Pathway results from the chicken data set using GOTM, Pathway studio and ingenuity softwares

Agnès Bonnet, Sandrine Lagarrigue, Laurence Liaubet, Christèle Robert Robert-Granié, Magali San Cristobal, Gwenola Tosser-Klopp

BMC Proceedings

BioMed Central

Research

Pathway results from the chicken data set using GOTM, Pathway Studio and Ingenuity softwares

Agnès Bonnet[1], Sandrine Lagarrigue[2,3], Laurence Liaubet[1], Christèle Robert-Granié[4], Magali SanCristobal[1] and Gwenola Tosser-Klopp*[1]

Address: [1]INRA, UMR444, Laboratoire de Génétique Cellulaire, F-31326 Castanet-Tolosan, France, [2]INRA, UMR 598, Génétique Animale, F-35000 Rennes, France, [3]Agrocampus Ouest, UMR 598 Génétique Animale, F-35000 Rennes, France and [4]INRA, UR631, Station d'Amélioration Génétique des Animaux, F-31326 Castanet-Tolosan, France

Email: Agnès Bonnet - agnes.bonnet@toulouse.inra.fr; Sandrine Lagarrigue - sandrine.lagarrigue@agrocampus-ouest.fr; Laurence Liaubet - Laurence.Liaubet@toulouse.inra.fr; Christèle Robert-Granié - Christele.Robert-Granie@toulouse.inra.fr; Magali SanCristobal - Magali.San-Cristobal@toulouse.inra.fr; Gwenola Tosser-Klopp* - Gwenola.Tosser@toulouse.inra.fr

* Corresponding author

from EADGENE and SABRE Post-analyses Workshop
Lelystad, The Netherlands. 12–14 November 2008

Published: 16 July 2009

BMC Proceedings 2009, **3**(Suppl 4):S11 doi:10.1186/1753-6561-3-S4-S11

This article is available from: http://www.biomedcentral.com/1753-6561/3/S4/S11

Abstract

Background: As presented in the introduction paper, three sets of differentially regulated genes were found after the analysis of the chicken infection data set from EADGENE. Different methods were used to interpret these results.

Results: GOTM, Pathway Studio and Ingenuity softwares were used to investigate the three lists of genes. The three softwares allowed the analysis of the data and highlighted different networks. However, only one set of genes, showing a differential expression between primary and secondary response gave significant biological interpretation.

Conclusion: Combining these databases that were developed independently on different annotation sources supplies a useful tool for a global biological interpretation of microarray data, even if they may contain some imperfections (e.g. gene not or not well annotated).

Background

Microarray data usually provide lists of genes, and often of hundreds of genes. The challenge is next to give sense to these lists and help to interpret the results. We used the three lists of differentially expressed genes presented in the introductive paper [1] as examples of the way to do such analyses. They address three different biological questions: primary or secondary response effect (MM8-PM8, 1736 genes), species effect (MM8-MA8, 85 genes) and time effect (MM8-MM24, 800 genes).

The biological interpretation of gene lists was performed by three approaches: study of GO enrichment terms with GOTM (Gene Ontology Tree Machine) and two pathway software with a demonstration version of Pathway Studio (Ariadne Genomics) and Ingenuity Pathway Analysis (IPA, Ingenuity Systems Inc., Redwood City, CA). These

approaches used various gene annotation databases: 1-Gene Ontology (GO) describes the most important ontologies in molecular biology, 2-Pathway Studio database is curated electronically using automated text-mining engines that are also available to investigators and, 3-Ingenuity's knowledge base is the first large database developed on all types of gene-gene and metabolite-gene interrelationships extracted manually from publications.

The purpose of our study is to evaluate three different *in silico* approaches in order to give sense to gene lists from microarray data.

Methods

GOTM

Gene ontology (GO) [2] constitutes a controlled vocabulary of about 20,000 terms organized in three independent hierarchies for cellular components, molecular functions, and biological processes [3]. The GO analyses of clusters were performed using the Gene Ontology Tree Machine (GOTM) software [4]. The GOTM web-based tool supplies statistical analysis to identify enriched Gene Ontology categories for the input gene sets and generates a GOTree, a tree-like structure to navigate the Gene Ontology Directed Acyclic Graph to help users visualize the GO-term relations. Hypergeometric test was used to select enriched GO terms for each cluster compared to the GO terms of the annotated genes present on the microarray (10451 genes, in this study). A GO category was considered as enriched for a level of p-value < 0.01.

Pathway Studio

We run the "Subnetwork Enrichment Analysis" (SNEA) option of the software, with default parameters, which builds small networks consisting of a single "regulator" gene and its targets [5]. The significance of the target expression levels in every built network is evaluated next. The algorithm finds the individual "regulators" which most likely affect differentially expressed targets. Thus, it is expected to provide the most plausible explanation for the observed expression changes. We then built the union of pathways from the selected networks.

Ingenuity Pathways Analysis

This system, a web-based interface [6], provides computational algorithms to identify and dynamically generate significant biological networks and pathways that are particularly enriched with our genes of interest called "focus genes". It also ranks networks by a score that takes into account the number of focus genes and the size of the networks, indicating the likelihood of the focus genes in a network being found together by chance. The higher the score (score = -log(p-value)), the lower is the probability of finding the observed Network Eligible Molecules in a given network by chance. IPA also gives information on biological functions and canonical pathways. The data set contains the three differentially expressed gene lists and the microarray gene list. It includes a column with mixed identifiers (HUGO symbols, REFSEQ...) and a column with gene fold changes. This data set was uploaded as a tab-delimited text file. We compared the 3 gene lists to the list of genes of the microarray and underlined the specific enrichment of regulated genes.

Results

GOTM

Among the 11538 oligonucleotides of the microarray with a proposed HUGO name, 10451 HUGO gene names were validated according to HGNC [7]. The HUGO gene name is a unique human ortholog and an abbreviation name provided by the annotation EADGENE Network of Excellence, funded by the EC and available on [8].

For the 3 condition contrasts MM8-MA8, MM8-MM24 and MM8-PM8, the gene numbers with validated HUGO among the differentially expressed genes were 38 among 85, 354 among 800 and 931 among 1736 respectively. Because GOTM does not use more than 500 genes at once, we had to split the MM8-PM8 list into 2 sets: up and down regulated genes (Table 1). Only the biological process GO category was reported in the present study (p-value < 0.01). As indicated in Table 2, six biological processes were enriched in MM8-MM24 and MM8-MA8 sets. Concerning the MM8-PM8 experimental condition, six and seven biological processes were enriched for the down-regulated and up-regulated gene sets respectively.

Pathway Studio

For the 3 condition contrasts MM8-MA8, MM8-MM24 and MM8-PM8, the gene numbers mapped with the software were 42 among 85, 382 among 800 and 1037 among 1736 respectively (Table 1).

Table 1: Number of validated identifiers with the three softwares

	MM8-MA8	MM8-MM24	MM8-PM8
Number of validated identifiers	85	800	1736
GOTM	38	354	931
Pathway Studio	42	382	1037
Ingenuity	43	386	1062

Table 2: Enriched biological process GO terms (P < 0.01) obtained by the Gene Ontology Tree Machine (GOTM).

Enriched biological process GO terms	P < 0.01	Time effect	Species effect	Injection effect
organelle organization and biogenesis: ARPC3 KATNB1 DCTN6 KIF3A JMJD2A HDAC4 MTSS1 CBX3 ADRB2 TIMM8A GOLGB1 RHOA RHOB NEDD9 SPTBN5 PEX1 H2AFY2 FMN2 RHOJ EP400 NSD1 SOX2 BRCA2 YWHAH DDX54 NOC4L ATG9A SUV39H2 SMC3 CDC42BPB	P = 0.005	MM8-MM24	MM8-MA8	
gluconeogenesis: ATF4 ACN9 TPI1	P = 0.008			
glutamine biosynthesis: GLUL CORO2A	P = 0.005			
nervous system development: SPON2 NCKAP1 CHRNA4 CNTF DAB1 TIMM8A EFNB1 NEUROG1 PAX5 RPS6KA3 YWHAH SEMA6D ST8SIA2 PARD6B LGI1 NOG CD9 ECE2 HDAC4 MTSS1	P = 0.005	MM8-MM24		
organelle organization and biogenesis: NSD1 L3MBTL DST NEDD9 SPTBN5 PFN2	P = 0.009		MM8-MA8	
pyruvate metabolism: ATF4 TPI1	P = 0.001			
G-protein signaling\, coupled to IP3 second messenger phospholipase C activating: NMUR1 DGKG P2RY4 AVPR1A DGKZ SPHK1	P = 0.008			
UDP-N-acetylglucosamine metabolism: UAP1 GNE	P = 0.006			
negative regulation of DNA replication: ENPP7 S100A11 CDT1	P = 0.0009			MM8-PM8
endocytosis: ADRB2 AP1B1 DOCK1 PACSIN3 PACSIN1 LRP6 NECAP2 ELMO3 SNX4	P = 0.0047			genes down-regulated
tissue development: APOA5 EDA ELF3 LAMB3 LAMC2 BMP4 SOX9 TFAP2A TGFB2 TUFT1	P = 0.004			
ectoderm development EDA ELF3 LAMB3 LAMC2 TFAP2A TGFB2	P = 0.003			
immune response-regulating signal transduction: FYN SPG21 PTPRC	P = 0.003			
caspase activation: DIABLO STAT1 CASP9 CASP8AP2	P = 0.004			
Golgi vesicle transport: SEC23A EXOC5 TMED2 COPE COPB1 SEC22A SAR1A MCFD2 COPB2 GOSR1	P = 0.006			MM8-PM8
protein targeting to mitochondrion: TIMM44 TOMM34 TSPO GRPEL1	P = 0.008			genes up-regulated
chromatin assembly or disassembly: CBX3 HMGB2 HP1BP3 HDAC8 SMARCE1 CHAF1B SMARCA5 TLK1	P = 0.002			
modification-dependent protein catabolism: FBXO21 RNF11 PSMA2 PSMA5 PSMB1 PSMB3 PSMC5 UBE2E1 UBE3A USP7 PSMF1 USP3	P = 0.008			
protein transport: 34 genes	P = 0.002			

For each set of data, the enriched biological processes found with GOTM analysis are listed, with their probability and representative genes.

The SNEA analysis gave significant results only for the MM8-PM8 gene list. Expression targets of JUN, CD8A, IL13 and SP1 were found as significantly over-represented (p-value <0.05). The combined pathway is represented in Figure 1. The software allowed to visualise the regulation of the target genes in this combined pathway, using the two other lists of genes (MM8-MA8) and (MM8-MM24) and showed 1 to 4 regulated targets (data not shown).

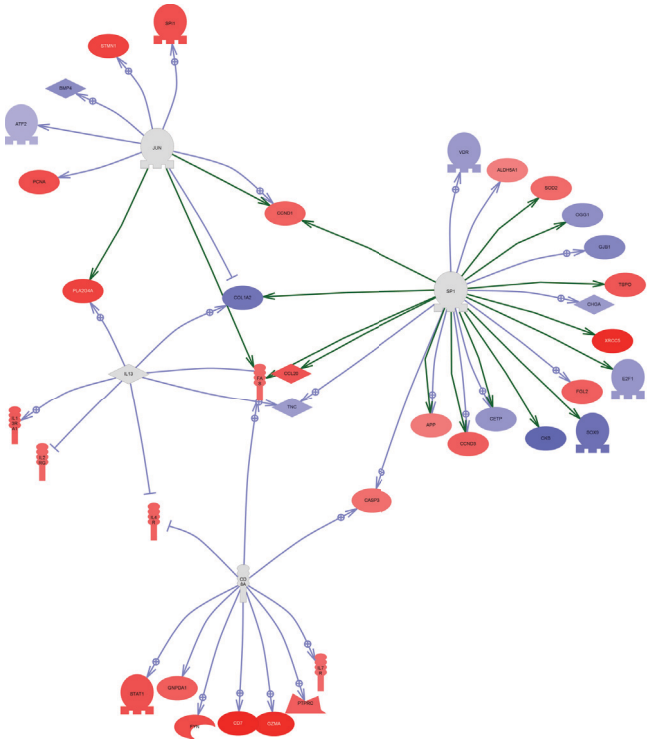

Figure 1
Combined pathway from the Sub Network Enrichment Analysis of Pathway Studio (primary or secondary response effect). Molecules are identified with their HUGO symbol. Red colour shows up regulated genes in the PM8 (primary response) condition versus MM8 (secondary response). Blue colour shows down regulated genes in the PM8 (primary response) condition versus MM8 (secondary response).

Ingenuity

Among the 13639 annotated genes from the microarray using the different identifiers, 12395 are recognized by the software and 7219 used to generate networks. For the 3 condition contrasts MM8-MA8, MM8-MM24 and MM8-PM8, the gene numbers mapped with Pathway Studio were 43 among 85, 396 among 800 and 1062 among 1736 respectively (Table 1).

First, we searched to identify the networks, functions or canonical pathways from the microarray gene list. One hundred networks are built from the whole microarray with a maximum score of 18. This score obtained for the microarray gene list is considered as a threshold and is used as the minimum score accepted to define a network for the next 3 gene list as significantly enriched.

Then, Ingenuity gives 22 significant enriched networks from the 3 gene lists with a score between 19 and 55 (Additional datafile 1). The enriched functions and canonical pathways were presented in Additional datafile 2.

Ingenuity underlines:

a) In the PM8-MM8 comparison, 14 networks are involved in significant biological functions compared to the microarray background. Biological functions as immune and inflammatory response, metabolism and proliferation are enriched. Particularly, the merged immunity network shows the down expression of a majority of genes during the secondary response. In accordance with previous results [9], canonical pathway shows specifically the down expression of genes involved in inflammatory cytokines signalisation as interleukin 4 and interferon signalisation during the secondary infection (Figure 2).

b) In the MM8-MM24 comparison, 3 networks (out of 7) and the canonical pathways results show mostly a down expression of the genes involved in the actin polymerisation.

c) In the MM8-MA8 comparison, only 1 significant network (score 22) was identified with down expressed genes involved in inflammatory diseases in *Eimeria acervulina* species compared to *Eimeria maxima*.

Conclusion

In this study, we used 3 lists of genes with 3 softwares. It is noticeable that slight differences (+/-10%) exist between the 3 softwares concerning the number of mapped genes. We identified two limits, using these programs: first, GOTM did not allow us to study all the genes at once; second, whatever program we used, the number of genes with an accepted name never reached 100% of the annotated list. The capability of the software to identify the gene names is a limiting step of the analysis.

Only primary vs. secondary response gene list (MM8-PM8) gave significant biological interpretation with the 3 softwares: immune function and inflammatory response were much higher in primary response (immune-system related processes).

Because these databases were developed on different annotation sources, the results can be considered as complementary. Combining these databases supplies a useful tool for a global biological interpretation of microarray data, even if they may contain some imperfections (gene not or not well annotated).

In another way, GOTM offers more accurate information for a list of genes but did not suggest any links between these genes. GOTM is more appropriate for description of the data with a statistical validation.

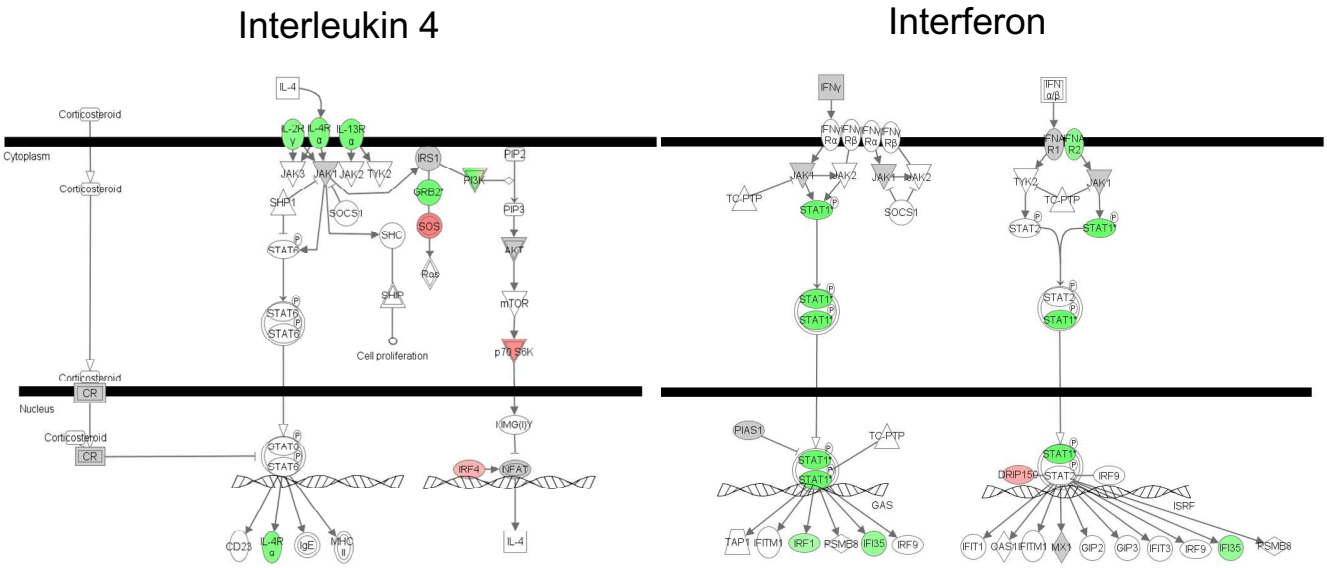

Figure 2
Canonical pathway underlines genes involved in inflammatory cytokines signalisation from the primary/secondary response comparison (PM8-MM8). We compared the secondary infection to the single ones. Green colour shows down regulated genes in the MM8 (secondary response) condition versus PM8 (primary response). Red colour shows up regulated genes in the MM8 (secondary response) condition versus PM8 (primary response).

Pathway Studio, in the way it was used in this work, presents two main characteristics that differentiate it from the other softwares. First, it tries to find links only between the genes in the proposed list and not in a more global context like Ingenuity. Secondly, it offers the possibility to identify major regulators for differentially regulated genes and this could be interesting in order to develop further studies. The proposed networks look like stars with the possible regulator in the node.

Ingenuity suggests more possible links between the differentially regulated genes and with some other genes in highly relevant networks. The networks are more complex than with Pathway Studio with more nodes, more genes without information about their expression. Ingenuity explores list of genes by very different ways, for example as very high confident information with canonical pathways and more exploratory ways as gene network.

These softwares give tools to functionally annotate the gene lists and help the biologists to give sense to huge data. Finally, these softwares offer the possibility to identify more interesting genes among a list in order to undertake further experiments.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
AB carried out the Ingenuity study, SL carried out the GOTM study. GTK carried out the Pathway Studio study and drafted the manuscript. All authors participated to a working group, the redaction work, read and approved the final manuscript.

Additional material

Additional file 1
***IPA networks**. For each of the 3 lists the networks are selected if their score is higher than 21 (the higher network score value generated by all the microarray genes (score >18). The table contains columns with the network number, the name of the comparison list, the names of the genes involved in the network, the score value and the top functions.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1753-6561-3-S4-S11-S1.doc]

Additional file 2

Data were filtered with 2 criterions: IPA threshold p.value (<0.05) and the corresponding microarray p.value. a – Biofunctions: we did not find significant functions for the MM8-MA8 gene list. b – Canonical Pathway. For each list, the pathways are ranked by the score (score = -log(p.value)) using the same criterions. The table includes also the ratio (number of focus molecules in a given pathway divided by the total number of the molecules that makes up that pathway).
Click here for file
[http://www.biomedcentral.com/content/supplementary/1753-6561-3-S4-S11-S2.doc]

Acknowledgements

Thanks are expressed to Fabrice Laurent (INRA Tours-Nouzilly, France) for his interesting discussion.

This article has been published as part of *BMC Proceedings* Volume 3 Supplement 4, 2009: EADGENE and SABRE Post-analyses Workshop. The full contents of the supplement are available online at http://www.biomedcentral.com/1753-6561/3?issue=S4.

References

1. Hedegaard J, Bicciato S, Bonnet A, Ramirez Boo M, Buitenhuis B, Collado-Romero M, Nagstrup Conley L, San Cristobal M, Ferrar iF, Groenen M, Hornshøj H, Hulsegge I, Jiang L, Arce Jiménez C, Jiménez-Marín A, Kommadath A, Lagarrigue S, Leunissen J, Liaubet L, Neerincx P, Nie H, Garrido Pavón J, Prickett D, Rebel J, Robert-Granié C, Skarman A, Smits M, Sørensen P, Tosser-Klopp G, Poel J van der, Watson M: **Methods for interpreting lists of affected genes obtained in a DNA microarray experiment.** *BMC Proceedings* 2009, **3(Suppl 4):**S5.
2. **the Gene Ontology** [http://www.geneontology.org/]
3. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25:**25-29.
4. Zhang B, Schmoyer D, Kirov S, Snoddy J: **GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies.** *BMC Bioinformatics* 2004, **5:**16.
5. **Ariadne** [http://www.ariadnegenomics.com/products/pathway-studio/expression-analysis/algorithms/]
6. **Ingenuity Systems** [http://www.ingenuity.com]
7. **HUGO Gene Nomenclature Committee** [http://www.genenames.org/]
8. **EADGENE** [http://www.eadgene.info/TheProject/IntegratioBiologicalresourcesandfacilitiesWP11/EADGENEOligoSetsAnnota tionFiles/tabid/324/Default.aspx]
9. Kim CH, Lillehoj HS, Hong YH, Keeler CL Jr: **Comparison of transcriptional changes associated with E. acervulina and E. maxima infections using cDNA microarray technology.** *Dev Biol (Basel)* 2008, **132:**121-130.