Bioinformatica e Biologia Computazionale per la Medicina Molecolare

Marco Masseroli, PhD  marco.masseroli@polimi.it
Integration and Analysis of Genomic Annotations: the DAVID and GFINDer Systems

Marco Masseroli, PhD
marco.masseroli@polimi.it
The DAVID and GFINDer Systems

Index

- Introduction and motivation
- DAVID (http://david.abcc.ncifcrf.gov/)
  - Functional annotation
  - Functional annotation chart
  - Functional annotation clustering
  - Gene functional classification
- GFINDer (http://www.bioinformatics.polimi.it/GFINDer/)
  - GFINDer modules
    - Uploading module
    - Annotation module
    - Exploration modules
    - Statistical analysis modules
- GFINDer in action
  - The Parkinson’s Disease case
- Integrated data quality checking
- Conclusions
- More on GFINDer …
The DAVID and GFINDer Systems

Introduction and motivation

• Many tasks in bioinformatics require comprehensive evaluation of many different types of data (e.g., structural, functional, phenotypic, …)
  - E.g. to identify the biomolecular phenomena involved in the differential expression of a gene set in a specific biological condition
• Such data are generally available in numerous, distributed and heterogeneous data sources
The DAVID and GFINDER Systems

Introduction and motivation

- In January 2010 more than 1,200 publicly accessible databanks
- **Increasing coverage** of both:
  - **Biomolecular entities** (genomic DNAs, genes, transcripts, proteins)
  - **Description of their structural and functional biomedical features** (sequences, expression in different tissues, involvement in biological processes and genetic disorders)
The DAVID and GFINDER Systems

Introduction and motivation

- Information about a given biomolecular entity is often scattered across many different databanks.
- Combining information from multiple databanks is paramount for biomolecular investigation.

- Several approaches have been proposed to integrate data from multiple sources, including:
  - data warehousing
  - multi-databases
  - federated databases
  - information linkage
  - mediator based systems
The DAVID and GFINDER Systems

Introduction and motivation

• Data warehousing is the most adequate when it is required:
  ▪ integration of numerous data
  ▪ efficient off-line data processing
  ▪ comprehensive mining of the integrated data

• It requires that information from the distributed databanks to be integrated are automatically retrieved and processed to create and maintain updated an integrated and consistent collection of originally distributed data

• Considering this scenario, a few tools have been developed to support biomedical interpretation of high-throughput gene lists
Main publicly available Web-tools aiding interpretation of biological meaning of high-throughput experimental results by:

- Complementing gene ID lists with controlled annotations
- Statistical analysis of relevance integrated annotations

**DAVID** - Database for Annotation, Visualization, and Integrated Discovery (http://david.abcc.ncifcrf.gov/)

- National Institute of Allergy and Infectious Diseases (NIAID), National Institute of Health (NIH) - US

**GFINDer** - Genome Function INtegrated Discoverer (http://www.bioinformatics.polimi.it/GFINDER/) and its Genomic and Proteomic Data Warehouse (GPDW)

- Politecnico di Milano

Both implement ICT to support biomolecular investigation with valuable spin-off for molecular medicine and health treatment
After grand opening in 2003, new enhanced version:

- DAVID Bioinformatics Resources 2007:
  - Data Warehouse
    - Numerous gene concept centered annotations
  - Functional Annotation
    - High number of controlled annotation integrated
  - Gene Functional Classification
    - Fuzzy algorithm for annotation category clustering
  - Gene Accession Conversion
    - Conversion of gene IDs also across species
  - Gene Name Batch Viewer
The DAVID and GFINDER Systems

DAVID 2007 - Functional annotation

Pay attention on which gene list, species and population background that the tool is being applied.

1. Upload List Background
   - Select to limit annotations by one or more species
   - Use All Species - Homo sapiens (171)
   - Select List:
     - Use
     - Rename
     - Remove
     - Combine
   - Show Gene List

2. View and select annotation categories of your interests. (7 of them is pre-selected as default)

3. Individual views/reports:
   - Percentage, e.g. 7/171 (involved genes / total genes)
   - Genes from your list involved in this annotation category
   - Single Chart Report ONLY for this annotation categories

4. Combined views/reports:
   - Clustered or non-redundant chart report of annotation terms for ALL selected annotation categories above
   - Linear or redundant chart report of annotation terms for ALL selected annotation categories above
   - Table report for ALL selected annotation categories.
The DAVID and GFINDeR Systems

**DAVID 2007 - Functional annotation chart**

- **Gene list and population background being analyzed**
- **Minimum number of genes for the corresponding term**
- **Maximum EASE Score/P-Value**
- **Maximum number of record per page**

### Options
- **Count Threshold**: 2
- **EASE Threshold**: 0.1
- **# of Records Displayed**: 1000

<table>
<thead>
<tr>
<th>Sublist</th>
<th>Category</th>
<th>Term</th>
<th>Count</th>
<th>%</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>SP_PIR_KEYWORDS</td>
<td>signal</td>
<td></td>
<td>47</td>
<td>27.5%</td>
<td>3.0E-10</td>
</tr>
<tr>
<td>SP_PIR_KEYWORDS</td>
<td>glycoprotein</td>
<td></td>
<td>51</td>
<td>29.8%</td>
<td>4.9E-8</td>
</tr>
<tr>
<td>GOTERM_CC_ALL</td>
<td>extracellular region</td>
<td></td>
<td>32</td>
<td>18.7%</td>
<td>1.1E-7</td>
</tr>
<tr>
<td>SP_PIR_KEYWORDS</td>
<td>alternative splicing</td>
<td></td>
<td>49</td>
<td>28.7%</td>
<td>6.4E-6</td>
</tr>
<tr>
<td>SP_PIR_KEYWORDS</td>
<td>chromoprotein</td>
<td></td>
<td>7</td>
<td>4.1%</td>
<td>1.1E-5</td>
</tr>
<tr>
<td>SP_PIR_KEYWORDS</td>
<td>direct protein sequencing</td>
<td></td>
<td>33</td>
<td>19.3%</td>
<td>1.2E-5</td>
</tr>
<tr>
<td>SP_PIR_KEYWORDS</td>
<td>phosphorylation</td>
<td></td>
<td>31</td>
<td>18.1%</td>
<td>1.6E-5</td>
</tr>
<tr>
<td>UP_SEQFEATURE</td>
<td>signal peptide</td>
<td></td>
<td>47</td>
<td>27.5%</td>
<td>3.7E-5</td>
</tr>
<tr>
<td>SP_PIR_KEYWORDS</td>
<td>metalloprotein</td>
<td></td>
<td>8</td>
<td>4.7%</td>
<td>4.7E-5</td>
</tr>
<tr>
<td>GOTERM_BP_ALL</td>
<td>response to chemical stimulus</td>
<td></td>
<td>14</td>
<td>8.2%</td>
<td>6.1E-5</td>
</tr>
</tbody>
</table>

- **Original database/resource where the terms orient**
- **Enriched terms associated with your gene list**
- **Related Term Search**
- **Genes involved in the term**
- **Percentage, e.g. 14/171 = 8.2% (involved genes/total genes)**
- **Modified Fisher Exact P-Value, EASE Score. The smaller, the more enriched.**
The DAVID and GFINDER Systems

DAVID 2007 - *Functional annotation clustering*

**Functional Annotation Clustering**

- **Current Gene List:** demoList
- **171 DAVID IDs**

**Options**
- Run using options
- Create Sublist

**Classification Stringency:** High

<table>
<thead>
<tr>
<th>Annotation Cluster 1</th>
<th>Enrichment Score: 3.60</th>
</tr>
</thead>
<tbody>
<tr>
<td>SP_PRIR_KEYWORDS</td>
<td>chromoprotein</td>
</tr>
<tr>
<td>SP_PRIR_KEYWORDS</td>
<td>metalloprotein</td>
</tr>
<tr>
<td>SP_PRIR_KEYWORDS</td>
<td>iron</td>
</tr>
<tr>
<td>GOTERM_MF_ALL</td>
<td>iron ion binding</td>
</tr>
<tr>
<td>SP_PRIR_KEYWORDS</td>
<td>heme</td>
</tr>
<tr>
<td>SP_PRIR_KEYWORDS</td>
<td>tetrapyrrole binding</td>
</tr>
<tr>
<td>GOTERM_MF_ALL</td>
<td>heme binding</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Annotation Cluster 2</th>
<th>Enrichment Score: 3.52</th>
</tr>
</thead>
<tbody>
<tr>
<td>SP_PRIR_KEYWORDS</td>
<td>antibiotic</td>
</tr>
<tr>
<td>SP_PRIR_KEYWORDS</td>
<td>antimicrobial</td>
</tr>
<tr>
<td>UP_SEQ_FEATURE</td>
<td>defense response to bacteria</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Annotation Cluster 3</th>
<th>Enrichment Score: 2.66</th>
</tr>
</thead>
<tbody>
<tr>
<td>UP_SEQ_FEATURE</td>
<td>domain: Ig-like C2-type 1</td>
</tr>
<tr>
<td>INTERPRO_NAME</td>
<td>Immunoglobulin</td>
</tr>
</tbody>
</table>

**EASE Score:**
- The overall enrichment score for the group based on the EASE scores of each term members. The higher, the more enriched.

**Related Term Search**
- Every term in the annotation cluster

**All genes involved in this annotation cluster**
- Genes involved in individual term

**A group of terms having similar biological meaning due to sharing similar gene members**

EASE Score, the modified Fisher Exact P-Value. They are identical to that in the Chart Report. The smaller, the more enriched.
The DAVID and GFINDER Systems

DAVID 2007 - Gene functional classification

Gene Functional Classification
Current Gene List: null
Current Background: Homo sapiens
15 DAVID IDs

Options: Classification Stringency: Medium
- Renrun using options
- Create Sublist
- Heatmap

3 Cluster(s)

Gene Group 1
Enrichment Score: 4.45
- 34375_AT, 875_G_AT
- chemokine (c-c motif) ligand 2
- 40385_AT
- chemokine (c-c motif) ligand 20
- 36103_AT
- chemokine (c-c motif) ligand 3
- 36674_AT
- chemokine (c-c motif) ligand 4
- 408_AT
- interleukin 8

Gene Group 2
Enrichment Score: 2.16
- 41446_F_AT
- metallothionein 1f (functional)
- 41446_F_AT
- metallothionein 1g
- 39232_AT
- lim and senescent cell antigen-like domains 1
- 41446_F_AT
- chromosome 20 open reading frame 127

Gene Group 3
Enrichment Score: 1.52
- 39748_AT
- solute carrier family 7 (cationic amino acid transporter, v+ system), member 1
- 32186_AT
- solute carrier family 7 (cationic amino acid transporter, v+ system), member 5
- 33143_S_AT
- solute carrier family 16 (monocarboxylic acid transporters), member 3
- 39841_AT
- solute carrier family 16 (monocarboxylic acid transporters), member 6

Save the output in your local disk.

What are the key biology for this gene group?

How do the gene members share common annotations/biology?

Are there any other genes in my gene list or in the genome functionally similar to this gene group?

1 genes from your list are not in the output.
Online interpretation of high-throughput experimental results (e.g. gene expression microarrays) by the scientific community

GFINDer biomolecular data warehouse:

- Data on 14 organisms:
  - 830,261 nucleotide sequences
  - 336,068 genes (24,682 human)
  - 1,125,161 proteins (312,161 human)
  - 3,149 protein domains
  - 1,269 biochemical pathways
  - 2,465 inherited disorders
- 8.5 GB of MySQL DB

Gene Ontology

- 24,493 concepts; 40,353 associations
Nearly 170,000 accesses from more than 9,300 IPs since 2004 (about 13,700 accesses from nearly 1,500 IPs in the last year)
The DAVID and GFINDer Systems
GFINDer modules - Uploading module

Upload Sequence IDs

Upload a list of sequence IDs in the **required format** by selecting the text file containing the list or by pasting it in the uploading area below.

- Check this box if the sequence IDs to upload refer to a previously uploaded background list. (For sequence IDs identified via microarray, their background list must include all IDs of the sequences on the microarray.)

If in the list at least a column with sequence ID’s membership classes are included, please specify column delimiter:

- TABULATION ( )
- COMMA (,)
- SEMI-COLON (;)

**File to upload:**

![File upload interface]

**Upload File**

<table>
<thead>
<tr>
<th>ID</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>U38073</td>
<td>ALL</td>
</tr>
<tr>
<td>U26266</td>
<td>ALL</td>
</tr>
<tr>
<td>K25303</td>
<td>ALL</td>
</tr>
<tr>
<td>T02612</td>
<td>ALL</td>
</tr>
<tr>
<td>U38451</td>
<td>ALL</td>
</tr>
<tr>
<td>K29566</td>
<td>ALL</td>
</tr>
<tr>
<td>K13792</td>
<td>ALL</td>
</tr>
<tr>
<td>M05700</td>
<td>ALL</td>
</tr>
<tr>
<td>M05194</td>
<td>ALL</td>
</tr>
<tr>
<td>M02759</td>
<td>ALL</td>
</tr>
</tbody>
</table>

**This is a classified sequence ID’s sample file**

Upload One of the following background lists:  
- Probe IDs from Affymetrix C. elegans

**Upload this list**
### The DAVID and GFINDer Systems

**GFINDer modules - Annotation module**

<table>
<thead>
<tr>
<th>Sequence ID</th>
<th>Type</th>
<th>Gene Name</th>
<th>MIM ID</th>
<th>Disease Category</th>
<th>Inheritance Mode</th>
<th>Phenotype ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>NIDDM</td>
<td>adenosine deaminase</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10076</td>
<td>NIDDM</td>
<td>protein tyrosine phosphatase, receptor type,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1009</td>
<td>NIDDM</td>
<td>cadherin 11, type 2, CB-cadherin (osteoblast)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10555</td>
<td>NIDDM</td>
<td>1-acylglycerol-3-phosphate O-acyltransferase</td>
<td>603100</td>
<td>Lipodystrophy, congenital generalized, type 1</td>
<td>Autosomal</td>
<td>608594</td>
</tr>
<tr>
<td>1071</td>
<td>NIDDM</td>
<td>cholesteryl ester transfer protein, plasma</td>
<td>118470</td>
<td>CETP deficiency</td>
<td>Autosomal</td>
<td>607322</td>
</tr>
<tr>
<td>1071</td>
<td>NIDDM</td>
<td>cholesteryl ester transfer protein, plasma</td>
<td>118470</td>
<td>Hyperalphaproteinemia</td>
<td>Autosomal</td>
<td>143470</td>
</tr>
<tr>
<td>1071</td>
<td>NIDDM</td>
<td>cholesteryl ester transfer protein, plasma</td>
<td>118470</td>
<td>Longevity, exceptional</td>
<td>Autosomal</td>
<td>152430</td>
</tr>
<tr>
<td>1234</td>
<td>IDDM</td>
<td>chemokine (C-C motif) receptor 5</td>
<td>601373</td>
<td>HIV infection, susceptibility/resistance to</td>
<td>Autosomal</td>
<td></td>
</tr>
<tr>
<td>1270</td>
<td>IDDM</td>
<td>ciliary neurotrophic factor</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1803</td>
<td>NIDDM</td>
<td>dipeptidylpeptidase 4 (CD26, adenosine deamin</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1934</td>
<td>NIDDM</td>
<td>eukaryotic translation elongation factor 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>207</td>
<td>NIDDM</td>
<td>v-akt murine thymoma viral oncogene homolog 1</td>
<td>164730</td>
<td>Schizophrenia, susceptibility to</td>
<td>Autosomal</td>
<td>181500</td>
</tr>
<tr>
<td>2169</td>
<td>NIDDM</td>
<td>fatty acid binding protein 2, intestinal</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2260</td>
<td>IDDM</td>
<td>fibroblast growth factor receptor 1 (fms-rela)</td>
<td>136350</td>
<td>Jackson-Weiss syndrome</td>
<td>Autosomal</td>
<td>123150</td>
</tr>
<tr>
<td>2260</td>
<td>IDDM</td>
<td>fibroblast growth factor receptor 1 (fms-rela)</td>
<td>136350</td>
<td>Kallmann syndrome 2</td>
<td>Autosomal</td>
<td>147950</td>
</tr>
<tr>
<td>2260</td>
<td>IDDM</td>
<td>fibroblast growth factor receptor 1 (fms-rela)</td>
<td>136350</td>
<td>Pfeiffer syndrome</td>
<td>Autosomal</td>
<td>101600</td>
</tr>
</tbody>
</table>
The DAVID and GFINDer Systems

**GFINDer modules** - Exploration modules

---

### GFINDer: Genome Function INtegrated Discoverer

- **Explore Protein Families & Domains**
- **Select sequence ID's list:** 1813: Cardio-Neuro
- **Source database:** Pfam
- **Column:** Value
- **NEURO CARDIO All**

---

### Table: Protein Families & Domains

<table>
<thead>
<tr>
<th>Type</th>
<th>Tree</th>
<th>Level</th>
<th>Category Name</th>
<th>Num. (%)</th>
<th>Histogram</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>29</td>
<td>0</td>
<td>Protein kinase-like</td>
<td>71 (10.3%)</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>29</td>
<td>1</td>
<td>Protein kinase</td>
<td>65 (9.4%)</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>29</td>
<td>2</td>
<td>EGF-like region</td>
<td>45 (6.5%)</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>69</td>
<td>0</td>
<td>EF-Hand type</td>
<td>37 (5.4%)</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>69</td>
<td>1</td>
<td>Serine/threonine protein kinase</td>
<td>37 (5.4%)</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>69</td>
<td>2</td>
<td>Calcium-binding EF-hand</td>
<td>36 (5.2%)</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>271</td>
<td>0</td>
<td>Immunoglobulin-like</td>
<td>28 (4.1%)</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>5</td>
<td>0</td>
<td>Tyrosine specific protein phosphatase and dual specificity protein phosphatase</td>
<td>28 (4.1%)</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>86</td>
<td>0</td>
<td>Fibronectin, type III</td>
<td>25 (3.6%)</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>53</td>
<td>0</td>
<td>EGF-like, type 3</td>
<td>25 (3.6%)</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>20</td>
<td>0</td>
<td>Steroid nuclear receptor, ligand-binding</td>
<td>25 (3.6%)</td>
<td></td>
</tr>
</tbody>
</table>

---

© Marco Masseroli, PhD
GFINDer uses controlled annotations to perform statistical functional and phenotypic evaluations of user classified gene lists.

User classification could come from:

- Microarray gene expression results
- Any classification method (e.g. different experimental conditions, different times, different patients, gene clustering, ...)

Considered annotations include: Gene Ontology categories, KEGG biochemical pathways, InterPro and Pfam protein families and domains, OMIM genetic disorders and phenotypes, eVOC expression ontology (anatomical system, cellular type, developmental stage, and pathology) categories.
Gene occurrences in controlled annotation categories are counted to evaluate the significance of the functional categories in a user classified gene list.

In the user gene list the classes can represent gene regulation or any gene classification.

Each gene can be considered as belonging to:
- A specific class (e.g. Class 1) or not
- A specific annotation category (e.g. Category A) or not
Gene counts are summarized in a 2x2 contingency table and Hypergeometric statistical test for equality of proportions or Fisher’s Exact test are applied.

\[
\begin{array}{ccc}
\text{Category A} & \text{Other Categories} & \text{Total} \\
\hline
\text{Class 1} & n_{11} & n_{12} & N_1 \\
\text{Other Classes} & n_{21} & n_{22} & N_2 \\
\hline
\text{Total} & N_{.1} & N_{.2} & N_{..}
\end{array}
\]

\(n_{ij} (i = 1, 2; j = 1, 2)\): number of genes of Category A, or Other Categories, for Class 1, or Other Classes

\(N_{i.}, N_{.j}\): the marginal total of number of genes for the table row or column

\(N_{..}\): the total number of considered genes (e.g. the number of genes on the microarray used)
The DAVID and GFINDer Systems

GFINDer modules - Statistical analysis modules

<table>
<thead>
<tr>
<th>Phenotype level</th>
<th>Phenotype category name</th>
<th>$p$-value $&lt;$ 0.05</th>
<th>Log(1/P) $&lt;$ 0.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bradykinesia [O:13, E:6.91, R:1.88] $T$</td>
<td>$p_{H}=0.00002$</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Parkinsonism [O:13, E:6.91, R:1.88] $T$</td>
<td>$p_{H}=0.00002$</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Rigidity [O:13, E:6.91, R:1.88] $T$</td>
<td>$p_{H}=0.00002$</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Autosomal dominant [O:12, E:6.38, R:1.88] $T$</td>
<td>$p_{H}=0.00005$</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Dementia [O:12, E:6.38, R:1.88] $T$</td>
<td>$p_{H}=0.00005$</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Neurofibrillary tangles composed of disordered microtubules [O:0, E:5.84, R:0] $T$</td>
<td>$p_{H}=0.00001$</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Dementia presenile and senile [O:0, E:6.38, R:0] $T$</td>
<td>$p_{H}=&lt;0.00001$</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Autosomal dominant [O:0, E:6.91, R:0] $T$</td>
<td>$p_{H}=&lt;0.00001$</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Long tract signs [O:0, E:6.91, R:0] $T$</td>
<td>$p_{H}=&lt;0.00001$</td>
<td></td>
</tr>
</tbody>
</table>
The DAVID and GFINDer Systems

GFINDer modules - Statistical analysis modules

Parkinson’s vs. Alzheimer’s disease related genes

© Marco Masseroli, PhD
GFIND*er in action - The Parkinson’s Disease case

- GFIND*er enriches experimental data with functional and phenotypic information from available ontologies
  - Experimentally selected genes candidate correlated with Parkinson’s diseases
  - Highlighting semantic ontological phenotypic features of selected genes
The DAVID and GFINDer Systems

**GFINDer in action - The Parkinson’s Disease case**

- GFINDer enriches experimental data with functional and phenotypic information from available ontologies
  - Highlighting semantic ontological functional features of selected genes
  - Inferring new associations

<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>Gene ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>APLP2</td>
<td>334</td>
</tr>
<tr>
<td>CYP2D6</td>
<td>1566</td>
</tr>
<tr>
<td>DBH</td>
<td>1621</td>
</tr>
<tr>
<td>EN2</td>
<td>2020</td>
</tr>
<tr>
<td>IREB2</td>
<td>3658</td>
</tr>
<tr>
<td>MAPT</td>
<td>4137</td>
</tr>
<tr>
<td>NDUFS2</td>
<td>4729</td>
</tr>
<tr>
<td>NGFB</td>
<td>4803</td>
</tr>
<tr>
<td>NR4A2</td>
<td>4929</td>
</tr>
<tr>
<td>PARK2</td>
<td>5071</td>
</tr>
<tr>
<td>PARK3</td>
<td>6072</td>
</tr>
<tr>
<td>SLC18A2</td>
<td>6571</td>
</tr>
<tr>
<td>SNCA</td>
<td>6622</td>
</tr>
<tr>
<td>TBP</td>
<td>6908</td>
</tr>
<tr>
<td>UCHL1</td>
<td>7345</td>
</tr>
<tr>
<td>SNCA1P</td>
<td>9627</td>
</tr>
<tr>
<td>PARK7</td>
<td>11315</td>
</tr>
<tr>
<td>PARK4</td>
<td>60464</td>
</tr>
<tr>
<td>PINK1</td>
<td>65018</td>
</tr>
<tr>
<td>LRRK2</td>
<td>120892</td>
</tr>
<tr>
<td>PARK9</td>
<td>140600</td>
</tr>
<tr>
<td>PARK10</td>
<td>170534</td>
</tr>
<tr>
<td>PARK11</td>
<td>352908</td>
</tr>
</tbody>
</table>
The DAVID and GFINDer Systems

GFINDer in action - The Parkinson’s Disease case

- GFINDer enriches experimental data with functional and phenotypic information from available ontologies
  - Highlighting semantic ontological functional features of selected genes
  - Inferring new associations

<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>Gene ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>APLP2</td>
<td>334</td>
</tr>
<tr>
<td>CYP2D6</td>
<td>1565</td>
</tr>
<tr>
<td>DBH</td>
<td>1621</td>
</tr>
<tr>
<td>EN12</td>
<td>2020</td>
</tr>
<tr>
<td>IREB2</td>
<td>3658</td>
</tr>
<tr>
<td>NAPRT</td>
<td>4137</td>
</tr>
<tr>
<td>NDUFV2</td>
<td>4729</td>
</tr>
<tr>
<td>NGFB</td>
<td>4803</td>
</tr>
<tr>
<td>NR4A2</td>
<td>4929</td>
</tr>
<tr>
<td>PARK2</td>
<td>5071</td>
</tr>
<tr>
<td>PARK3</td>
<td>5072</td>
</tr>
<tr>
<td>SLC18A2</td>
<td>6571</td>
</tr>
<tr>
<td>SNCA</td>
<td>6622</td>
</tr>
<tr>
<td>TBP</td>
<td>6908</td>
</tr>
<tr>
<td>UCHL1</td>
<td>7345</td>
</tr>
<tr>
<td>SNCAIP</td>
<td>9627</td>
</tr>
<tr>
<td>PARK7</td>
<td>11315</td>
</tr>
<tr>
<td>PARK4</td>
<td>60454</td>
</tr>
<tr>
<td>PINK1</td>
<td>65018</td>
</tr>
<tr>
<td>LRRK2</td>
<td>120892</td>
</tr>
<tr>
<td>PARK9</td>
<td>140600</td>
</tr>
<tr>
<td>PARK10</td>
<td>170534</td>
</tr>
<tr>
<td>PARK11</td>
<td>352906</td>
</tr>
</tbody>
</table>
GFINDER in Action
The Parkinson's Disease Case (2)
The DAVID and GFINDER Systems

GFINDER in action - The Parkinson's Disease case

- EN2
  - Gene Ontology
  - biological process
  - regulation of biological process
  - development
  - cellular process

- NGFB
  - regulation of cellular physiological process
  - establishment of localization
  - cell communication
  - cell-cell signaling
  - signal transduction

- SLC18A2
  - regulation of metabolism
  - regulation of cellular metabolism
  - establishment of localization
  - transport
  - GMP transport
  - neuroreceptor transport
  - monoamine transport
  - dopamine transport

- APLP2
  - cell surface receptor
  - linked signal transduction
  - G-protein coupled receptor
  - protein signaling pathway
  - dopamine receptor signaling pathway
The DAVID and GFINDer Systems

GFINDer - Integrated data quality checking

• In integrated data, GFINDer automatic processing identifies:
  ▪ Data structure differences in new updates
  ▪ Inconsistencies among data from different or the same databank

• This is ensured in each updating of source data files by:
  ▪ Strict checking of data parsed from source data files
  ▪ Checking absence of null data and data structure modifications
  ▪ Use of regular expressions to identify and check IDs
    - Grants correct use of alias and historic or obsolete IDs
  ▪ Checking and cross-validation of data imported from different sources to identify redundant and mismatching information
The DAVID and GFINDer Systems

GFINDer - Integrated data quality checking

- Analysis of overlaps and relationship loops among integrated data:
  - Can help in verifying data consistency and unveiling unexpected information pattern
  - Can possibly lead to biological discoveries
- E.g. by checking cross-references between Gene Ontology, Entrez Gene and UniProt databanks, we tested consistency of GO annotations of proteins and their codifying genes
  - Found 321 (1.83%) GO annotations (about 197 different GO terms) of 81 human proteins not comprised in the GO annotations of their codifying genes from Entrez Gene
  - Including 143 (44.56%) annotations with evidence stronger than inferred from electronic annotation (IEA)
The DAVID and GFINDer Systems

Conclusions

• Computational systems and data warehouses, such as DAVID and GFINDer, provide support for:
  ▪ Comprehensive use and analysis of sparsely available genomic structural, functional and phenotypic data
  ▪ Answering biological questions requiring integrated access to many biomolecular information and knowledge

• Data quality controls of the integrated data can reveal inconsistencies or missing information in public databanks
  ▪ Reporting them to the databank curators can improve the quality of data available to the scientific community
  ▪ Allowing their correct use in support of high-throughput data driven biological discoveries
http://www.bioinformatics.polimi.it/GFINDer/


