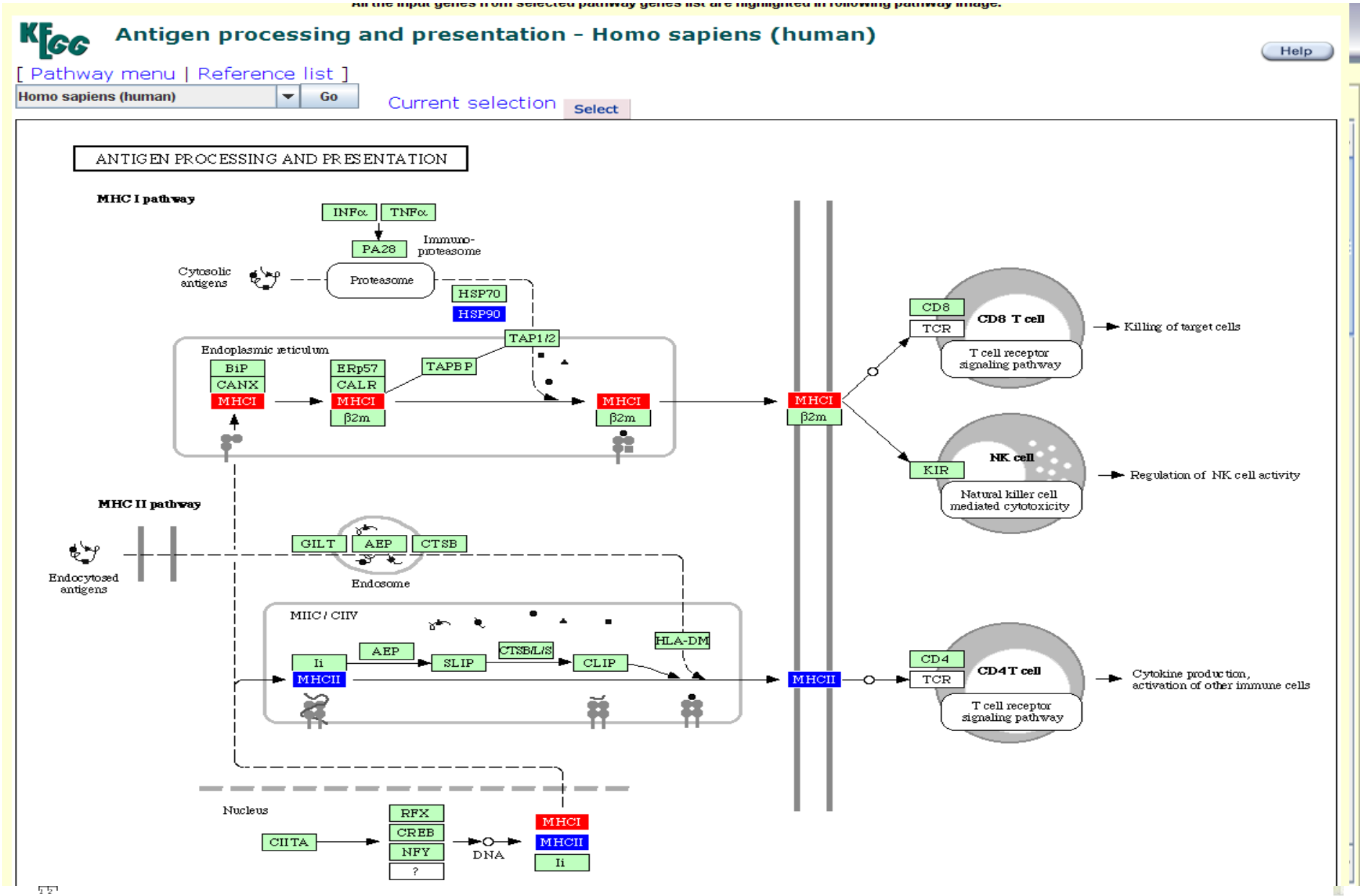
  - valutazione attraverso metodiche alternative (real time RT-PCR)del livello di espressione in un sottoinsieme di geni differenzialmente espressi

- Interpretazione:

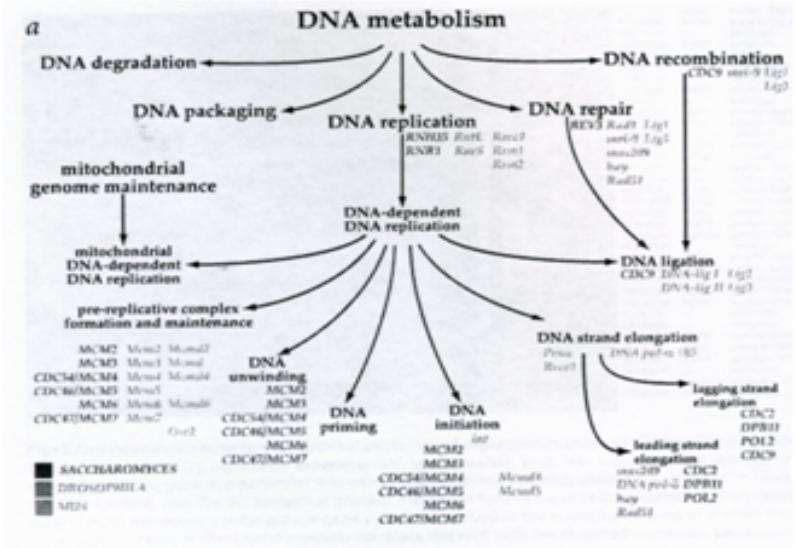
  - analisi della lista dei geni DE per individuare l'effetto a livello molecolare del fenomeno biologico indagato
    - informazioni sui singoli geni
    - reti biochimiche (pathway) di trasmissione del segnale

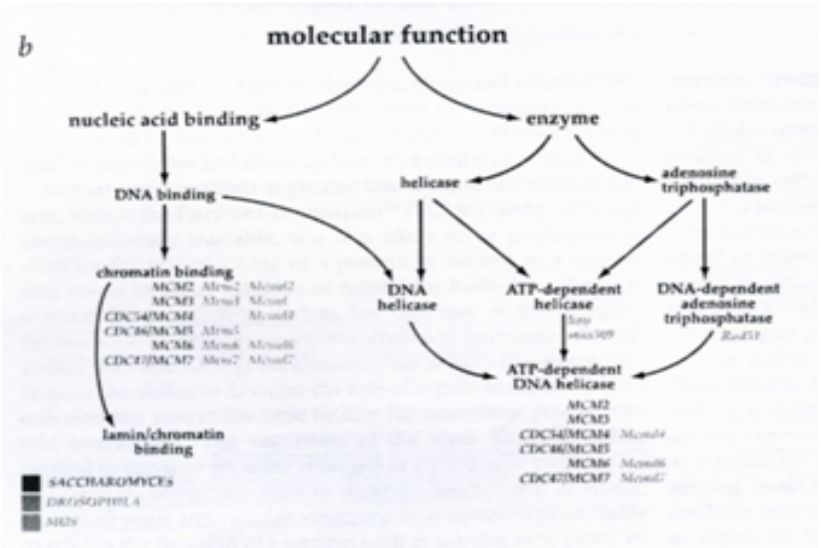
  - formulazione di un'ipotesi sugli SNP che risultano associati al fenotipo

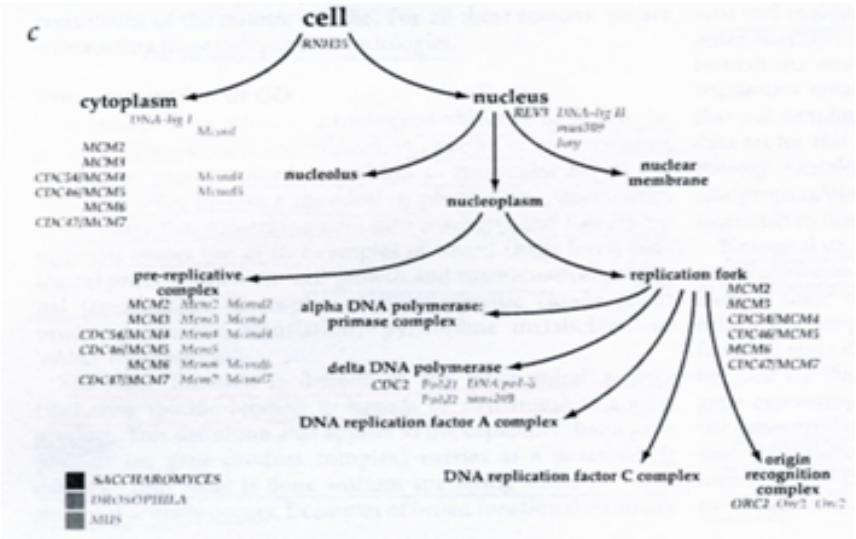
Banche dati


Antigen processing and presentation - Homo sapiens (human)
Help
[Pathway menu | Reference list]
Homo sapiens (human)
Go
Current selection
Select
ANTIGEN PROCESSING AND PRESENTATION
MHC I pathway
INFα
TNFα
PA28
Immuno-proteasome
Cytosolic antigens
Proteasome
HSP70
HSP90
TAP1/2
Endoplasmic reticulum
BiP
CANX
MHCI
ERp57
CALR
MHCI
β2m
TAPBP
MHCI
β2m
MHCI
β2m
CD8
TCR
CD8 T cell
T cell receptor signaling pathway
Killing of target cells
KIR
NK cell
Natural killer cell mediated cytotoxicity
Regulation of NK cell activity
MHC II pathway
Endocytosed antigens
GILT
AEP
CTSB
Endosome
MIIC / CIIV
Ii
MHCII
AEP
SLIP
CTSB/L/S
CLIP
HLA-DM
MHCII
CD4
TCR
CD4 T cell
T cell receptor signaling pathway
Cytokine production, activation of other immune cells
Nucleus
CIITA
RFX
CREB
NFY
?
DNA
MHCI
MHCII
Ii

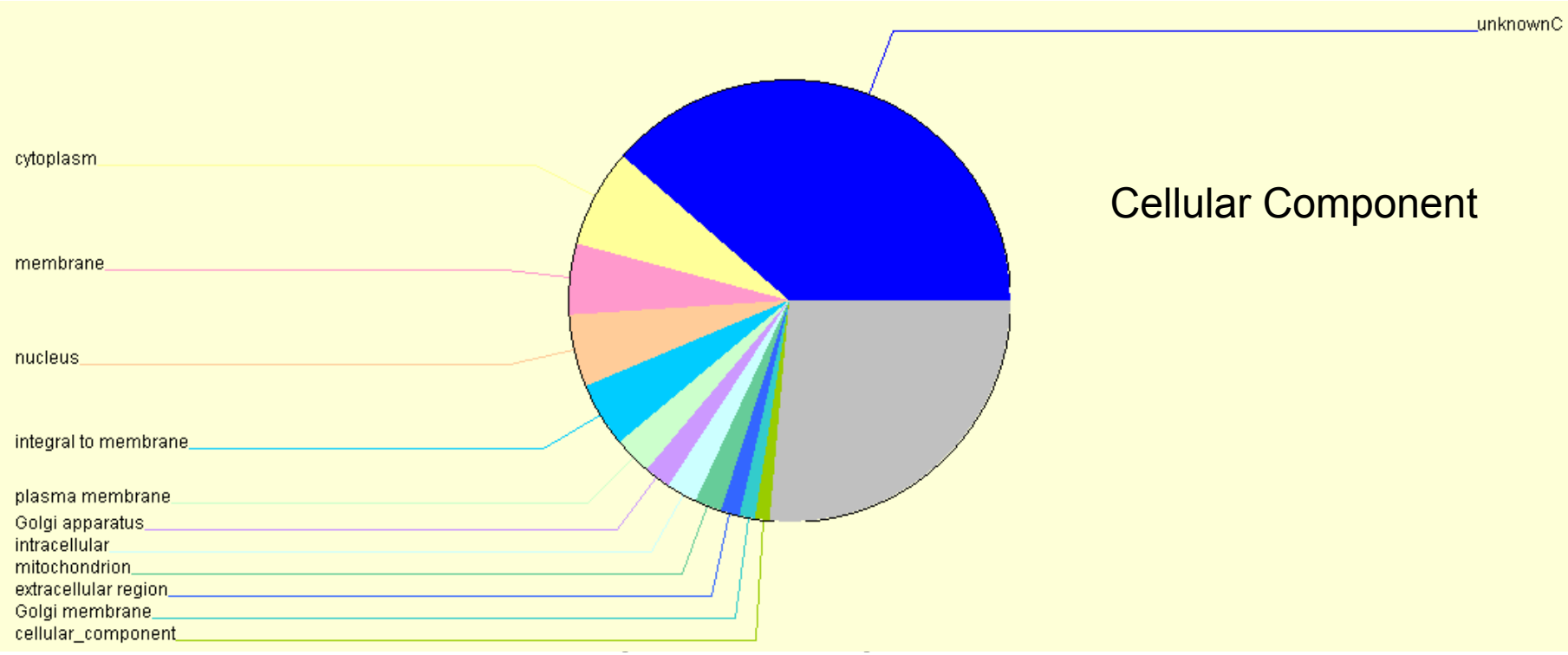
Banche dati


a
DNA metabolism
DNA degradation
DNA recombination
DNA packaging
DNA replication
DNA repair
mitochondrial genome maintenance
DNA-dependent DNA replication
mitochondrial DNA-dependent DNA replication
DNA ligation
pre-replicative complex formation and maintenance
DNA strand elongation
DNA unwinding
DNA priming
DNA initiation
lagging strand elongation
leading strand elongation
SACCHAROMYCES


b
molecular function
nucleic acid binding
enzyme
DNA binding
helicase
adenosine triphosphatase
chromatin binding
DNA helicase
ATP-dependent helicase
DNA-dependent adenosine triphosphatase
ATP-dependent DNA helicase
lamin/chromatin binding
SACCHAROMYCES


c
cell
cytoplasm
nucleus
nucleolus
nucleoplasm
nuclear membrane
pre-replicative complex
replication fork
alpha DNA polymerase: primase complex
delta DNA polymerase
DNA replication factor A complex
DNA replication factor C complex
origin recognition complex
SACCHAROMYCES

Banche dati



- funzione molecolare
- processo biologico
- componente cellulare

http://vortex.cs.wayne.edu/projects.htm

- SW per la caratterizzazione ontologica in dataset

Software per l'interpretazione dei dati

- http://www.genecards.org/index.shtml
- http://www.ihop-net.org/UniPub/iHOP/
- http://www.pubgene.org/

Identificazione di geni candidati per studi di genotipizzazione

Bioinformatics 2008 24(13):i277-i285; doi:10.1093/bioinformatics/btn182

Identifying gene-disease associations using centrality on a literature mined gene-interaction network

Arzucan Özgür [1], Thuy Vu [1], Güneş Erkan [1] and Dragomir R. Radev [1,2,*]

[1]Electrical Engineering and Computer Science and [2]School of Information, University of Michigan, Ann Arbor, MI 48109, USA

Abstract

Motivation: Understanding the role of genetics in diseases is one of the most important aims of the biological sciences. The completion of the Human Genome Project has led to a rapid increase in the number of publications in this area. However, the coverage of curated databases that provide information manually extracted from the literature is limited. Another challenge is that determining disease-related genes requires laborious experiments. Therefore, predicting good candidate genes before experimental analysis will save time and effort. We introduce an automatic approach based on text mining and network analysis to predict gene-disease associations. We collected an initial set of known disease-related genes and built an interaction network by automatic literature mining based on dependency parsing and support vector machines. Our hypothesis is that the central genes in this disease-specific network are likely to be related to the disease. We used the degree, eigenvector, betweenness and closeness centrality metrics to rank the genes in the network.
Results: The proposed approach can be used to extract known and to infer unknown gene-disease associations. We evaluated the approach for prostate cancer. Eigenvector and degree centrality achieved high accuracy. A total of 95% of the top 20 genes ranked by these methods are confirmed to be related to prostate cancer. On the other hand, betweenness and closeness centrality predicted more genes whose relation to the disease is currently unknown and are candidates for experimental study.
Availability: A web-based system for browsing the disease-specific gene-interaction networks is available at: http://gin.ncibi.org