

Functional profiling from *Babelomics 5*

Dan Crespo

dcrespo@cipf.es

September 29th, 2016



GDA

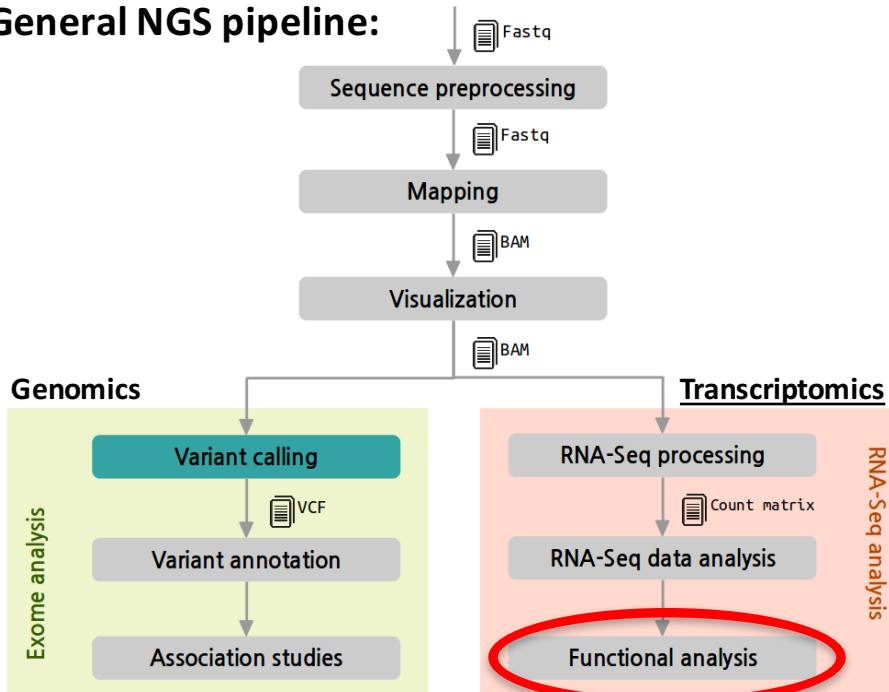
International Course on
Genomic Data Analysis



PRINCIPE FELIPE
CENTRO DE INVESTIGACION

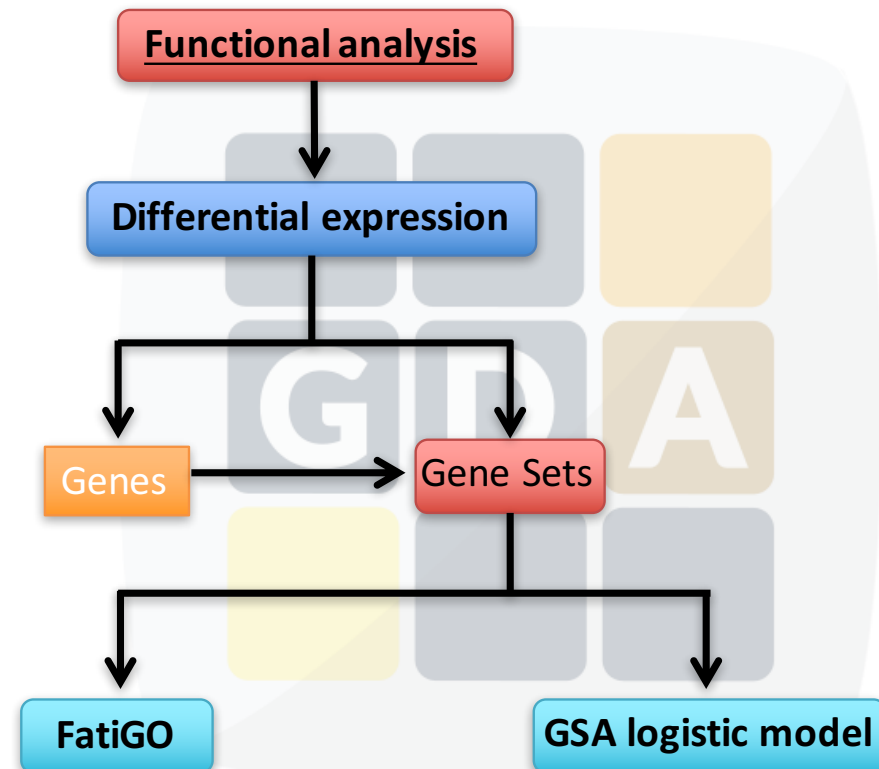
Where are we?

General NGS pipeline:



Functional analysis:

- Single enrichment: **FatiGO**
- Gene set enrichment: **Logistic Model**



Questions we try to answer

- Is there any significant functional enrichment in my gene list / gene sets?



Questions we try to answer

- Is there any significant functional enrichment in my gene list / gene sets?
- Are these genes involved in common pathways?



Questions we try to answer

- Is there any significant functional enrichment in my gene list / gene sets?
- Are these genes involved in common pathways?
- Do they share specific regulation?



Questions we try to answer

- Is there any significant functional enrichment in my gene list / gene sets?
- Are these genes involved in common pathways?
- Do they share specific regulation?
- Are they involved in the same disease?



Babelomics 5: FatiGO



GDA
International Course on
Genomic Data Analysis



PRINCIPE FELIPE
CENTRO DE INVESTIGACION

Single enrichment

Comparison between two list of genes with user defined annotations



Single enrichment

Comparison between two list of genes with user defined annotations

List 1: User provided

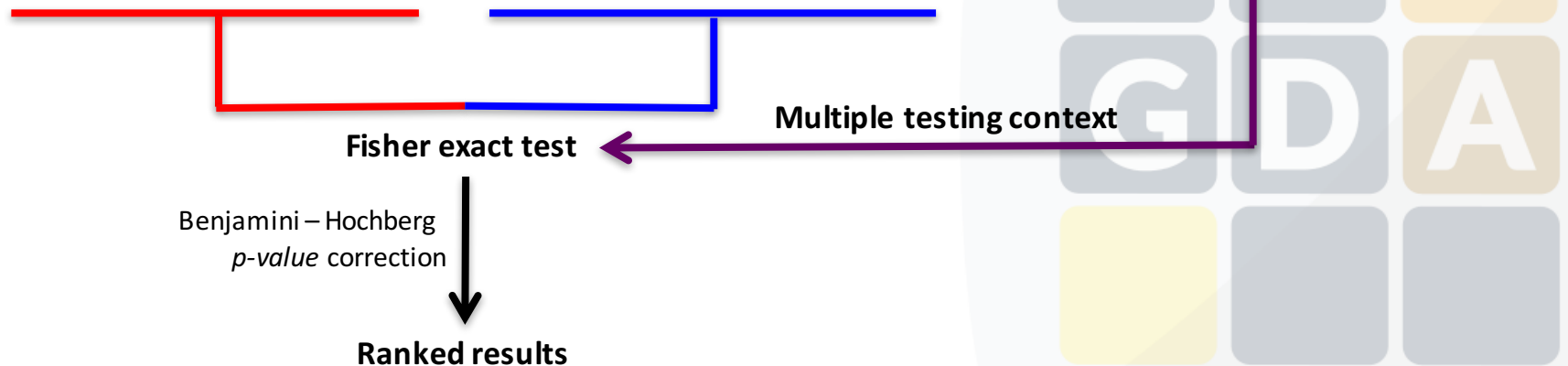
- Gene symbols
- Probe Ids
- Entrez Ids ...

List 2:

- User provided
- Rest of genome
- Complementary list

Functional annotations:

- Pre-defined annotations
- Custom annotations



Single enrichment

List 1

AATF	BIRC7	CIAS1	PEA15
APAF1	BIRC8	CIAS1	PNUTL2
API5	BNIP1	CIDEA	PRKAA1
AVEN	BNIP2	CIDEA	RIPK2
BAG1	BNIP3	CIDEB	RTN4
BAG2	BOK	CRADD	SON
BAG3	BRCA1	CRADD	SPHK2
BAG4	CARD10	DAD1	SPP1
BAG5	CARD11	DEDD	SPP1
BAK1	CARD12	DFFA	TNFAIP3
BAX	CARD12	DFFB	TNFRSF10A
BCL10	CARD14	EBAG9	TNFRSF10B
BCL2	CARD15	EBAG9	TNFRSF25
BCL2	CARD4	FADD	TNFRSF6
BCL2A1	CARD4	FAIM2	TNFRSF6B
BCL2L1	CARD6	HDAC1	
BCL2L10	CARD9	HDAC3	
BCL2L13	CASP1	HRK	
BCL2L2	CASP10	HTATIP2	
BCL2L7P1	CASP10	IER3	
BFAR	CASP10	IER3	
BID	CASP2	IER3	
BIK	CASP2	IL10	
BIRC1	CASP2	IL18	
BIRC1	CASP4	IL2	
BIRC2	CASP5	MALT1	
BIRC3	CASP8	MCL1	
BIRC4	CASP9	NME6	
BIRC5	CBX4	NOL3	
BIRC6	CFLAR	PDCD2	

List 2

ACTA1	DNAI1	KIF3B	MYO15A
ACTA2	DNAI2	KIF3C	MYO15B
ACTB	DNAI2	KIF4A	MYO18B
ACTC	DNAL4	KIF5A	MYO1A
ACTG1	DNALI1	KIF5B	MYO1A
ACTG2	DNCH1	KIF5B	MYO1B
ACTL6	DNC1	KIF5C	MYO1C
ACTR1A	DNCI2	KIF9	MYO1D
ACTR1B	DNCL1	KIFC1	MYO1E
APPBP2	DNCL2A	KIFC3	MYO1E
ATSV	DNCL2B	KNS2	MYO1F
C20orf23	DNCLI2	KNL7	MYO1G
CENPE	DNM1	MPHOSPH1	MYO3A
DCTN1	DNM2	MYH1	MYO3B
DCTN2	KIF11	MYH1	MYO3B
DNAH1	KIF13A	MYH10	MYO5A
DNAH11	KIF13B	MYH11	MYO5B
DNAH12	KIF14	MYH13	MYO5C
DNAH14	KIF17	MYH13	MYO6
DNAH14	KIF1B	MYH2	MYO6
DNAH14	KIF1C	MYH2	MYO7A
DNAH14	KIF2	MYH3	MYO7B
DNAH14	KIF20A	MYH4	MYO9A
DNAH17	KIF22	MYH4	MYO9B
DNAH2	KIF22	MYH6	OPA1
DNAH3	KIF23	MYH6	RPLP0P1
DNAH5	KIF23	MYH7	RPLP0P1
DNAH7	KIF25	MYH7B	TCTE1
DNAH8	KIF26A	MYH8	TCTEL1
DNAH9	KIF2C	MYH9	TIAF1
DNAH9	KIF3A	MYO10	

Functional annotations

Apoptotic process GO:0006915

CBX4	BRCA1	AVEN	MCL1
CRADD	HDAC3	FAIM2	BCL2L2
IL10	HDAC1	BID	BCL2L1
BAK1	RTN4	TNFRSF25	PEA15
BIRC8	IER3	BCL2	CFLAR
API5	HTATIP2	DAD1	CARD9
BCL10	DEDD	PRKAA1	SPHK2
OPA1	BNIP3	ACTC	CIDEB
DFFA	PDCD2	BIRC6	CASP2
DFFB	CASP5	BIRC5	EBAG9
BCL2A1	BAG5	BIRC3	NOL3
HRK	BAG4	BIRC2	CARD10
FADD	CASP4	BIRC1	TNFRSF10A
BCL2L13	CASP9	CASP10	CARD11
BCL2L10	BAG1	SON	CARD14
NME6	BOK	BAX	KIF1B
RIPK2	BAG3	BIK	DNM2
TNFAIP3	CASP8	APAF1	TNFRSF10B
TNFRSF6B	CASP1	TIAF1	BNIP1
BNIP2	CARD6	MALT1	CIDEA

Single enrichment

List 1

<u>AATF</u>	<u>BIRC7</u>	<u>CIAS1</u>	<u>PEA15</u>
<u>APAF1</u>	<u>BIRC8</u>	<u>CIAS1</u>	<u>PNUTL2</u>
<u>API5</u>	<u>BNIP1</u>	<u>CIDEA</u>	<u>PRKAA1</u>
<u>AVEN</u>	<u>BNIP2</u>	<u>CIDEA</u>	<u>RIPK2</u>
<u>BAG1</u>	<u>BNIP3</u>	<u>CIDEB</u>	<u>RTN4</u>
<u>BAG2</u>	<u>BOK</u>	<u>CRADD</u>	<u>SON</u>
<u>BAG3</u>	<u>BRCA1</u>	<u>CRADD</u>	<u>SPHK2</u>
<u>BAG4</u>	<u>CARD10</u>	<u>DAD1</u>	<u>SPP1</u>
<u>BAG5</u>	<u>CARD11</u>	<u>DEDD</u>	<u>SPP1</u>
<u>BAK1</u>	<u>CARD12</u>	<u>DFFA</u>	<u>TNFAIP3</u>
<u>BAX</u>	<u>CARD12</u>	<u>DFFB</u>	<u>TNFRSF10A</u>
<u>BCL10</u>	<u>CARD14</u>	<u>EBAG9</u>	<u>TNFRSF10B</u>
<u>BCL2</u>	<u>CARD15</u>	<u>EBAG9</u>	<u>TNFRSF25</u>
<u>BCL2</u>	<u>CARD4</u>	<u>FADD</u>	<u>TNFRSF6</u>
<u>BCL2A1</u>	<u>CARD4</u>	<u>FAIM2</u>	<u>TNFRSF6B</u>
<u>BCL2L1</u>	<u>CARD6</u>	<u>HDAC1</u>	
<u>BCL2L10</u>	<u>CARD9</u>	<u>HDAC3</u>	
<u>BCL2L13</u>	<u>CASP1</u>	<u>HRK</u>	
<u>BCL2L2</u>	<u>CASP10</u>	<u>HTATIP2</u>	
<u>BCL2L7P1</u>	<u>CASP10</u>	<u>IER3</u>	
<u>BFAR</u>	<u>CASP10</u>	<u>IER3</u>	
<u>BID</u>	<u>CASP2</u>	<u>IER3</u>	
<u>BIK</u>	<u>CASP2</u>	<u>IL10</u>	
<u>BIRC1</u>	<u>CASP2</u>	<u>IL18</u>	
<u>BIRC1</u>	<u>CASP4</u>	<u>IL2</u>	
<u>BIRC2</u>	<u>CASP5</u>	<u>MALT1</u>	
<u>BIRC3</u>	<u>CASP8</u>	<u>MCL1</u>	
<u>BIRC4</u>	<u>CASP9</u>	<u>NME6</u>	
<u>BIRC5</u>	<u>CBX4</u>	<u>NOL3</u>	
<u>BIRC6</u>	<u>CFLAR</u>	<u>PDCD2</u>	

List 2

<u>ACTA1</u>	<u>DNAI1</u>	<u>KIF3B</u>	<u>MYO15A</u>
<u>ACTA2</u>	<u>DNAI2</u>	<u>KIF3C</u>	<u>MYO15B</u>
<u>ACTB</u>	<u>DNAI2</u>	<u>KIF4A</u>	<u>MYO18B</u>
<u>ACTC</u>	<u>DNAL4</u>	<u>KIF5A</u>	<u>MYO1A</u>
<u>ACTG1</u>	<u>DNALI1</u>	<u>KIF5B</u>	<u>MYO1A</u>
<u>ACTG2</u>	<u>DNCH1</u>	<u>KIF5B</u>	<u>MYO1B</u>
<u>ACTL6</u>	<u>DNCI1</u>	<u>KIF5C</u>	<u>MYO1C</u>
<u>ACTR1A</u>	<u>DNCI2</u>	<u>KIF9</u>	<u>MYO1D</u>
<u>ACTR1B</u>	<u>DNCL1</u>	<u>KIFC1</u>	<u>MYO1E</u>
<u>APPBP2</u>	<u>DNCL2A</u>	<u>KIFC3</u>	<u>MYO1E</u>
<u>ATSV</u>	<u>DNCL2B</u>	<u>KNS2</u>	<u>MYO1F</u>
<u>C20orf23</u>	<u>DNCL2</u>	<u>KNL7</u>	<u>MYO1G</u>
<u>CENPE</u>	<u>DNM1</u>	<u>MPHOSPH1</u>	<u>MYO3A</u>
<u>DCTN1</u>	<u>DNM2</u>	<u>MYH1</u>	<u>MYO3B</u>
<u>DCTN2</u>	<u>KIF11</u>	<u>MYH1</u>	<u>MYO3B</u>
<u>DNAH1</u>	<u>KIF13A</u>	<u>MYH10</u>	<u>MYO5A</u>
<u>DNAH11</u>	<u>KIF13B</u>	<u>MYH11</u>	<u>MYO5B</u>
<u>DNAH12</u>	<u>KIF14</u>	<u>MYH13</u>	<u>MYO5C</u>
<u>DNAH14</u>	<u>KIF17</u>	<u>MYH13</u>	<u>MYO6</u>
<u>DNAH14</u>	<u>KIF1B</u>	<u>MYH2</u>	<u>MYO6</u>
<u>DNAH14</u>	<u>KIF1C</u>	<u>MYH2</u>	<u>MYO7A</u>
<u>DNAH14</u>	<u>KIF2</u>	<u>MYH3</u>	<u>MYO7B</u>
<u>DNAH14</u>	<u>KIF20A</u>	<u>MYH4</u>	<u>MYO9A</u>
<u>DNAH17</u>	<u>KIF22</u>	<u>MYH4</u>	<u>MYO9B</u>
<u>DNAH2</u>	<u>KIF22</u>	<u>MYH6</u>	<u>OPA1</u>
<u>DNAH3</u>	<u>KIF23</u>	<u>MYH6</u>	<u>RPLP0P1</u>
<u>DNAH5</u>	<u>KIF23</u>	<u>MYH7</u>	<u>RPLP0P1</u>
<u>DNAH7</u>	<u>KIF25</u>	<u>MYH7B</u>	<u>TCTE1</u>
<u>DNAH8</u>	<u>KIF26A</u>	<u>MYH8</u>	<u>TCTEL1</u>
<u>DNAH9</u>	<u>KIF2C</u>	<u>MYH9</u>	<u>TIAF1</u>
<u>DNAH9</u>	<u>KIF3A</u>	<u>MYO10</u>	

Functional annotations

Apoptotic process GO:0006915

<u>CBX4</u>	<u>BRCA1</u>	<u>AVEN</u>	<u>MCL1</u>
<u>CRADD</u>	<u>HDAC3</u>	<u>FAIM2</u>	<u>BCL2L2</u>
<u>IL10</u>	<u>HDAC1</u>	<u>BID</u>	<u>BCL2L1</u>
<u>BAK1</u>	<u>RTN4</u>	<u>TNFRSF25</u>	<u>PEA15</u>
<u>BIRC8</u>	<u>IER3</u>	<u>BCL2</u>	<u>CFLAR</u>
<u>API5</u>	<u>HTATIP2</u>	<u>DAD1</u>	<u>CARD9</u>
<u>BCL10</u>	<u>DEDD</u>	<u>PRKAA1</u>	<u>SPHK2</u>
<u>OPA1</u>	<u>BNIP3</u>	<u>ACTC</u>	<u>CIDEB</u>
<u>DFFA</u>	<u>PDCD2</u>	<u>BIRC6</u>	<u>CASP2</u>
<u>DFFB</u>	<u>CASP5</u>	<u>BIRC5</u>	<u>EBAG9</u>
<u>BCL2A1</u>	<u>BAG5</u>	<u>BIRC3</u>	<u>NOL3</u>
<u>HRK</u>	<u>BAG4</u>	<u>BIRC2</u>	<u>CARD10</u>
<u>FADD</u>	<u>CASP4</u>	<u>BIRC1</u>	<u>TNFRSF10A</u>
<u>BCL2L13</u>	<u>CASP9</u>	<u>CASP10</u>	<u>CARD11</u>
<u>BCL2L10</u>	<u>BAG1</u>	<u>SON</u>	<u>CARD14</u>
<u>NME6</u>	<u>BOK</u>	<u>BAX</u>	<u>KIF1B</u>
<u>RIPK2</u>	<u>BAG3</u>	<u>BIK</u>	<u>DNM2</u>
<u>TNFAIP3</u>	<u>CASP8</u>	<u>APAF1</u>	<u>TNFRSF10B</u>
<u>TNFRSF6B</u>	<u>CASP1</u>	<u>TIAF1</u>	<u>BNIP1</u>
<u>BNIP2</u>	<u>CARD6</u>	<u>MALT1</u>	<u>CIDEA</u>

Single enrichment

List 1

<u>AATF</u>	<u>BIRC7</u>	<u>CIAS1</u>	<u>PEA15</u>
<u>APAF1</u>	<u>BIRC8</u>	<u>CIAS1</u>	<u>PNUTL2</u>
<u>API5</u>	<u>BNIP1</u>	<u>CIDEA</u>	<u>PRKAA1</u>
<u>AVEN</u>	<u>BNIP2</u>	<u>CIDEA</u>	<u>RIPK2</u>
<u>BAG1</u>	<u>BNIP3</u>	<u>CIDEB</u>	<u>RTN4</u>
<u>BAG2</u>	<u>BOK</u>	<u>CRADD</u>	<u>SON</u>
<u>BAG3</u>	<u>BRCA1</u>	<u>CRADD</u>	<u>SPHK2</u>
<u>BAG4</u>	<u>CARD10</u>	<u>DAD1</u>	<u>SPP1</u>
<u>BAG5</u>	<u>CARD11</u>	<u>DEDD</u>	<u>SPP1</u>
<u>BAK1</u>	<u>CARD12</u>	<u>DFFA</u>	<u>TNFAIP3</u>
<u>BAX</u>	<u>CARD12</u>	<u>DFFB</u>	<u>TNFRSF10A</u>
<u>BCL10</u>	<u>CARD14</u>	<u>EBAG9</u>	<u>TNFRSF10B</u>
<u>BCL2</u>	<u>CARD15</u>	<u>EBAG9</u>	<u>TNFRSF25</u>
<u>BCL2</u>	<u>CARD4</u>	<u>FADD</u>	<u>TNFRSF6</u>
<u>BCL2A1</u>	<u>CARD4</u>	<u>FAIM2</u>	<u>TNFRSF6B</u>
<u>BCL2L1</u>	<u>CARD6</u>	<u>HDAC1</u>	
<u>BCL2L10</u>	<u>CARD9</u>	<u>HDAC3</u>	
<u>BCL2L13</u>	<u>CASP1</u>	<u>HRK</u>	
<u>BCL2L2</u>	<u>CASP10</u>	<u>HTATIP2</u>	
<u>BCL2L7P1</u>	<u>CASP10</u>	<u>IER3</u>	
<u>BFAR</u>	<u>CASP10</u>	<u>IER3</u>	
<u>BID</u>	<u>CASP2</u>	<u>IER3</u>	
<u>BIK</u>	<u>CASP2</u>	<u>IL10</u>	
<u>BIRC1</u>	<u>CASP2</u>	<u>IL18</u>	
<u>BIRC1</u>	<u>CASP4</u>	<u>IL2</u>	
<u>BIRC2</u>	<u>CASP5</u>	<u>MALT1</u>	
<u>BIRC3</u>	<u>CASP8</u>	<u>MCL1</u>	
<u>BIRC4</u>	<u>CASP9</u>	<u>NME6</u>	
<u>BIRC5</u>	<u>CBX4</u>	<u>NOL3</u>	
<u>BIRC6</u>	<u>CFLAR</u>	<u>PDCD2</u>	

List 2

<u>ACTA1</u>	<u>DNAI1</u>	<u>KIF3B</u>	<u>MYO15A</u>
<u>ACTA2</u>	<u>DNAI2</u>	<u>KIF3C</u>	<u>MYO15B</u>
<u>ACTB</u>	<u>DNAI2</u>	<u>KIF4A</u>	<u>MYO18B</u>
<u>ACTC</u>	<u>DNAL4</u>	<u>KIF5A</u>	<u>MYO1A</u>
<u>ACTG1</u>	<u>DNALI1</u>	<u>KIF5B</u>	<u>MYO1A</u>
<u>ACTG2</u>	<u>DNCH1</u>	<u>KIF5B</u>	<u>MYO1B</u>
<u>ACTL6</u>	<u>DNCI1</u>	<u>KIF5C</u>	<u>MYO1C</u>
<u>ACTR1A</u>	<u>DNCI2</u>	<u>KIF9</u>	<u>MYO1D</u>
<u>ACTR1B</u>	<u>DNCL1</u>	<u>KIFC1</u>	<u>MYO1E</u>
<u>APPBP2</u>	<u>DNCL2A</u>	<u>KIFC3</u>	<u>MYO1E</u>
<u>ATSV</u>	<u>DNCL2B</u>	<u>KNS2</u>	<u>MYO1F</u>
<u>C20orf23</u>	<u>DNCL2</u>	<u>KNL1</u>	<u>MYO1G</u>
<u>CENPE</u>	<u>DNM1</u>	<u>MPHOSPH1</u>	<u>MYO3A</u>
<u>DCTN1</u>	<u>DNM2</u>	<u>MYH1</u>	<u>MYO3B</u>
<u>DCTN2</u>	<u>KIF11</u>	<u>MYH1</u>	<u>MYO3B</u>
<u>DNAH1</u>	<u>KIF13A</u>	<u>MYH10</u>	<u>MYO5A</u>
<u>DNAH11</u>	<u>KIF13B</u>	<u>MYH11</u>	<u>MYO5B</u>
<u>DNAH12</u>	<u>KIF14</u>	<u>MYH13</u>	<u>MYO5C</u>
<u>DNAH14</u>	<u>KIF17</u>	<u>MYH13</u>	<u>MYO6</u>
<u>DNAH14</u>	<u>KIF1B</u>	<u>MYH2</u>	<u>MYO6</u>
<u>DNAH14</u>	<u>KIF1C</u>	<u>MYH2</u>	<u>MYO7A</u>
<u>DNAH14</u>	<u>KIF2</u>	<u>MYH3</u>	<u>MYO7B</u>
<u>DNAH14</u>	<u>KIF20A</u>	<u>MYH4</u>	<u>MYO9A</u>
<u>DNAH17</u>	<u>KIF22</u>	<u>MYH4</u>	<u>MYO9B</u>
<u>DNAH2</u>	<u>KIF22</u>	<u>MYH6</u>	<u>OPA1</u>
<u>DNAH3</u>	<u>KIF23</u>	<u>MYH6</u>	<u>RPLP0P1</u>
<u>DNAH5</u>	<u>KIF23</u>	<u>MYH7</u>	<u>RPLP0P1</u>
<u>DNAH7</u>	<u>KIF25</u>	<u>MYH7B</u>	<u>TCTE1</u>
<u>DNAH8</u>	<u>KIF26A</u>	<u>MYH8</u>	<u>TCTEL1</u>
<u>DNAH9</u>	<u>KIF2C</u>	<u>MYH9</u>	<u>TIAF1</u>
<u>DNAH9</u>	<u>KIF3A</u>	<u>MYO10</u>	

Functional annotations

Apoptotic process GO:0006915

<u>CBX4</u>	<u>BRCA1</u>	<u>AVEN</u>	<u>MCL1</u>
<u>CRADD</u>	<u>HDAC3</u>	<u>FAIM2</u>	<u>BCL2L2</u>
<u>IL10</u>	<u>HDAC1</u>	<u>BID</u>	<u>BCL2L1</u>
<u>BAK1</u>	<u>RTN4</u>	<u>TNFRSF25</u>	<u>PEA15</u>
<u>BIRC8</u>	<u>IER3</u>	<u>BCL2</u>	<u>CFLAR</u>
<u>API5</u>	<u>HTATIP2</u>	<u>DAD1</u>	<u>CARD9</u>
<u>BCL10</u>	<u>DEDD</u>	<u>PRKAA1</u>	<u>SPHK2</u>
<u>OPA1</u>	<u>BNIP3</u>	<u>ACTC</u>	<u>CIDEB</u>
<u>DFFA</u>	<u>PDCD2</u>	<u>BIRC6</u>	<u>CASP2</u>
<u>DFFB</u>	<u>CASP5</u>	<u>BIRC5</u>	<u>EBAG9</u>
<u>BCL2A1</u>	<u>BAG5</u>	<u>BIRC3</u>	<u>NOL3</u>
<u>HRK</u>	<u>BAG4</u>	<u>BIRC2</u>	<u>CARD10</u>
<u>FADD</u>	<u>CASP4</u>	<u>BIRC1</u>	<u>TNFRSF10A</u>
<u>BCL2L13</u>	<u>CASP9</u>	<u>CASP10</u>	<u>CARD11</u>
<u>BCL2L10</u>	<u>BAG1</u>	<u>SON</u>	<u>CARD14</u>
<u>NME6</u>	<u>BOK</u>	<u>BAX</u>	<u>KIF1B</u>
<u>RIPK2</u>	<u>BAG3</u>	<u>BIK</u>	<u>DNM2</u>
<u>TNFAIP3</u>	<u>CASP8</u>	<u>APAF1</u>	<u>TNFRSF10B</u>
<u>TNFRSF6B</u>	<u>CASP1</u>	<u>TIAF1</u>	<u>BNIP1</u>
<u>BNIP2</u>	<u>CARD6</u>	<u>MALT1</u>	<u>CIDEA</u>

Single enrichment

List 1

AATF	BIRC7	CIAS1	PEA15
APAF1	BIRC8	CIAS1	PNUTL2
API5	BNIP1	CIDEA	PRKAA1
AVEN	BNIP2	CIDEA	RIPK2
BAG1	BNIP3	CIDEB	RTN4
BAG2	BOK	CRADD	SON
BAG3	BRCA1	CRADD	SPHK2
BAG4	CARD10	DAD1	SPP1
BAG5	CARD11	DEDD	SPP1
BAK1	CARD12	DFFA	TNFAIP3
BAX	CARD12	DFFB	TNFRSF10A
BCL10	CARD14	EBAG9	TNFRSF10B
BCL2	CARD15	EBAG9	TNFRSF25
BCL2	CARD4	FADD	TNFRSF6
BCL2A1	CARD4	FAIM2	TNFRSF6B
BCL2L1	CARD6	HDAC1	
BCL2L10	CARD9	HDAC3	
BCL2L13	CASP1	HRK	
BCL2L2	CASP10	HTATIP2	
BCL2L7P1	CASP10	IER3	
BFAR	CASP10	IER3	
BID	CASP2	IER3	
BIK	CASP2	IL10	
BIRC1	CASP2	IL18	
BIRC1	CASP4	IL2	
BIRC2	CASP5	MALT1	
BIRC3	CASP8	MCL1	
BIRC4	CASP9	NME6	
BIRC5	CBX4	NOL3	
BIRC6	CFLAR	PDCD2	

List 2

ACTA1	DNAI1	KIF3B	MYO15A
ACTA2	DNAI2	KIF3C	MYO15B
ACTB	DNAI2	KIF4A	MYO18B
ACTC	DNAL4	KIF5A	MYO1A
ACTG1	DNALI1	KIF5B	MYO1A
ACTG2	DNCH1	KIF5B	MYO1B
ACTL6	DNC1	KIF5C	MYO1C
ACTR1A	DNCI2	KIF9	MYO1D
ACTR1B	DNCL1	KIFC1	MYO1E
APPBP2	DNCL2A	KIFC3	MYO1E
ATSV	DNCL2B	KNS2	MYO1F
C20orf23	DNCLI2	KNSL7	MYO1G
CENPE	DNM1	MPHOSPH1	MYO3A
DCTN1	DNM2	MYH1	MYO3B
DCTN2	KIF11	MYH1	MYO3B
DNAH1	KIF13A	MYH10	MYO5A
DNAH11	KIF13B	MYH11	MYO5B
DNAH12	KIF14	MYH13	MYO5C
DNAH14	KIF17	MYH13	MYO6
DNAH14	KIF1B	MYH2	MYO6
DNAH14	KIF1C	MYH2	MYO7A
DNAH14	KIF2	MYH3	MYO7B
DNAH14	KIF20A	MYH4	MYO9A
DNAH17	KIF22	MYH4	MYO9B
DNAH2	KIF22	MYH6	OPA1
DNAH3	KIF23	MYH6	OPA1
DNAH5	KIF23	MYH7	RPLP0P1
DNAH7	KIF25	MYH7B	RPLP0P1
DNAH8	KIF26A	MYH8	TCTE1
DNAH9	KIF2C	MYH9	TCTEL1
DNAH9	KIF3A	MYO10	TIAF1

Functional annotations

Positive regulation of immune system process GO0002684

IL18	RIPK2	MALT1	ACTB
ACTG1	TNFAIP3	MYO10	BIRC3
BCL10	CARD9	CASP8	BIRC2
MYH2	MYO1C	CARD11	BAX
FADD	MYO1G	BCL2	IL2



Single enrichment

List 1

AATF	BIRC7	CIAS1	PEA15
APAF1	BIRC8	CIAS1	PNUTL2
API5	BNIP1	CIDEA	PRKAA1
AVEN	BNIP2	CIDEA	<u>RIPK2</u>
BAG1	BNIP3	CIDEB	RTN4
BAG2	BOK	CRADD	SON
BAG3	BRCA1	CRADD	SPHK2
BAG4	CARD10	DAD1	SPP1
BAG5	<u>CARD11</u>	DEDD	SPP1
BAK1	<u>CARD12</u>	DFFA	<u>TNFAIP3</u>
BAX	CARD12	DFFB	TNFRSF10A
<u>BCL10</u>	CARD14	EBAG9	TNFRSF10B
<u>BCL2</u>	CARD15	EBAG9	TNFRSF25
BCL2	CARD4	<u>FADD</u>	TNFRSF6
BCL2A1	CARD4	<u>FAIM2</u>	TNFRSF6B
BCL2L1	CARD6	HDAC1	
BCL2L10	<u>CARD9</u>	HDAC3	
BCL2L13	CASP1	HRK	
BCL2L2	CASP10	HTATIP2	
BCL2L7P1	CASP10	IER3	
BFAR	CASP10	IER3	
BID	CASP2	IER3	
BIK	CASP2	IL10	
BIRC1	CASP2	<u>IL18</u>	
BIRC1	CASP4	<u>IL2</u>	
<u>BIRC2</u>	CASP5	<u>MALT1</u>	
<u>BIRC3</u>	<u>CASP8</u>	<u>MCL1</u>	
BIRC4	CASP9	NME6	
BIRC5	CBX4	NOL3	
BIRC6	CFLAR	PDCD2	

List 2

ACTA1	DNAI1	KIF3B	MYO15A
ACTA2	DNAI2	KIF3C	MYO15B
<u>ACTB</u>	DNAI2	KIF4A	MYO18B
ACTC	DNAL4	KIF5A	MYO1A
<u>ACTG1</u>	DNALI1	KIF5B	MYO1A
ACTG2	DNCH1	KIF5B	MYO1B
ACTL6	DNC1	KIF5C	<u>MYO1C</u>
ACTR1A	DNCI2	KIF9	MYO1D
ACTR1B	DNCL1	KIFC1	MYO1E
APPBP2	DNCL2A	KIFC3	MYO1E
ATSV	DNCL2B	KNS2	MYO1F
C20orf23	DNCLI2	KNSL7	<u>MYO1G</u>
CENPE	DNM1	MPHOSPH1	MYO3A
DCTN1	DNM2	MYH1	MYO3B
DCTN2	KIF11	MYH1	MYO3B
DNAH1	KIF13A	MYH10	MYO5A
DNAH11	KIF13B	MYH11	MYO5B
DNAH12	KIF14	MYH13	MYO5C
DNAH14	KIF17	MYH13	MYO6
DNAH14	KIF1B	MYH2	MYO6
DNAH14	KIF1C	<u>MYH2</u>	MYO7A
DNAH14	KIF2	<u>MYH3</u>	MYO7B
DNAH14	KIF20A	MYH4	MYO9A
DNAH17	KIF22	MYH4	MYO9B
DNAH2	KIF22	MYH6	OPA1
DNAH3	KIF23	MYH6	OPA1
DNAH5	KIF23	MYH7	RPLP0P1
DNAH7	KIF25	MYH7B	RPLP0P1
DNAH8	KIF26A	MYH8	TCTE1
DNAH9	KIF2C	MYH9	TCTEL1
DNAH9	KIF3A	<u>MYO10</u>	TIAF1

Functional annotations

Positive regulation of immune system process GO0002684

<u>IL18</u>	<u>RIPK2</u>	<u>MALT1</u>	<u>ACTB</u>
<u>ACTG1</u>	<u>TNFAIP3</u>	<u>MYO10</u>	<u>BIRC3</u>
<u>BCL10</u>	<u>CARD9</u>	<u>CASP8</u>	<u>BIRC2</u>
<u>MYH2</u>	<u>MYO1C</u>	<u>CARD11</u>	<u>BAX</u>
<u>FADD</u>	<u>MYO1G</u>	<u>BCL2</u>	<u>IL2</u>



Single enrichment

List 1

AATF	BIRC7	CIAS1	PEA15
APAF1	BIRC8	CIAS1	PNUTL2
API5	BNIP1	CIDEA	PRKAA1
AVEN	BNIP2	CIDEA	RIPK2
BAG1	BNIP3	CIDEB	RTN4
BAG2	BOK	CRADD	SON
BAG3	BRCA1	CRADD	SPHK2
BAG4	CARD10	DAD1	SPP1
BAG5	CARD11	DEDD	SPP1
BAK1	CARD12	DFFA	TNFAIP3
BAX	CARD12	DFFB	TNFRSF10A
BCL10	CARD14	EBAG9	TNFRSF10B
BCL2	CARD15	EBAG9	TNFRSF25
BCL2	CARD4	FADD	TNFRSF6
BCL2A1	CARD4	FAIM2	TNFRSF6B
BCL2L1	CARD6	HDAC1	
BCL2L10	CARD9	HDAC3	
BCL2L13	CASP1	HRK	
BCL2L2	CASP10	HTATIP2	
BCL2L7P1	CASP10	IER3	
BFAR	CASP10	IER3	
BID	CASP2	IER3	
BIK	CASP2	IL10	
BIRC1	CASP2	IL18	
BIRC1	CASP4	IL2	
BIRC2	CASP5	MALT1	
BIRC3	CASP8	MCL1	
BIRC4	CASP9	NME6	
BIRC5	CBX4	NOL3	
BIRC6	CFLAR	PDCD2	

List 2

ACTA1	DNAI1	KIF3B	MYO15A
ACTA2	DNAI2	KIF3C	MYO15B
ACTB	DNAI2	KIF4A	MYO18B
ACTC	DNAL4	KIF5A	MYO1A
ACTG1	DNALI1	KIF5B	MYO1A
ACTG2	DNCH1	KIF5B	MYO1B
ACTL6	DNCI1	KIF5C	MYO1C
ACTR1A	DNCI2	KIF9	MYO1D
ACTR1B	DNCL1	KIFC1	MYO1E
APPBP2	DNCL2A	KIFC3	MYO1E
ATSV	DNCL2B	KNS2	MYO1F
C20orf23	DNCL2	KNL1	MYO1G
CENPE	DNM1	MPHOSPH1	MYO3A
DCTN1	DNM2	MYH1	MYO3B
DCTN2	KIF11	MYH1	MYO3B
DNAH1	KIF13A	MYH10	MYO5A
DNAH11	KIF13B	MYH11	MYO5B
DNAH12	KIF14	MYH13	MYO5C
DNAH14	KIF17	MYH13	MYO6
DNAH14	KIF1B	MYH2	MYO6
DNAH14	KIF1C	MYH2	MYO7A
DNAH14	KIF2	MYH3	MYO7B
DNAH14	KIF20A	MYH4	MYO9A
DNAH17	KIF22	MYH4	MYO9B
DNAH2	KIF22	MYH6	OPA1
DNAH3	KIF23	MYH6	OPA1
DNAH5	KIF23	MYH7	RPLP0P1
DNAH7	KIF25	MYH7B	RPLP0P1
DNAH8	KIF26A	MYH8	TCTE1
DNAH9	KIF2C	MYH9	TCTE1
DNAH9	KIF3A	MYO10	TIAF1

Functional annotations

Positive regulation of immune system process GO0002684

<u>IL18</u>	<u>RIPK2</u>	<u>MALT1</u>	<u>ACTB</u>
<u>ACTG1</u>	<u>TNFAIP3</u>	<u>MYO10</u>	<u>BIRC3</u>
<u>BCL10</u>	<u>CARD9</u>	<u>CASP8</u>	<u>BIRC2</u>
<u>MYH2</u>	<u>MYO1C</u>	<u>CARD11</u>	<u>BAX</u>
<u>FADD</u>	<u>MYO1G</u>	<u>BCL2</u>	<u>IL2</u>

Organ development GO0048515

IL18	MYH6	KIF3A	MYO7A
IL10	MYH9	SPHK2	CASP5
BAK1	MYO18B	MYO1E	CASP9
MYO6	RIPK2	MALT1	BOK
MYO3A	BCL2L2	HDAC1	BAG3
MYH3	BCL2L1	MYH11	CASP8
DFFB	MYO15A	MYH10	CASP2
MYH7	SPP1	RTN4	ACTA1
FADD	CFLAR	DEDD	ACTA2
ACTB	DNAH5	MYO5A	CARD11
ACTC	BCL2	BID	FAIM2
BIRC6	BAX	BIK	APAF1
IL2			

Single enrichment

List 1

AATF	BIRC7	CIAS1	PEA15
<u>APAF1</u>	BIRC8	CIAS1	PNUTL2
API5	BNIP1	CIDEA	PRKAA1
AVEN	BNIP2	CIDEA	<u>RIPK2</u>
BAG1	BNIP3	CIDEB	<u>RTN4</u>
BAG2	<u>BOK</u>	CRADD	SON
<u>BAG3</u>	BRCA1	CRADD	SPHK2
BAG4	CARD10	DAD1	<u>SPP1</u>
BAG5	<u>CARD11</u>	<u>DEDD</u>	SPP1
<u>BAK1</u>	CARD12	DDFA	TNFAIP3
<u>BAX</u>	CARD12	<u>DFFB</u>	TNFRSF10A
BCL10	CARD14	<u>EBAG9</u>	TNFRSF10B
<u>BCL2</u>	CARD15	EBAG9	TNFRSF25
BCL2	CARD4	FADD	TNFRSF6
BCL2A1	CARD4	<u>FAIM2</u>	TNFRSF6B
<u>BCL2L1</u>	CARD6	<u>HDAC1</u>	
BCL2L10	CARD9	HDAC3	
BCL2L13	CASP1	HRK	
<u>BCL2L2</u>	CASP10	HTATIP2	
BCL2L7P1	CASP10	IER3	
BFAR	CASP10	IER3	
<u>BID</u>	<u>CASP2</u>	IER3	
<u>BIK</u>	CASP2	<u>IL10</u>	
BIRC1	CASP2	<u>IL18</u>	
BIRC1	CASP4	<u>IL2</u>	
BIRC2	<u>CASP5</u>	<u>MALT1</u>	
BIRC3	<u>CASP8</u>	MCL1	
BIRC4	<u>CASP9</u>	NME6	
BIRC5	CBX4	NOL3	
<u>BIRC6</u>	<u>CELFAR</u>	PDCD2	

List 2

<u>ACTA1</u>	DNAI1	KIF3B	<u>MYO15A</u>
<u>ACTA2</u>	DNAI2	KIF3C	<u>MYO15B</u>
<u>ACTB</u>	DNAI2	KIF4A	<u>MYO18B</u>
<u>ACTC</u>	DNAL4	KIF5A	MYO1A
ACTG1	DNALI1	KIF5B	MYO1A
ACTG2	DNCH1	KIF5B	MYO1B
ACTL6	DNC1	KIF5C	MYO1C
ACTR1A	DNCI2	KIF9	MYO1D
ACTR1B	DNCL1	KIFC1	<u>MYO1E</u>
APPBP2	DNCL2A	KIFC3	MYO1E
ATSV	DNCL2B	KNS2	MYO1F
C20orf23	DNCLI2	KNL57	MYO1G
CENPE	DNM1	MPHOSPH1	<u>MYO3A</u>
DCTN1	DNM2	MYH1	MYO3B
DCTN2	KIF11	MYH1	MYO3B
DNAH1	KIF13A	<u>MYH10</u>	<u>MYO5A</u>
DNAH11	KIF13B	<u>MYH11</u>	MYO5B
DNAH12	KIF14	MYH13	MYO5C
DNAH14	KIF17	MYH13	<u>MYO6</u>
DNAH14	KIF1B	MYH2	MYO6
DNAH14	KIF1C	MYH2	<u>MYO7A</u>
DNAH14	KIF2	<u>MYH3</u>	MYO7B
DNAH14	KIF20A	MYH4	MYO9A
DNAH17	KIF22	MYH4	MYO9B
DNAH2	KIF22	<u>MYH6</u>	OPA1
DNAH3	KIF23	MYH6	OPA1
<u>DNAH5</u>	KIF23	<u>MYH7</u>	RPLP0P1
DNAH7	KIF25	MYH7B	RPLP0P1
DNAH8	KIF26A	MYH8	TCTE1
DNAH9	KIF2C	<u>MYH9</u>	TCTEL1
DNAH9	<u>KIF3A</u>	MYO10	TIAF1

Functional annotations

Positive regulation of immune system process GO0002684

IL18	RIPK2	MALT1	ACTB
ACTG1	TNFAIP3	MYO10	BIRC3
BCL10	CARD9	CASP8	BIRC2
MYH2	MYO1C	CARD11	BAX
FADD	MYO1G	BCL2	IL2

Organ development GO0048515

<u>IL18</u>	<u>MYH6</u>	<u>KIF3A</u>	<u>MYO7A</u>
<u>IL10</u>	<u>MYH9</u>	SPHK2	<u>CASP5</u>
<u>BAK1</u>	<u>MYO18B</u>	<u>MYO1E</u>	<u>CASP9</u>
<u>MYO6</u>	<u>RIPK2</u>	<u>MALT1</u>	<u>BOK</u>
<u>MYO3A</u>	<u>BCL2L2</u>	<u>HDAC1</u>	<u>BAG3</u>
<u>MYH3</u>	<u>BCL2L1</u>	<u>MYH11</u>	<u>CASP8</u>
<u>DFFB</u>	<u>MYO15A</u>	<u>MYH10</u>	<u>CASP2</u>
<u>MYH7</u>	<u>SPP1</u>	<u>RTN4</u>	<u>ACTA1</u>
<u>FADD</u>	<u>CELFAR</u>	<u>DEDD</u>	<u>ACTA2</u>
<u>ACTB</u>	<u>DNAH5</u>	<u>MYO5A</u>	<u>CARD11</u>
<u>ACTC</u>	<u>BCL2</u>	<u>BID</u>	<u>FAIM2</u>
<u>BIRC6</u>	<u>BAX</u>	<u>BIK</u>	<u>APAF1</u>
<u>IL2</u>			

FatiGO

Nucleic Acids Research, 2007, Vol. 35, Web Server Issue: W91-W96
doi:10.1093/nar/gkm260

FatiGO+: a functional profiling tool for genomic data.
Integration of functional annotation, regulatory motifs
and interaction data with microarray experiments

BIOINFORMATICS APPLICATIONS NOTE Vol. 20 no. 4 2004, pages 578-585
DOI: 10.1093/bioinformatics/btg045



**FatiGO: a web tool for finding significant
associations of Gene Ontology terms with
groups of genes**

Fátima Al-Shahrour, Ramón Díaz-Uribe and Joaquín Dopazo*

Bioinformatics Unit, Centro Nacional de Investigaciones Oncológicas (CNIO), Madrid
Feminólez Almagro 2, 28002 Madrid, Spain

Received on August 16, 2003; initial version on September 30, 2003; accepted on October 1, 2003
Advance Access publication January 22, 2004

ABSTRACT

Summary: We present a simple but powerful procedure to extract Gene Ontology (GO) terms that are significantly over- or under-represented in sets of genes within the context of a genome-scale experiment (DNA microarray, proteomics, etc.). Such procedure has been implemented as a web application, FatiGO, allowing for easy and interactive querying. FatiGO, which takes the multiple-testing nature of statistical contrast into account, currently includes GO associations for diverse organisms (human, mouse, fly, worm and yeast) and the TrEMBL/Swissprot GOAnnotations@EBI correspondences from the European Bioinformatics Institute. **Availability:** <http://fati.go.bioinformatics.org>
Contact: jpdopazo@cnio.es

Most resources available that collect information regarding gene or protein function, biological properties, etc., are based on the pre-genomic design in which the information is accorded and displayed in a one-gene-at-a-time format. Nevertheless, many problems related to functional genomics involve the detection of biological properties, functions, etc., shared by a set of genes, that are then aside from the remaining ones. The practical application of methods of pre-genomic design to these problems is drastically limited when thousands of genes are involved in the comparative study. The use of techniques of automatic management of biological information, such as text mining, in studying the coherence of gene groups obtained from different methodologies has only recently been addressed (Olivero *et al.*, 2000; Raychaudhuri *et al.*, 2002; Pavlidis *et al.*, 2003), although its practical application still has many drawbacks (Blanchier *et al.*, 2002). Furthermore, real implementations are often scarce and beyond the reach of many users.

An alternative to extracting information from scientific text sources is by using ontologies. In its most simple representation, ontologies provide a structured description

of biological information that is extremely useful for computational management. One of the most widely accepted ontologies is Gene Ontology (GO; Ashburner *et al.*, 2000), which organizes information for molecular function, biological processes and cellular components for a number of different organisms. The potential of GO terms as a structured source of information however, has yet to be fully exploited.

Here we present FatiGO, a web-based application (<http://fati.go.bioinformatics.org>). Since the publication of FatiGO in the GO consortium web page (<http://www.geneontology.org>) less than a year ago, a number of tools have been implemented based on the same idea of mapping biological knowledge on sets of genes. Thus, OntoExpress (Kucan *et al.*, 2002), which generates tables that correlate groups of genes to biochemical and molecular functions or MAP-Primer (Daigler *et al.*, 2003), which, using a searchable web interface, identifies GO terms over-represented in the data. A similar tool, FunPep (Bakken *et al.*, 2002), evaluates groups of yeast genes in terms of their annotations in diverse databases. Many of these tools are stand-alone applications with user-friendly interfaces, but obviously suffer limitations in processing large amounts of data. Moreover, important issues such as the multiple-testing nature of the statistical contrast are not well addressed.

FatiGO is used to extract relevant GO terms for a group of genes with respect to a set of genes of reference (typically the rest of genes). The terms are considered to be relevant by the application of a Fisher's exact test that considers the multiple-testing nature of the statistical contrast performed. Multiple testing is an important issue that is, nevertheless, scarcely addressed (Shimizu, 2002). If the multiple-testing nature of the statistical contrast is not taken into account an increase in the rate of false positives (i.e. terms identified as over- or under-represented whose proportions, in reality, are not significantly different, occur). FatiGO can deal with thousands of genes from different organisms (currently human, mouse, *Drosophila*, worm, yeast, as well as genes whose proteins are included in Swissprot database), and can be queried using

*To whom correspondence should be addressed.

Annotation

GO:0006915
GO:0002684
GO:0048515

List 1

76
14
30

List 2

5
6
19



FatiGO

Nucleic Acids Research, 2007, Vol. 35, Web Server issue: W91-W96
doi:10.1093/nar/gkm260

FatiGO+: a functional profiling tool for genomic data.
Integration of functional annotation, regulatory motifs
and interaction data with microarray experiments

BIOINFORMATICS APPLICATIONS NOTE doi:10.1093/bioinformatics/btl045



**FatiGO: a web tool for finding significant
associations of Gene Ontology terms with
groups of genes**

Fátima Al-Shahrour, Ramón Díaz-Uriarte and Joaquín Dopazo*

Bioinformatics Unit, Centro Nacional de Investigaciones Oncológicas (CNIO), Madrid
Feminólez Almagro 2, 28002 Madrid, Spain

Received on August 16, 2005; revised on September 30, 2005; accepted on October 1, 2005
Advance Access publication January 22, 2006

ABSTRACT
Summary: We present a simple but powerful procedure to extract Gene Ontology (GO) terms that are significantly over- or under-represented in sets of genes within the context of a genome-scale experiment (DNA microarray, proteomics, etc.). Said procedure has been implemented as a web application, FatiGO, allowing for easy and interactive querying. FatiGO, which takes the multiple-testing nature of statistical contrast into account, currently includes GO associations for diverse organisms (human, mouse, fly, worm and yeast) and the TrEMBL/Swissprot GOAnnotations@EBI correspondences from the European Bioinformatics Institute. Availability: <http://fatigo.bioinformatics.univ-se.es>
Contact: jopazo@cnio.es

Most resources available that collect information regarding gene or protein function, biological properties, etc., are based on the pre-genomic design in which the information is accessed and displayed in a one-gene-at-a-time format. Nevertheless, many problems related to functional genomics involve the detection of biological properties, functions, etc., shared by a set of genes, that arise from within the remaining ones. The practical application of methods of pre-genomic design to these problems is drastically limited when thousands of genes are involved in the comparative study. The use of techniques of automatic management of biological information, such as text mining, in studying the coherence of gene groups obtained from different methodologies has only recently been addressed (Olivero *et al.*, 2000; Raychaudhuri *et al.*, 2002; Pavlidis *et al.*, 2003), although its practical application still has many drawbacks (Blancher *et al.*, 2002). Furthermore, real implementations are often scarce and beyond the reach of many users.

An alternative to extracting information from scientific text sources is by using ontologies. In its most simple representation, ontologies provide a structured description

of biological information that is extremely useful for computational management. One of the most widely accepted ontologies is Gene Ontology (GO; Ashburner *et al.*, 2000), which organizes information for molecular function, biological processes and cellular components for a number of different organisms. The potential of GO terms as a structured source of information however, has yet to be fully exploited. Here we present FatiGO, a web-based application (<http://fatigo.bioinformatics.univ-se.es>). Since the publication of FatiGO in the GO consortium web page (<http://www.geneontology.org>) less than a year ago, a number of tools have been implemented based on the same idea of mapping biological knowledge on sets of genes. Thus, OntoExpress (Kucan *et al.*, 2002), which generates tables that correlate groups of genes to biochemical and molecular functions or MAP-Primer (Daigler *et al.*, 2003), which, using a searchable web interface, identifies GO terms over-represented in the data. A similar tool, FatiSpot (Bilimoria *et al.*, 2002), evaluates groups of gene genes in terms of their annotations in diverse databases. Many of these tools are stand-alone applications with user-friendly interfaces, but obviously suffer limitations in processing large amounts of data. Moreover, important issues such as the multiple-testing nature of the statistical contrast are not well addressed.

FatiGO is used to extract relevant GO terms for a group of genes with respect to a set of genes of reference (typically the rest of genes). The terms are considered to be relevant by the application of a Fisher's exact test that considers the multiple-testing nature of the statistical contrast performed. Multiple testing is an important issue that is, nevertheless, scarcely addressed (Shimin, 2002). If the multiple-testing nature of the statistical contrast is not taken into account an increase in the rate of false positives (i.e. genes identified as over- or under-represented whose proportions, in reality, are not significantly different, occur). FatiGO can deal with thousands of genes from different organisms (currently human, mouse, *Drosophila*, worm, yeast, as well as genes whose proteins are included in Swissprot database), and can be queried using

*To whom correspondence should be addressed.

Annotation

GO:0006915
GO:0002684
GO:0048515

List 1

76
14
30

List 2

5
6
19

Fisher exact test

Annotation

GO:0006915
GO:0002684
GO:0048515

LOR

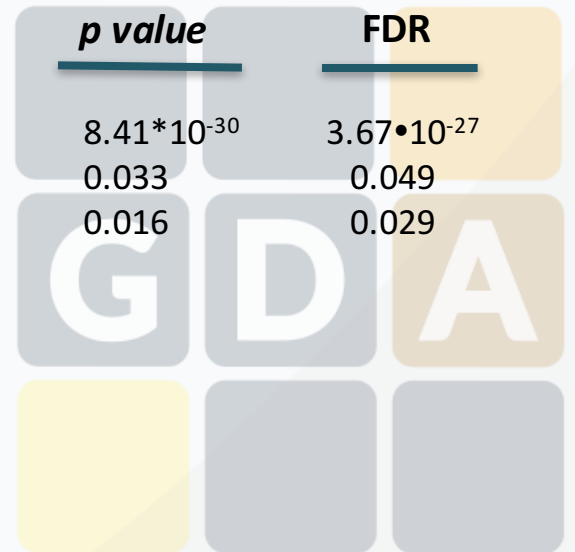
-4.13
-1.11
-0.79

p value

$8.41 \cdot 10^{-30}$
0.033
0.016

FDR

$3.67 \cdot 10^{-27}$
0.049
0.029



FatiGO

Nucleic Acids Research, 2007, Vol. 35, Web Server Issue: WWW-1896
doi:10.1093/nar/gkm260

FatiGO+: a functional profiling tool for genomic data.
Integration of functional annotation, regulatory motifs
and interaction data with microarray experiments

BIOINFORMATICS APPLICATIONS NOTE Vol. 20 no. 4 2004, pages 578-585
DOI: 10.1093/bioinformatics/btg045



**FatiGO: a web tool for finding significant
associations of Gene Ontology terms with
groups of genes**

Fátima Al-Shahrour, Ramón Díaz-Uriarte and Joaquín Dopazo*

Bioinformática Unit, Centro Nacional de Investigaciones Oncológicas (CNIO), Madrid
Feminólez Almagro 2, 28029 Madrid, Spain

Received on August 16, 2003; revised on September 30, 2003; accepted on October 1, 2003
Advance Access publication January 22, 2004

ABSTRACT
Summary: We present a simple but powerful procedure to extract Gene Ontology (GO) terms that are significantly over- or under-represented in sets of genes within the context of a genome-scale experiment (DNA microarray, proteomics, etc.). This procedure has been implemented as a web application, FatiGO, allowing for easy and interactive querying. FatiGO, which takes the multiple-testing nature of statistical contrast into account, currently includes GO associations for diverse organisms (human, mouse, fly, worm and yeast) and the TrEMBL/Swissprot/GenProt/UniProtKB/TrEMBL correspondence from the European Bioinformatics Institute. Availability: <http://fatigo.bioinfo.cnio.es>
Contact: jdopazo@cnio.es

Most resources available that collect information regarding gene or protein function, biological properties, etc., are based on the pre-genomic design in which the information is accessed and displayed in a one-gene-at-a-time format. Nevertheless, many problems related to functional genomics involve the detection of biological properties, functions, etc., shared by a set of genes, that arise there aside from the remaining ones. The practical application of methods of pre-genomic design to these problems is drastically limited when thousands of genes are involved in the comparative study. The use of techniques of automatic management of biological information, such as text mining, in studying the coherence of gene groups obtained from different methodologies has only recently been addressed (Olivero *et al.*, 2000; Raychaudhuri *et al.*, 2002; Pavlidis *et al.*, 2003), although its practical application still has many drawbacks (Blanchier *et al.*, 2002). Furthermore, real implementations are often source and beyond the reach of many users.

An alternative to extracting information from scientific text sources is by using ontologies. In its most simple representation, ontologies provide a structured description

of biological information that is extremely useful for computational management. One of the most widely accepted ontologies is Gene Ontology (GO; Ashburner *et al.*, 2000), which organizes information for molecular function, biological processes and cellular components for a number of different organisms. The potential of GO terms as a structured source of information however, has yet to be fully exploited. Here we present FatiGO, a web-based application (<http://fatigo.bioinfo.cnio.es>). Since the publication of FatiGO in the GO consortium web page (<http://www.geneontology.org>) less than a year ago, a number of tools have been implemented based on the same idea of mapping biological knowledge on sets of genes. Thus, OntoExpress (Kucan *et al.*, 2002), which generates tables that correlate groups of genes to biochemical and molecular functions or MAP-Filter (Daigler *et al.*, 2003), which, using a searchable web interface, identifies GO terms over-represented in the data. A similar tool, FatiSpot (Blanchier *et al.*, 2002), evaluates groups of gene sets in terms of their annotations in diverse databases. Many of these tools are stand-alone applications with user-friendly interfaces, but obviously suffer limitations in processing large amounts of data. Moreover, important issues such as the multiple-testing nature of the statistical contrast are not well addressed.

FatiGO is used to extract relevant GO terms for a group of genes with respect to a set of genes of reference (typically the rest of genes). The terms are considered to be relevant by the application of a Fisher's exact test that considers the multiple-testing nature of the statistical contrast performed. Multiple testing is an important issue that is, nevertheless, scarcely addressed (Sham, 2002). If the multiple-testing nature of the statistical contrast is not taken into account an increase in the rate of false positives (i.e. genes identified as over- or under-represented whose proportions, in reality, are not significantly different), occurs. FatiGO can deal with thousands of genes from different organisms (currently human, mouse, *Drosophila*, worm, yeast, as well as genes whose proteins are included in Swissprot database), and can be queried using

*To whom correspondence should be addressed.

Annotation

GO:0006915
GO:0002684
GO:0048515

List 1

76
14
30

List 2

5
6
19

Fisher exact test

Annotation

GO:0006915
GO:0002684
GO:0048515

LOR

-4.13
-1.11
-0.79

p value

8.41 * 10⁻³⁰
0.033
0.016

FDR

3.67 * 10⁻²⁷
0.049
0.029



FatiGO on *Babelomics 5*

Tool selection

Functional ▾ dcrespo

Single enrichment

- ▶ FatiGO
- ▶ Gene set enrichment
- ▶ Logistic model

Network enrichment

- ▶ Snow
- ▶ Gene set network enrichment
- ▶ Network Miner

Comparison

Define your comparison

Id list vs Id list Id List vs Rest of genome Id List vs Rest of ids contained in your annotations (complementary list)

First list selection

Select your data

File Text area

The files must be on the server to select them.
You can upload files using the button  inside file browser.

Second list selection

<p>List 2</p> <p><input checked="" type="radio"/> File <input type="radio"/> Text area</p> <p>The files must be on the server to select them. You can upload files using the button  inside file browser.</p> <p><input type="button" value="File browser"/> <input type="button" value="WorkSpace/"/></p>	<p>List 2</p> <p>Rest of the genome</p> <p>List 2</p> <p>Rest of ids contained in your annotations (complementary list)</p>
---	---

Test options

Fisher exact test

Two tailed

Over-represent

Over-represent


Remove duplicates

- Never
- Remove on each list separately
- Remove on each list and common ids
- Remove from list2 those appearing in list1 (complementary list)

Job information

Job information

Output folder

You can create folders using the button  inside file browser.

Job name

Description

Databases

Select an organism

- Human (homo sapiens)
- Mouse (mus musculus)
- Rat (rattus norvegicus)
- Fruitfly (drosophila melanogaster)
- Zebrafish (danio rerio)
- Saccharomyces cerevisiae
- Caenorhabditis elegans
- Arabidopsis thaliana

Databases

Select an organism ▾

GO biological process

GO molecular function


GO cellular component

GOSlim GOA

Interpro

Genome-Scale Metabolic Network

Your annotations

The files must be on the server to select them.
You can upload files using the button  inside file browser.

Data input

Gene list: bioformat "ID"

```
MBP-de-Dan-2:example datocresbe$  
MY05B  
DNAL4  
MY018B  
MYH9  
C20orf23  
RPLP0P1  
RPLP0P1  
DNCL2A  
KIF3B  
MYH7B  
MYH6  
MYH6  
MYH7  
KIF13B  
KIF26A  
KNS2  
DNCH1  
MY09A  
MY05C  
MY05A  
MY01E
```

1 Column: Gene ID

Custom annotation: bioformat Gene vs annotation

```
MBP-de-Dan-2:example datocresbe$ head -30 ../annotation  
"rno04014_47 49" "plasma membrane"  
"rno04014_47 49" "positive regulation of cell prolifer  
"rno04014_47 49" "extrinsic component of cytoplasmic s  
"rno04014_47 49" "GTP binding"  
"rno04014_47 49" "Golgi apparatus"  
"rno04014_47 49" "positive regulation of angiogenesis"  
"rno04014_47 49" "molecular_function"  
"rno04014_47 49" "biological_process"  
"rno04014_47 49" "cellular_component"  
"rno04014_42" "lipid catabolic process"  
"rno04014_42" "phospholipase A2 activity"  
"rno04014_42" "molecular_function"  
"rno04014_42" "mitochondrial inner membrane"  
"rno04014_42" "cellular response to insulin stimulus"  
"rno04014_42" "glucose transport"  
"rno04014_43" "RNA polymerase II transcription factor  
"rno04014_43" "transcription from RNA polymerase II pr  
"rno04014_43" "cell differentiation"  
"rno04014_44" "nucleus"  
"rno04014_44" "RNA polymerase II transcription factor  
"rno04014_44" "transcription from RNA polymerase II pr  
"rno04014_44" "molecular_function"
```

Column 1:
Gene ID

Column 2:
Function

Reading results

Analysis overview

Job information

Name: *motor vs apop*
 Description: [Description](#)
 Tool: *fatigo*
 Output folder: [Workspace/analysis/20160224155426/](#)

Input data

Species: *hsa*
 Duplicates management: *Never remove*
 Fisher exact test: *Two tailed*
 List 1 (after duplicates managing) [clean_list1.txt](#)
 List 2 (after duplicates managing) [clean_list2.txt](#)

Summary

Id annotations per DB [annotations_per_db.txt](#)

#DB	List1	List 2
GO biological process propagated	85 of 124 (68.55%) 63.66 annotations/id	79 of 105 (75.24%) 99.28 annotations/id
1 Results < 1 of 1 >		

Results

Significant Results

Number of significant terms per DB [significant_count_0.05.txt](#)

Select p-value:

#DB	N° of significant terms
GO biological process propagated	588
1 Results < 1 of 1 >	

GO biological process propagated

GO biological process propagated significant terms (pvalue<0.05) [significant_go_biological_process_propagated_0.05.txt](#)

Term	Term size	Term size (in genome)	Annotations lists	Annotated ids list	Odds ratio (log e)	Pvalue	Adj. Pvalue
protein secretion(GO:0009306)	5	500	List 1: 1.0% List 2: 4.76%	List 1: IL10 List 2: CIDEA CASP5 CASP1	-1.7976931349e+304	0.019	0.03
regulation of release of cytochrome c from mitochondria(GO:0090199)	8	139	List 1: 0.81% List 2: 6.67%	List 1: OPA1 List 2: BAK1 HRK BCL2L1 BNIP3	-2.17	0.025	0.039
positive regulation of protein ubiquitination(GO:0031398)	6	376	List 1: 1.0% List 2: 5.71%	List 1: BCL10 List 2: RIPK2 MALT1 BRCA1	-1.7976931349e+304	0.0086	0.017
regulation of protein modification process(GO:0031399)	21	2965	List 1: 1.61% List 2: 18.1%	List 1: DCTN1 List 2: BAK1 CENPE BCL10 RIPK2 TNFAIP3	-2.6	1.2e-5	6.7e-5
regulation of protein ubiquitination(GO:0031396)	8	559	List 1: 1.0% List 2: 7.62%	List 1: BCL10 List 2: RIPK2 TNFAIP3 MALT1	-1.7976931349e+304	0.0017	0.0046
immune system development(GO:0002520)	14	1449	List 1: 1.61% List 2: 11.43%	List 1: MYH9 List 2: MYO1E BAK1 FADD RIPK2	-2.06	0.0039	0.0089
leukocyte differentiation(GO:0002521)	11	815	List 1: 0.81% List 2: 9.52%	List 1: MYH9 List 2: IL10 BAK1 FADD RIPK2	-2.56	0.0031	0.0081
positive regulation of intracellular signal transduction(GO:0030230)	22	1709	List 1: 1.0% List 2: 20.95%	List 1: BCL10 List 2: FADD RIPK2	-1.7976931349e+304	9.4e-9	8.9e-8

Download results

All results

GO biological process propagated [go_biological_process_propagated.txt](#)

Annotation files

Annotations for GO biological process propagated [go_biological_process_propagated.annot](#)

Reading results

Analysis overview

Job information

Name: *motor vs apop*
 Description: *Description*
 Tool: *fatigo*
 Output folder: *Workspace/analysis/20160224155426/*

Input data

Species: *hsa*
 Duplicates management: *Never remove*
 Fisher exact test: *Two tailed*
 List 1 (after duplicates managing) [clean_list1.txt](#)
 List 2 (after duplicates managing) [clean_list2.txt](#)

Summary

Id annotations per DB [annotations_per_db.txt](#)

#DB	List1	List 2
GO biological process propagated	85 of 124 (68.55%) 63.66 annotations/id	79 of 105 (75.24%) 99.28 annotations/id
1 Results < 1 of 1 >		

Results

Significant Results

Number of significant terms per DB [significant_count_0.05.txt](#)

Select p-value:

#DB	N° of significant terms
GO biological process propagated	588
1 Results < 1 of 1 >	

GO biological process propagated

GO biological process propagated significant terms (pvalue<0.05) [significant_go_biological_process_propagated_0.05.txt](#)

Term	Term size	Term size (in genome)	Annotations lists	Annotated ids list	Odds ratio (log e)	Pvalue	Adj. Pvalue
protein secretion(GO:0009306)	5	500	List 1: 1.0% List 2: 4.76%	List 1: IL10 List 2: CIDEA CASP5 CASP1	-1.7976931349e+304	0.019	0.03
regulation of release of cytochrome c from mitochondria(GO:0090199)	8	139	List 1: 0.81% List 2: 6.67%	List 1: OPA1 List 2: BAK1 HRK BCL2L1 BNIP3	-2.17	0.025	0.039
positive regulation of protein ubiquitination(GO:0031398)	6	376	List 1: 1.0% List 2: 5.7%	List 1: BCL10 List 2: RIPK2 MALT1 BRCA1	-1.7976931349e+304	0.0086	0.017
regulation of protein modification process(GO:0031399)	21	2965	List 1: 1.61% List 2: 18.1%	List 1: DCTN1 List 2: BAK1 CENPE BCL10 RIPK2 TNFAIP3	-2.6	1.2e-5	6.7e-5
regulation of protein ubiquitination(GO:0031396)	8	559	List 1: 1.0% List 2: 7.62%	List 1: BCL10 List 2: RIPK2 TNFAIP3 MALT1	-1.7976931349e+304	0.0017	0.0046
immune system development(GO:0002520)	14	1449	List 1: 1.61% List 2: 11.43%	List 1: MYH9 List 2: MYO1E BAK1 FADD RIPK2	-2.06	0.0039	0.0089
leukocyte differentiation(GO:0002521)	11	815	List 1: 0.81% List 2: 9.52%	List 1: MYH9 List 2: IL10 BAK1 FADD RIPK2	-2.56	0.0031	0.0081
positive regulation of intracellular signal transduction(GO:0030233)	22	1709	List 1: 1.0% List 2: 20.95%	List 1: BCL10 List 2: FADD RIPK2	-1.7976931349e+304	9.4e-9	8.9e-8

Enriched class

Annotated genes from each list

False Discovery Rate

Download results

All results

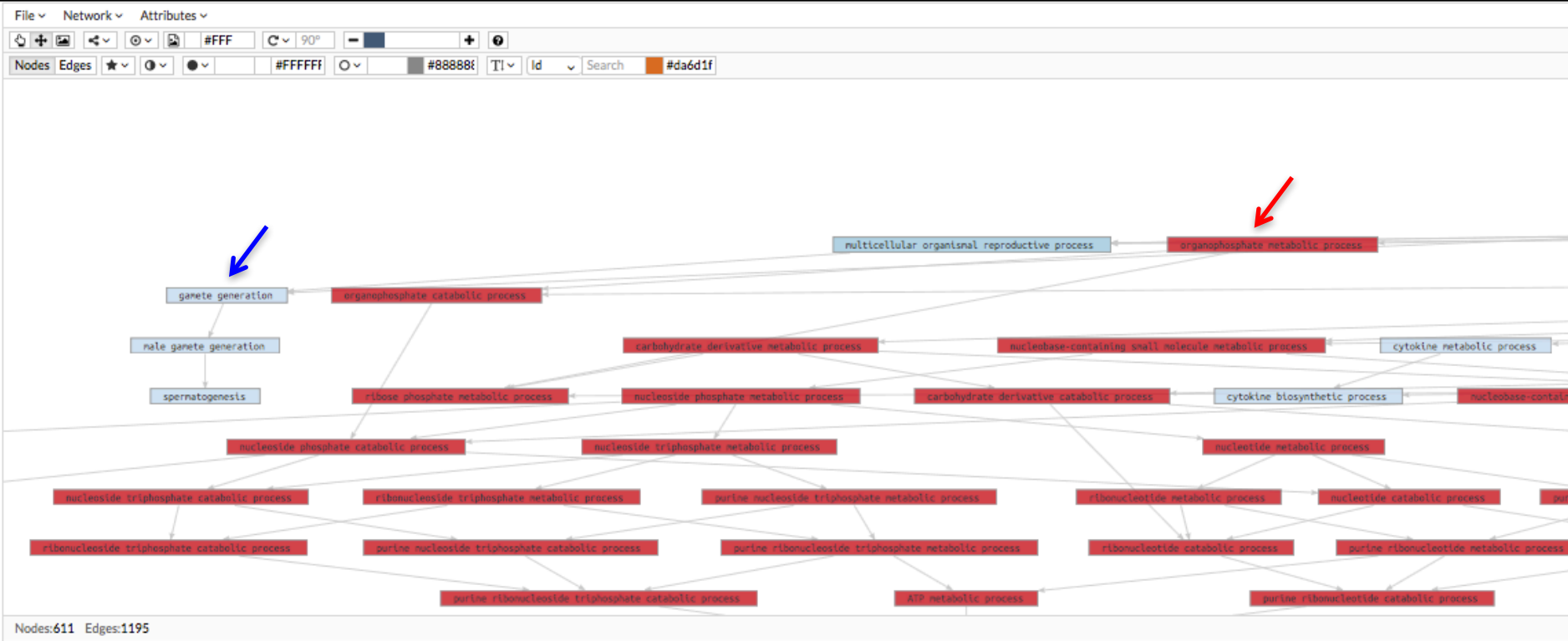
GO biological process propagated [go_biological_process_propagated.txt](#)

Annotation files

Annotations for GO biological process propagated [go_biological_process_propagated.annot](#)

Reading results

Results visualization



Babelomics 5: Gene Set Analysis



GDA

International Course on
Genomic Data Analysis



PRINCIPE FELIPE
CENTRO DE INVESTIGACION

Gene set enrichment

Comparison between a **ranked list** with user defined annotations



Gene set enrichment

Comparison between a **ranked list** with user defined annotations

Ranked List: User provided

- Gene symbols
- Probe Ids
- Entrez Ids ...

Functional annotations:

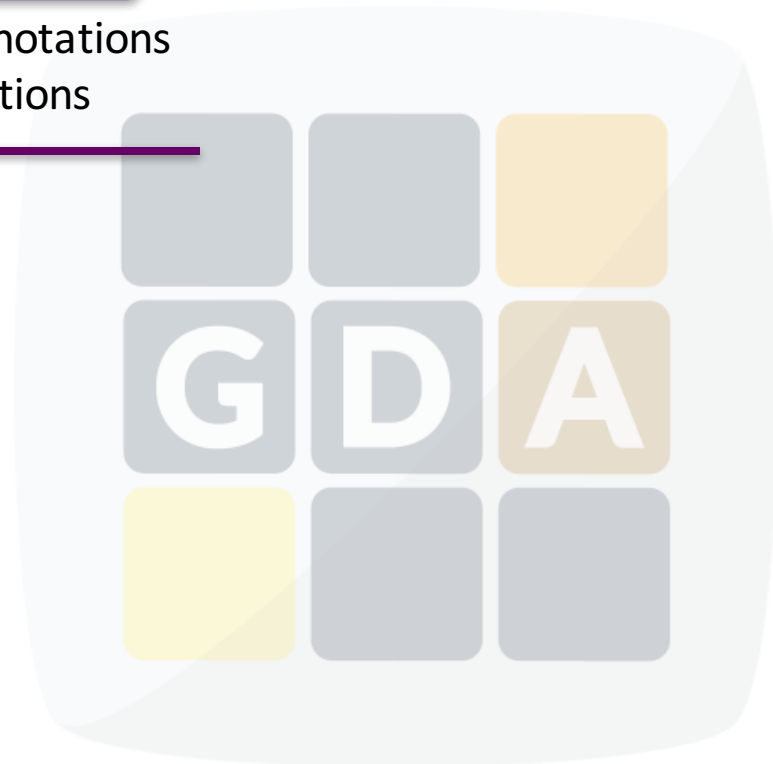
- Pre-defined annotations
- Custom annotations

Multiple testing context

Logistic regression model

Benjamini – Hochberg
p-value correction

Ranked results



GSA: Logistic model

BMC Bioinformatics

Software
From genes to functional classes in the study of biological systems
 Fátima Al-Shahrour¹, Leonardo Arbiza¹, Hernán Dopazo³, Jaime Huerta-Cepas^{1,2}, Pablo Mínguez², David Montaner^{1,2} and Joaquín Dopazo^{*1,2}

BIOINFORMATICS ORIGINAL PAPER

Gene expression

Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information

Fátima Al-Shahrour, Ramón Díaz-Uriarte¹ and Joaquín Dopazo^{*}

OPEN ACCESS Freely available online

PLoS ONE

Multidimensional Gene Set Analysis of Genomic Data

David Montaner^{1,2}, Joaquín Dopazo^{1,2,3,4}

¹Departamento de Estadística y Genética, Centro de Investigación y Desarrollo en Biología Celular (CIBIC), Universidad Nacional de Córdoba (UNC), Centro de Investigación y Desarrollo en Biología Celular, Córdoba, Argentina, ²Centro de Investigación y Desarrollo en Biología Celular (CIBIC), Universidad Nacional de Córdoba (UNC), Córdoba, Argentina, ³Centro de Investigación y Desarrollo en Biología Celular (CIBIC), Universidad Nacional de Córdoba (UNC), Córdoba, Argentina, ⁴Centro de Investigación y Desarrollo en Biología Celular (CIBIC), Universidad Nacional de Córdoba (UNC), Córdoba, Argentina

Abstract

Understanding the functional implications of changes in gene expression, mutations, etc., is the aim of most genomic experiments. To achieve this, several functional profiling methods have been proposed. Such methods study the behaviour of different gene modules (e.g. gene ontology terms) in response to one particular variable (e.g. differential gene expression). In spite of the wealth of information provided by functional profiling methods, a common limitation to all of them is their inherent unidimensional nature. In order to overcome this restriction we present a multidimensional logistic model that allows studying the relationship of gene modules with different genomic scale measurements (e.g. differential expression, genotyping association, methylation, copy number alterations, heterozygosity, etc.) simultaneously. Moreover, the relationship of such functional modules with the interaction of the variables can also be studied, which produces novel results impossible to be derived from the conventional unidimensional functional profiling methods. We report novel results of gene set associations that remained undetected by the conventional one-dimensional gene set analysis in several examples. Our findings demonstrate the potential of the proposed approach for the discovery of new cell functionalities with complex dependencies on more than one variable.

David Montaner D, Dopazo J (2010) Multidimensional Gene Set Analysis of Genomic Data. PLoS ONE 5(4): e12348. doi:10.1371/journal.pone.012348

Editor: Jörg Hübner, Heidelberg Institute of Evolutionary Biology, GERMANY

Received December 1, 2009; Accepted March 20, 2010; Published April 27, 2010

Copyright: © 2010 Montaner, Dopazo. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by grants from the Spanish Ministerio de Sanidad y Consumo, the National Institute of Health Research (ISCIII) and the Spanish Ministerio de Educación y Ciencia. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

*Email: jdopazo@ibic.unco.edu.ar

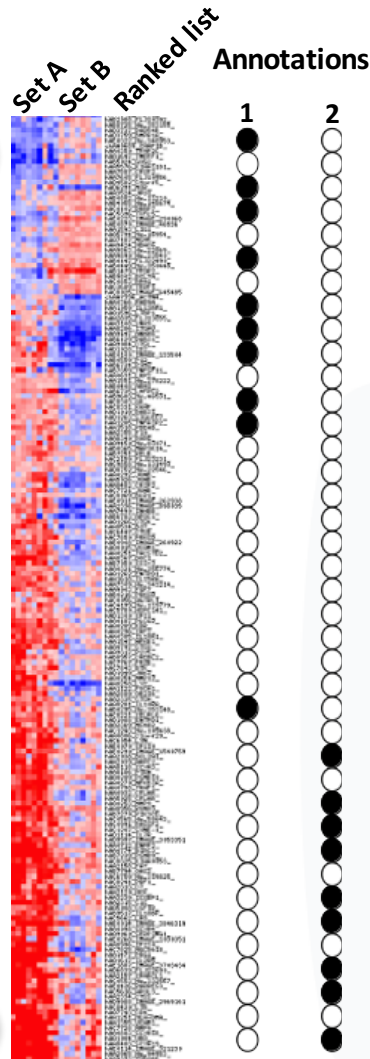
Introduction

The development of new genomic technologies, such as microarrays of gene expression, genotyping or array-CGH, along with the new microarraying sequencing techniques is increasing the volume of data throughout annually. As a direct consequence of this, the bottleneck in functional genomics has shifted from the data production phase to the data analysis steps. In particular, the necessity for providing a functional interpretation at molecular level data accounts for the genome-wide experimental designs has promoted the development of different methods for the functional analysis of this type of data in the last years [1–3]. It is widely agreed that most of the biological functionality of the cell arises from complex interactions among their molecular components that differ operational interacting entities or modules [4]. Functions collectively performed by such modules have conceptually been represented in different ways. Gene ontology (GO) [5] and KEGG pathway [6] are the most popular and widely used module definitions although many other are available in different repositories (e.g. Reactome [7], BioCarta, etc.). For practical purposes, functional modules are hierarchically defined as sets of genes sharing functional annotations extracted from any of these repositories. Functional profiling methods exploit different definitions of modules in an attempt of understanding the functional basis of high-throughput experimental results [7] by linking functional modules to different experimental conditions [8], have been used for this purpose. More sensitive approaches, generally known as gene set analysis (GSA)

methods, pioneered by the Gene Set Enrichment Analysis (GSEA) [9] were later proposed [10]. In the original formulation, GSA methods aimed to identify distinct sets of functionally related genes (modules) with a coordinate and significant over- or under-expression across the complete list of genes ranked according to the differential expression [11,12,13,14,15]. GSA methods can detect such modules even if their gene components are not significantly differentially expressed when tested individually. GSA has been successfully applied to the analysis of microarray experiments and has contributed to the adaptation of a systems biology perspective in distinct fields such as cancer [16]. Recent findings brought about by the application of GSA methods to microarray experiments [17] are consistent with the idea that pathways, rather than individual genes, appear to govern the course of tumorigenesis [18]. The use of GSA has been extended to other sets beyond microarrays, such as methylation [19], QTL analysis [20] or genotyping [21]. Nevertheless, the different versions of GSA published to date [12,10] are inherently one-dimensional. Its application to the analysis of genomic datasets is as prone limited to the study of a unique variable measured for the genes. The experimental conditions tested, even if controlled by other variables (e.g. age, gender, treatment, etc.), are typically summarized into a single value for each gene (e.g. differential expression in a case-control, risk in the case of survival analysis, etc.) which is then tested through GSA. However, the complex set of different high-throughput methodologies allows the observation of different measurements for the genes such as methylation status, typing status, linkage

Logistic regression

Statistic



GSA: Logistic model

BMC Bioinformatics

Software
From genes to functional classes in the study of biological systems
 Fátima Al-Shahrour¹, Leonardo Arbiza¹, Hernán Dopazo¹, Jaime Huerta-Cepas^{1,2}, Pablo Minguéz², David Montaner^{1,2} and Joaquín Dopazo^{1,2}

BIOINFORMATICS ORIGINAL PAPER

Gene expression
Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information

Fátima Al-Shahrour, Ramón Díaz-Uriarte¹ and Joaquín Dopazo^{1,2}

OPEN ACCESS Freely available online

Multidimensional Gene Set Analysis of Genomic Data

David Montaner^{1,2}, Joaquín Dopazo^{1,2,3,4}

¹Departamento de Estadística y Genética, Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain; ²Functional Genomics Unit (FGU), Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain; ³CEBPG de Informática Biológica (CIBIB), Valencia, Spain

Abstract

Understanding the functional implications of changes in gene expression, mutations, etc. is the aim of most genomic experiments. To achieve this, several functional profiling methods have been proposed. Such methods study the behaviour of different gene modules (e.g. gene ontology terms) in response to one particular variable (e.g. differential gene expression). In spite of the wealth of information provided by functional profiling methods, a common limitation to all of them is their inherent unidimensional nature. In order to overcome this restriction we present a multidimensional logistic model that allows studying the relationship of gene modules with different genome scale measurements (e.g. differential expression, genotyping association, methylation, copy number alterations, heterozygosity, etc.) simultaneously. Moreover, the relationship of such functional modules with the interaction among the variables can also be studied, which produces novel results impossible to be derived from the conventional unidimensional functional profiling methods. We report sound results of gene set associations that remained undetected by the conventional one-dimensional gene set analysis in several examples. Our findings demonstrate the potential of the proposed approach for the discovery of new cell functionalities with complex dependencies on more than one variable.

Keywords: Network; D. Dopazo; D. Montaner; Multidimensional Gene Set Analysis of Genomic Data; PLoS ONE; e112148; doi:10.1371/journal.pone.0112148

Editor: Jörg Hübner, Heidelberg Institute for Theoretical Studies, Germany

Received December 1, 2010; Accepted March 20, 2011; Published April 27, 2011

Copyright: © 2011 Montaner, Dopazo. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by grants from the project B2010BMD0001 from the Spanish Ministry of Science and Innovation. The National Institute of Health Research (NIHR) is a part of Genomics England. The FUGU de Informática Biológica is an initiative of the FGU. The FGU is an initiative of the CIPF. The work was partially supported by grants from the Spanish Ministry of Science and Innovation and the Spanish Ministry of Health. We thank the reviewers for their comments and suggestions. We thank the reviewers for their comments and suggestions. We thank the reviewers for their comments and suggestions.

Competing Interests: The authors have declared that no competing interests exist.

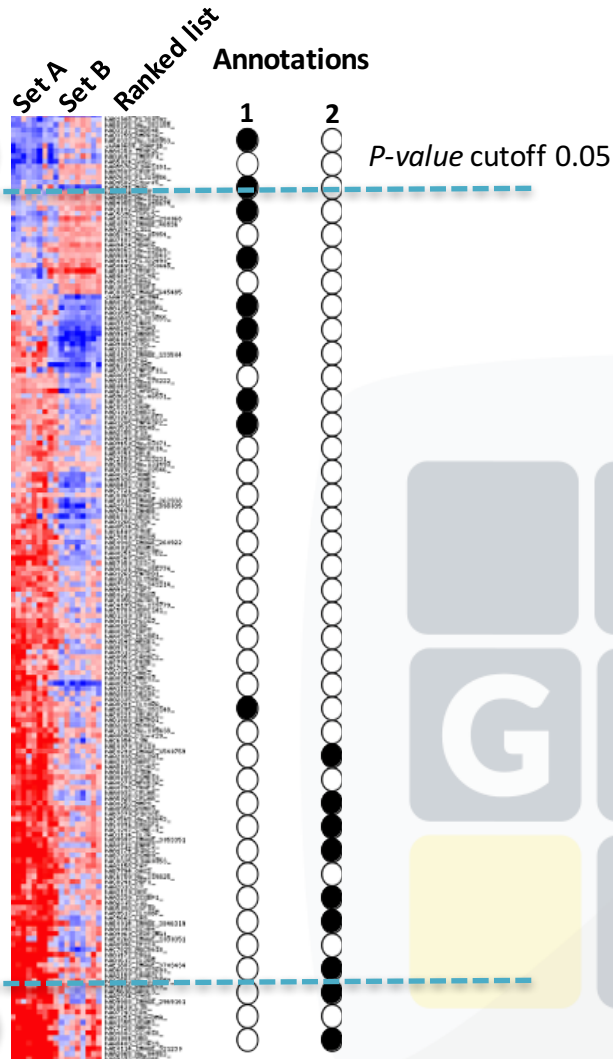
*Email: dopazo@cipf.es

Introduction

The development of new genomic technologies, such as microarrays of gene expression, genotyping or array-CGH, along with the new annotation-enrichment techniques is increasing the volume of data throughout genomics. As a direct consequence of this, the bottleneck in functional genomics has shifted from the data production phase to the data analysis steps. In particular, the necessity for providing a functional interpretation at molecular level has become one of the main challenges in genomics. Gene set analysis (GSA) and KEGG pathway [5] are the most popular and widely used module definition methods, among other are available in different repositories (e.g. Reactome [6], BioCarta, etc.). For practical purposes, functional modules are hierarchically defined as sets of genes sharing functional annotations extracted from any of these repositories. Functional profiling methods exploit different definitions of modules in an attempt of understanding the functional basis of high-throughput experimental results [7]. Initially, functional enrichment methods, in different implementations [8], have been used for this purpose. More advanced approaches, generally known as gene set analysis (GSA)

methods, pioneered by the Gene Set Enrichment Analysis (GSEA) [9] were later proposed [10]. In the original formulation, GSA methods aimed to identify distinct sets of functionally related genes (modules) with a coordinate and significant over- or under-expression across the complete list of genes ranked according to the differential expression [11,12,13,14,15]. GSA methods can detect such modules even if their gene composition are not significantly differentially expressed when tested individually. GSA has been successfully applied to the analysis of microarray experiments and has contributed to the adaptation of a systems biology perspective in distinct fields such as cancer [16]. Recent findings brought about by the application of GSA methods to microarray experiments [17] are consistent with the idea that pathways, rather than individual genes, appear to govern the course of tumorigenesis [18]. The use of GSA has been extended to other sets beyond microarrays, such as methylation [19], QTL analysis [20] or genotyping [21]. Nevertheless, the different variants of GSA published to date [12,20] are inherently one-dimensional. Its application to the analysis of genomic datasets is as present limited to the study of a unique variable measured for the genes. The experimental conditions tested, even if controlled by other variables (e.g. age, gender, treatment, etc.), are typically summarized into a unique value for each gene (e.g. differential expression in a case-control, risk in the case of survival analysis, etc.) which is then tested through GSA. However, the capture set of different high-throughput methodologies allows the observation of different measurements for the genes such as methylation status, genotyping status, linkage

Logistic regression
 Statistic



GSA: Logistic model

BMC Bioinformatics

Software
From genes to functional classes in the study of biological systems
 Fátima Al-Shahrour¹, Leonardo Arbiza¹, Hernán Dopazo¹, Jaime Huerta-Cepas^{1,2}, Pablo Minguéz², David Montaner^{1,2} and Joaquín Dopazo^{1,2}

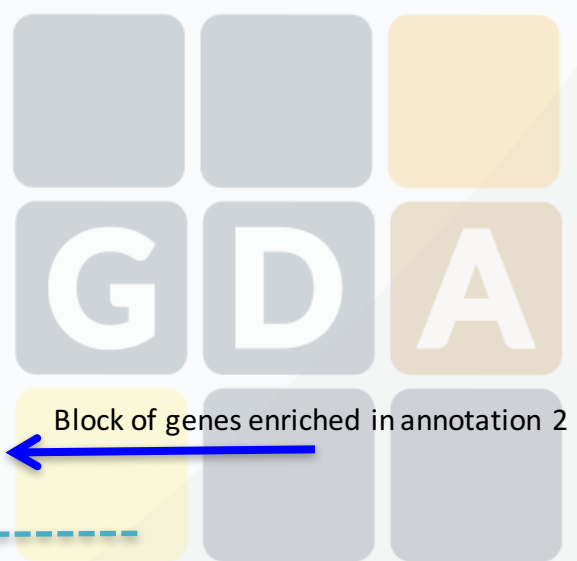
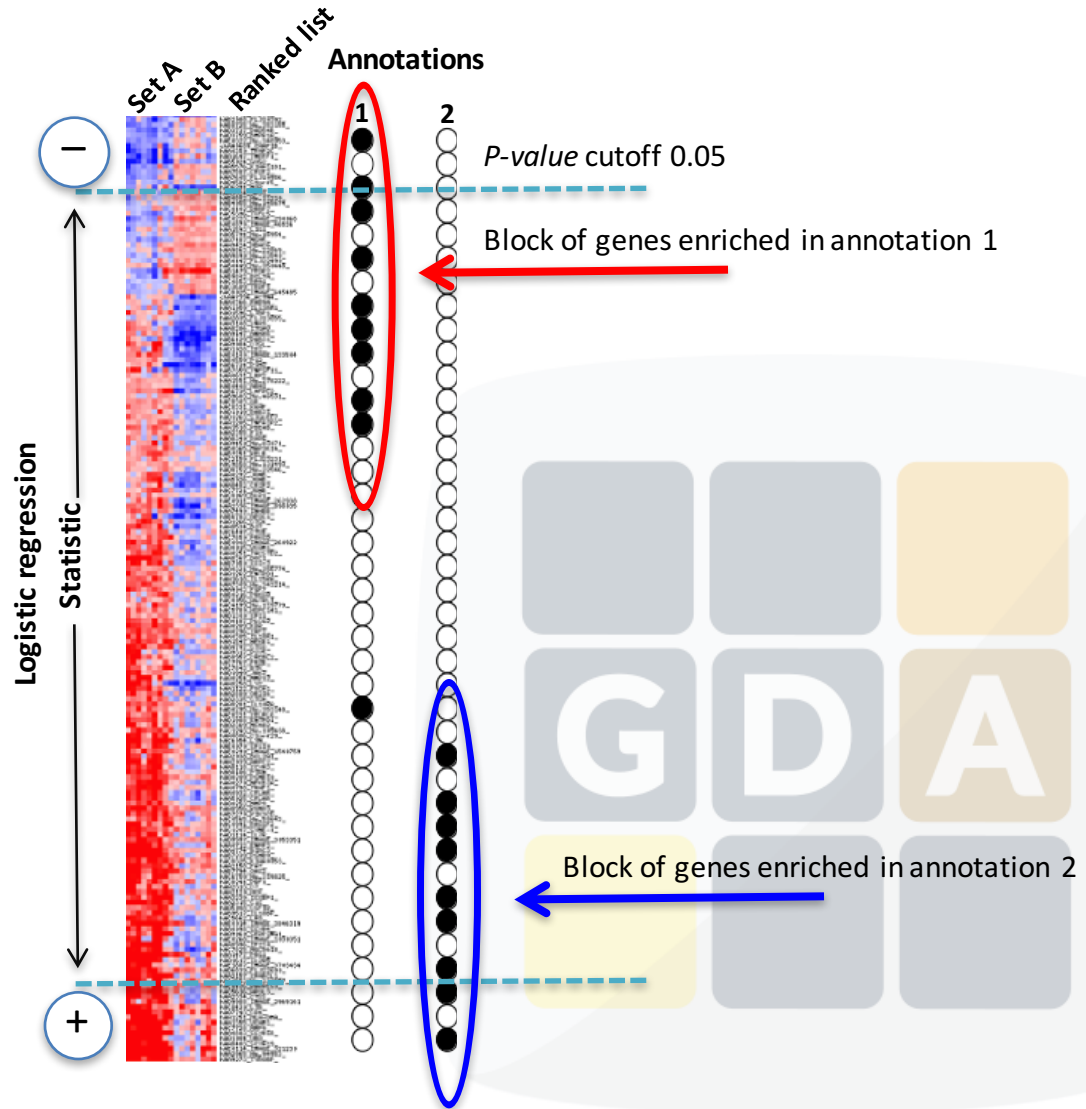
BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 13 2020, pages 2989–3000 doi:10.1093/bioinformatics/btaa267

Gene expression
Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information
 Fátima Al-Shahrour, Ramón Díaz-Uriarte¹ and Joaquín Dopazo^{1,2}

Multidimensional Gene Set Analysis of Genomic Data
 David Montaner^{1,2}, Joaquín Dopazo^{1,2,3,4}

Abstract
 Understanding the functional implications of changes in gene expression, mutations, etc., is the aim of most genomic experiments. To achieve this, several functional profiling methods have been proposed. Such methods study the behaviour of different gene modules (i.e. gene ontology terms) in response to one particular variable (e.g. differential gene expression). In spite of the wealth of information provided by functional profiling methods, a common limitation to all of them is their inherent unidimensional nature. In order to overcome this restriction we present a multidimensional logistic model that allows studying the relationship of gene modules with different genomic scale measurements (e.g. differential expression, genotyping association, methylation, copy number alterations, heterozygosity, etc.) simultaneously. Moreover, the relationship of such functional modules with the interaction among the variables can also be studied, which produces novel results impossible to be derived from the conventional unidimensional functional profiling methods. We report sound results of gene set associations that remained undetected by the conventional one-dimensional gene set analysis in several examples. Our findings demonstrate the potential of the proposed approach for the discovery of new cell functionalities with complex dependencies on more than one variable.

Introduction
 The development of new genomic technologies, such as microarrays of gene expression, genotyping or array-CGH, along with the new microarray expression technologies is increasing the volume of data throughout annually. As a direct consequence of this, the bottleneck in functional genomics has shifted from the data production phase to the data analysis steps. In particular, the necessity for providing a functional interpretation at molecule level data accounts for the genome-wide experimental designs has promoted the development of different methods for the functional analysis of this type of data in the last years [1–5]. It is widely accepted that most of the biological functionality of the cell arises from complex interactions among their molecular components that differ operationally interacting entities or modules [6]. Functions collectively performed by such modules have conceptually been represented in different ways. Gene ontology (GO) [7] and KEGG pathway [8] are the most popular and widely used module definitions although many other are available in different repositories (e.g. Reactome [9], BioCarta, etc.). For practical purposes, functional modules are hierarchically defined as sets of genes sharing functional annotations extracted from any of these repositories. Functional profiling methods exploit different definitions of modules in an attempt of understanding the functional basis of high-throughput experimental results [7] (i.e. bulk), functional enrichment methods, in different implementations [10], have been used for this purpose. More sensitive approaches, generally known as gene set analysis (GSA) methods, pioneered by the Gene Set Enrichment Analysis (GSEA) [5] were later proposed [1, 10]. In the original formulation, GSA methods aimed to identify distinct sets of functionally related genes (modules) with a coordinate and significant over- or under-expression across the complete list of genes ranked according to the differential expression [11, 12, 13, 14, 15]. GSA methods can detect such modules even if their gene components are not significantly differentially expressed when tested individually. GSA has been successfully applied to the analysis of microarray experiments and has contributed to the adaptation of a systems biology perspective in distinct fields such as cancer [16]. Recent findings brought about by the application of GSA methods to microarray experiments [17] are consistent with the idea that pathways, rather than individual genes, appear to govern the course of tumorigenesis [18]. The use of GSA has been extended to other data beyond microarrays, such as methylation [19], QTL analysis [20] or genotyping [21]. Nevertheless, the different variants of GSA published to date [12, 19] are inherently unidimensional. Its application to the analysis of genomic datasets is as present limited to the study of a unique variable measured for the genes. The experimental conditions tested, even if controlled by other variables (e.g. age, gender, treatment, etc.), are typically summarized into a single value for each gene (e.g. differential expression in a case-control, risk in the case of survival analysis, etc.) which is used to rank them accordingly. Moreover, the capture set of different high-throughput methodologies allows the observation of different measurements for the genes such as methylation status, linking between linkage



GSA on Babelomics 5


Tool selection

Functional ▾ dcrespo

- Single enrichment
- ▶ Fatigo
- Gene set enrichment
- ▶ Logistic model
- Network enrichment
- ▶ Snow
- Gene set network enrichment
- ▶ Network Miner

Data selection

Select your data

The files must be on the server to select them.
You can upload files using the button  inside file browser.

Remove duplicates


Databases

✓ Select an organism

- Human (homo sapiens)
- Mouse (mus musculus)
- Rat (rattus norvegicus)
- Fruitfly (drosophila melanogaster)
- Zebrafish (danio rerio)
- Saccharomyces cerevisiae
- Caenorhabditis elegans
- Arabidopsis thaliana


Databases

Select an organism ▾

- GO biological process
- GO molecular function
- GO cellular component
- GOSlim GOA
- Interpro
- Genome-Scale Metabolic Network
- Your annotations
The files must be on the server to select them.
You can upload files using the button  inside file browser.

Job information

Job information

Output folder
You can create folders using the button  + inside file browser.

✕

Job name

Description

Data input

Gene list: bioformat "ID"

```
MBP-de-Dan-2: data_dan
DYNLT1 14.29
RAB13 9.42
COMMD6 7.3
RPL35 6.99
ZNF124 6.74
NDUFA4 6.74
COX7A2 6.67
UBR2 6.66
YWHAQ 6.65
JAK1 6.65
USMG5 6.51
NDUFA12 6.44
AKT3 6.43
SARS 6.31
STRN 6.28
STK4 6.24
COX6A1 6.14
PIK3R4 6.08
NDUFS4 6.08
FBXO38 6.05
VPS8 6.01
CASP4 5.99
SPATS2 5.97
IL6ST 5.93
```

Column 1: Ranked Gene ID
Column 2: Rank

Custom annotation: bioformat Gene vs annotation

```
MBP-de-Dan-2: example dataocresbe$ head -30 ../annotation
"rno04014_47 49" "plasma membrane"
"rno04014_47 49" "positive regulation of cell prolifer
"rno04014_47 49" "extrinsic component of cytoplasmic s
"rno04014_47 49" "GTP binding"
"rno04014_47 49" "Golgi apparatus"
"rno04014_47 49" "positive regulation of angiogenesis"
"rno04014_47 49" "molecular_function"
"rno04014_47 49" "biological_process"
"rno04014_47 49" "cellular_component"
"rno04014_42" "lipid catabolic process"
"rno04014_42" "phospholipase A2 activity"
"rno04014_42" "molecular_function"
"rno04014_42" "mitochondrial inner membrane"
"rno04014_42" "cellular response to insulin stimulus"
"rno04014_42" "glucose transport"
"rno04014_43" "RNA polymerase II transcription factor
"rno04014_43" "transcription from RNA polymerase II pr
"rno04014_43" "cell differentiation"
"rno04014_44" "nucleus"
"rno04014_44" "RNA polymerase II transcription factor
"rno04014_44" "transcription from RNA polymerase II pr
"rno04014_44" "transcription from RNA polymerase II pr
```

Column 1: Gene ID
Column 2: Function

Reading results

Analysis overview

Job information

Name: [fluorouridine_dataset](#)
 Description: [fluorouridine_dataset](#)
 Tool: [fatiscan](#)
 Output folder: [WorkSpace/analysis/20151120155138/](#)

Input data

Ranked list [ranked_list.txt](#)
 Id list (sorted) [id_list.txt](#)
 Statistic (sorted) [statistic.txt](#)

Summary

Id annotations per DB [annotations_per_db.txt](#)

#DB	Number of annotations
GO biological process propagated	706 of 1000 (70.6%) 59.11 annotations/id
1 Results < 1 of 1 >	

Results

Significant Results

Number of significant terms per DB [significant_count_0.05.txt](#)

Select p-value: 0.1 0.05 0.01 0.005

#DB	N° of significant terms
GO biological process propagated	209
1 Results < 1 of 1 >	

GO biological process propagated

GO biological process propagated significant terms (pvalue<0.05) [significant_go_biological_process_propagated_0.05.txt](#)

Term	Term size	Term size(in genome)	annotated_genes lists	converged ids list	lor	adj_pvalue
regulation of myeloid cell differentiation(GO:0045637)	15	355	OGT CSF1 PTK2B ATP6AP1 MEIS2 CCNA1	true	0.28	3.2e-6
mitotic cell cycle(GO:000278)	51	2418	RCC1 GORASP1 NLDC GAS1 PTK2B MEIS2	true	0.17	7.3e-6
negative regulation of myeloid cell differentiation(GO:0045638)	7	148	INPP4B INPP5D TLR4 CCNA1	true	0.36	1.8e-5
mitotic cell cycle process(GO:1903047)	40	2106	RCC1 NLDC RPS6 NUP205 DLC1	true	0.17	3.6e-5
single-organism organelle organization(GO:1902589)	91	4963	OGT CCNA1 CRYAB RCC1 OGT	true	0.13	4.2e-5
regulation of leukocyte differentiation(GO:1902105)	21	435	IL4R IL12A CSF1 ATP6AP1 RPL19	true	0.21	9.9e-5
RNA metabolic process(GO:0016070)	197	13296	HNRPF HOXD11 AGTR2 RPL11 IL4R	true	0.095	1.5e-4
negative regulation of leukocyte differentiation(GO:1902106)	9	148	INPP4B INPP5D RPL19	true	0.28	1.9e-4

Download results

All results

GO biological process propagated [go_biological_process_propagated.txt](#)

Annotation files

Annotations for GO biological process propagated [go_biological_process_propagated_annot](#)

Reading results

Analysis overview

Job information

Name: [fluorouridine_dataset](#)
 Description: [fluorouridine_dataset](#)
 Tool: [fatiscan](#)
 Output folder: [WorkSpace/analysis/20151120155138/](#)

Input data

Ranked list [ranked_list.txt](#)
 Id list (sorted) [id_list.txt](#)
 Statistic (sorted) [statistic.txt](#)

Summary

Id annotations per DB [annotations_per_db.txt](#)

#DB	Number of annotations
GO biological process propagated	706 of 1000 (70.6%) 59.11 annotations/id
1 Results < 1 of 1 >	

Results

Significant Results

Number of significant terms per DB [significant_count_0.05.txt](#)

Select p-value: 0.1 0.05 0.01 0.005

#DB	N° of significant terms
GO biological process propagated	209
1 Results < 1 of 1 >	

GO biological process propagated

GO biological process propagated significant terms (pvalue<0.05) [significant_go_biological_process_propagated_0.05.txt](#)

Term	Term size	Term size(in genome)	annotated_genes lists	converged ids list	Ior	adj_pvalue
regulation of myeloid cell differentiation(GO:0045637)	15	355	OGT CSF1 PTK2B ATP6AP1 MEIS2 CCNA1 RCC1	true	0.28	3.2e-6
mitotic cell cycle(GO:000278)	51	2418	GORASP1 NLUDC GAS1 PTK2B MEIS2	true	0.17	7.3e-6
negative regulation of myeloid cell differentiation(GO:0045638)	7	148	INPP4B INPP5D TLR4 CCNA1 RCC1 NLUDC	true	0.36	1.8e-5
mitotic cell cycle(GO:000278)	40	2106	RPS6 NUP205 DLC1 OGT CCNA1 CRYAB RCC1 OGT	true	0.17	3.6e-5
single-organism organelle organization(GO:1902589)	91	4963	IL4R IL12A CSF1 ATP6AP1 RPL19 HNRPF HOXD11 AGTR2 RPL11 IL4R INPP4B INPP5D	true	0.13	4.2e-5
regulation of leukocyte differentiation(GO:1902105)	21	435		true	0.21	9.9e-5
RNA metabolic process(GO:0016070)	197	13296		true	0.095	1.5e-4
negative regulation of leukocyte differentiation(GO:1902106)	9	148		true	0.28	1.9e-4

Enriched class

Annotated genes list

False Discovery Rate

Download results

All results

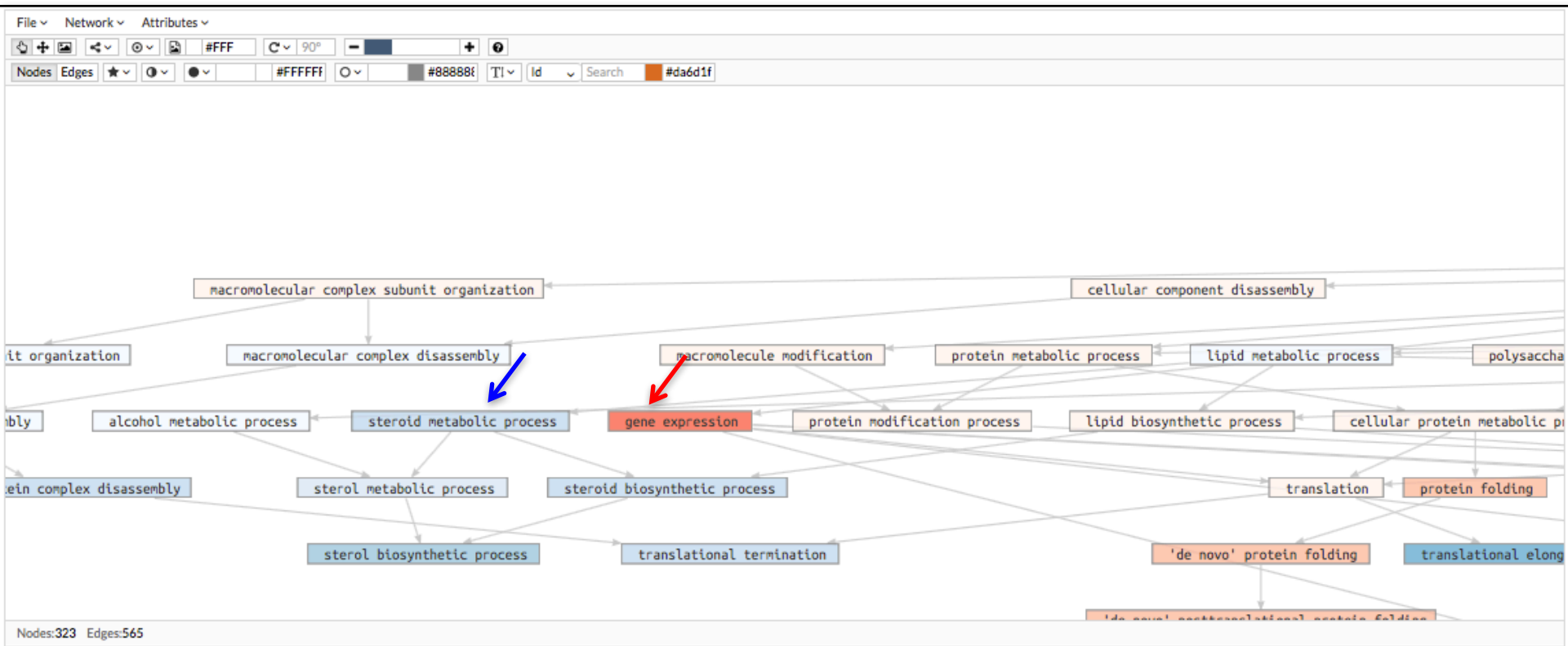
GO biological process propagated [go_biological_process_propagated.txt](#)

Annotation files

Annotations for GO biological process propagated [go_biological_process_propagated_annot](#)

Reading results

Results visualization



Keypoints

- **FatiGO:** Detects significant over representation of functional annotations in one gene set respect to the other one.



Keypoints

- **FatiGO:** Detects significant over representation of functional annotations in one gene set respect to the other one.
 - Critical step: Gene list selection. Typically from differential gene expression.



Keypoints

- **FatiGO:** Detects significant over representation of functional annotations in one gene set respect to the other one.
 - Critical step: Gene list selection. Typically from differential gene expression.
- **GSA:** Detects gene sets (functional annotations) that are consistently associated to high or low values in a ranked list of genes.

Training data and resources



<http://courses.babelomics.org>
<http://babelomics.org>

FatiGO

<http://www.ncbi.nlm.nih.gov/pubmed/17478504>
<http://www.ncbi.nlm.nih.gov/pubmed/14990455>
wiki: <https://github.com/babelomics/babelomics/wiki/Single%20Enrichment>

Gene set analysis: Logistic model

<http://www.ncbi.nlm.nih.gov/pubmed/20436964>
<http://www.ncbi.nlm.nih.gov/pubmed/17407596>
<http://www.ncbi.nlm.nih.gov/pubmed/15840702>
wiki: <https://github.com/babelomics/babelomics/wiki/Gene%20Set%20Enrichment>

Any questions?

User:	gda16	gda16b	gda16c
Password:	gda16	gda16b	gda16c



Thank you very much for your attention!

Systems genomics team



PRINCIPE FELIPE
CENTRO DE INVESTIGACION

