A Functional and Regulatory Map of Asthma

Noa Novershtern, Zohar Itzhaki, Ohad Manor,

Nir Friedman, and Naftali Kaminski

Online Data Supplement

Supplementary Information

Compendium Data Selection

We started to build the compendium having DEA dataset (1). We searched NCBI Gene Expression Omnibus (GEO) for other in-vivo asthma murine models. Four datasets (as true for time of analysis - August 2005) passed our inclusion criterion and were added to the compendium. The inclusion criterion requires a dataset to have at least two biological replicates for each treatment and at least 85% present calls. The list of all murine asthma-related models which can be found in GEO, along with the criteria for inclusion and exclusion, is available in Supplementary Table E4.

Gene Set Experimental Data Selection

To build the experimental gene sets we looked for human gene expression that was measured in various lung cell types. The initial set included data that was generated by Kaminski et. al. (2-4). We then searched GEO for all human asthma models. Four datasets that were available at the time of analysis were chosen to generate the gene sets HAH, HCL, HBE and HAE. See supplementary table E5 for the dataset details.

Module Network Evaluation

To evaluate the quality of the module network we examined the nature of the splits in the regulation program. A good split distinguishes between two coherent groups of experiments, such as treatment vs. control, or between sub-populations of such groups. We therefore measured for each split, how well it distinguishes between each possible classification of the experiments, using the mutual information score (5). Briefly, the mutual information between a given pair of random variables *X* and *Y* measures how knowing one of variable reduces the uncertainty about the other. Formally, the mutual information is defined as the following:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p(x)\,p(y)} \right),$$

If X and Y are independent, then knowing X does not give any information about Y and vice versa, and the mutual information is zero. At the other extreme, if X and Y are identical then knowing X determines the value of Y (and vice versa), and the mutual information is one.

We calculated the mutual information for each split and classification by representing them as two binary random variables. The first denotes the classification of experiments according to the split, and the second according to their attributes (e.g. treatment or control). If the split carries information about the classification samples, its mutual information will be high (Figure E1).

Figure E2 shows the percentage of splits that separate (mutual information > 0.5) between different classes of experiments. The network inferred from the data excluding FTM dataset, contains 318 splits, of which 44% are informative about treatment vs. control. However, when the network inferred from data including the FTM dataset, only 17% of the 473 splits distinguish between treatment and control, and 24% distinguish between the FTM tissue subgroups (Whole Lung and TP), or between the whole FTM dataset and the rest of the datasets. The FTM dataset therefore largely determines the structure of the network, and introduce a bias which is not relevant to the rest of the data. The strong bias can be explained by the substantial difference between TP and WL, which masks the other signals in the data, and also by the fact that FTM data were generated on two-channel UCSF platform, whereas the rest of the data are affymetrix single-channel hybridizations.
As a conclusion, we have decided to exclude the FTM dataset from the module network analysis, as presented in the paper.

Legends for supplementary figures

Figure E1: The average Bayesian Score per gene, as a function of module number, inferred with Module Network algorithm. The model with the highest Bayesian score has 61 modules.

Figure E2: The mutual information score indicates how informative one random variable (e.g. the split) on another variable (e.g. the sample classification). In one extreme (Figure 2E-A) the split does not carry any information about the classification - the "treatment" samples are evenly distributed between the two split sides and the mutual information is zero. At the other extreme (Figure 2E-B) the split fully corresponds to the classification, and the mutual information is one. If there is some correspondence between the split and the classification, the mutual information will be between 0 and 1 (Figure 2E-C).

Percentage of splits with substantial mutual information (> 0.5) is shown in 2E-D for various sample classifications. Split that separates coherent groups of samples has high mutual information. Separation between treatments or strain types is biologically meaningful, whereas separation between datasets indicates a technical artifact. The informative split percentage is shown for two compendiums – one that includes FTM dataset (grey), and one that excludes FTM dataset (black).

References

1. Zimmermann, N., N. E. King, J. Laporte, M. Yang, A. Mishra, S. M. Pope, E. E. Muntel, D. P. Witte, A. A. Pegg, P. S. Foster, Q. Hamid, and M. E. Rothenberg. 2003. Dissection of experimental asthma with DNA microarray analysis identifies arginase in asthma pathogenesis. *J Clin Invest* 111(12):1863-74.
2. Lee, J. H., N. Kaminski, G. Dolganov, G. Grunig, L. Koth, C. Solomon, D. J. Erle, and D. Sheppard. 2001. Interleukin-13 induces dramatically different transcriptional programs in three human airway cell types. *Am J Respir Cell Mol Biol* 25(4):474-85.
3. Gal, N., A. Pardo, Z. Yakhini, C. Becerril, A. Ben-Dor, N. Friedman, I. Ben-Dov, and N. Kaminski. 2002. Gene Expression Analysis Of Lung Fibroblasts Derived From Idiopathic Pulmonary Fibrosis Patients. *Am J Respir Crit Care Med* 165(8):A171.
4. Kaminski, N., J. H. Lee, J. Allard, R. A. Heller, and D. Sheppard. 2000. TGF induces distinct transcriptional programs in airway epithelial and airway smooth muscle cells. *Am J Respir Crit Care Med* 161(3):A667.
5. Shannon, C.E and Weaver W. *The Mathematical Theory of Communication.* Univ of Illinois Press, 1949. ISBN 0-252-72548-4

Figure E1

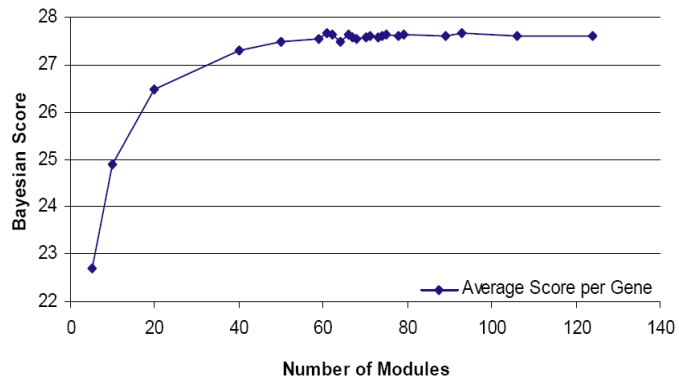
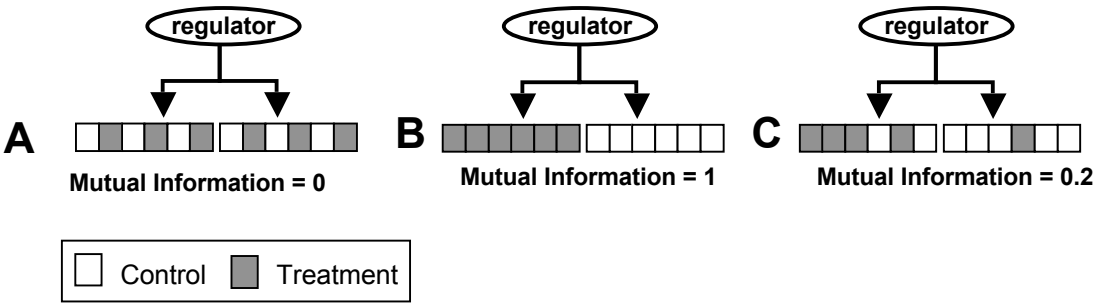
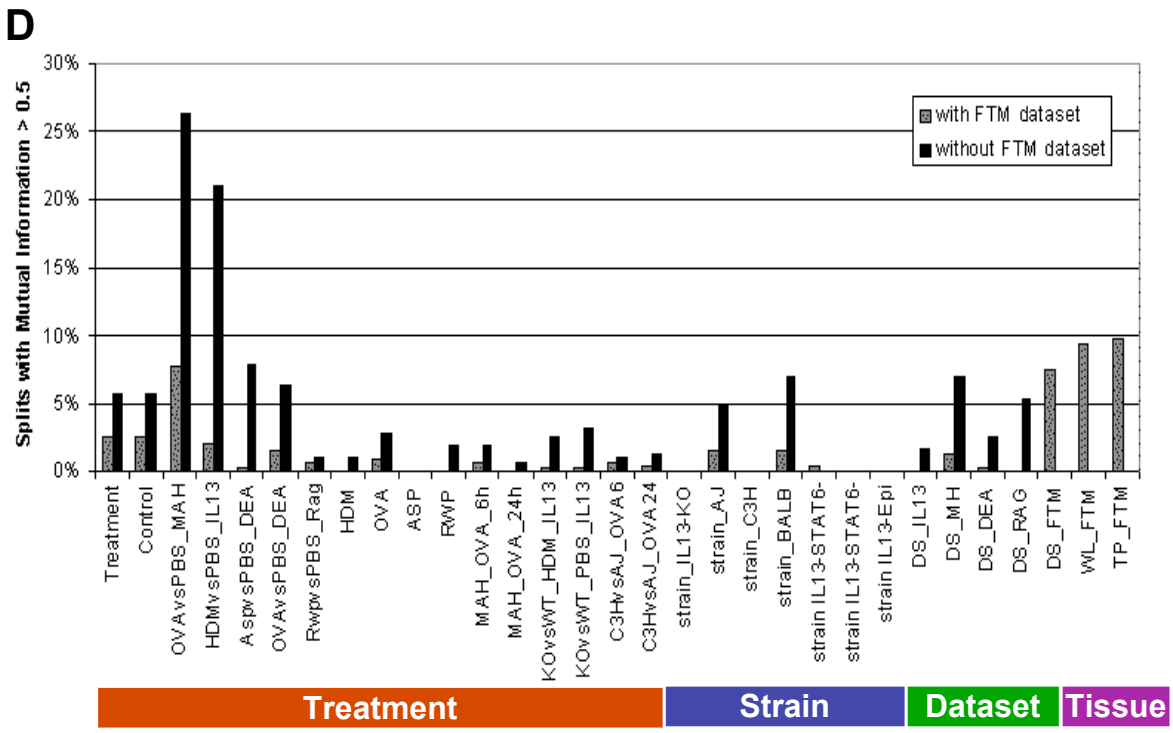

Bayesian Score
28
27
26
25
24
23
22
0
20
40
60
80
100
120
140
Number of Modules
Average Score per Gene

Figure E2


A
regulator
Mutual Information = 0
B
regulator
Mutual Information = 1
C
regulator
Mutual Information = 0.2
Control
Treatment


D
Splits with Mutual Information > 0.5
30%
25%
20%
15%
10%
5%
0%
with FTM dataset
without FTM dataset
Treatment
Control
OVAvsPBS_MAH
HDMvsPBS_IL13
AspwsPBS_DEA
OVAvsPBS_DEA
RwpwsPBS_Rag
HDM
OVA
ASP
RWP
MAH_OVA_6h
MAH_OVA_24h
KOvsWT_HDM_IL13
KOwsWT_PBS_IL13
C3HwsAJ_OVA6
C3HwsAJ_OVA24
strain_IL13-KO
strain_AJ
strain_C3H
strain_BALB
strain IL13-STAT6-
strain IL13-STAT6-
strain IL13-Epi
DS_IL13
DS_MH
DS_DEA
DS_RAG
DS_FTM
WL_FTM
TP_FTM
Treatment
Strain
Dataset
Tissue