Exploratory analysis of biological data

Department of
Physics of Complex Systems

▪ **Prof. Eytan Domany**

Assif Yitzhaky, Yair Horesh,
Noam Shental, Hilah Gal,
Libi Hertzberg, Itai Kela, Tal Shay,
Michal Sheffer, Yuval Tabach,
Amit Zeisel, Or Zuk, Lior Harpaz,
Mark Koudritzky, Temima Schnitzer.

☎ 972 8 934 3964
FAX 972 8 934 4109
@ eytan.domany@weizmann.ac.il
www.weizmann.ac.il/complex/compphys/

We develop exploratory methods and use them for the analysis of high-throughput biological data, to study complex biological processes, especially the onset and progression of human cancer. Gene expression, copy number information, SNP and clinical data are combined together, to enhance our understanding of cancer and to provide insights about the effect of a tumor's molecular profile on its prognosis.

Chromosomal instabilities play key roles in several types of cancer. We studied how DNA copy number (measured by SNP chips or array CGH), and gene expression are related to survival in colorectal cancer, brain tumors and Leukemia. We found that the currently available SNP chip preprocessing algorithms were lacking in precision and reliability, and thus we developed a new allele-specific SNP chip preprocessing algorithm. We also study regulation of gene expression and the deregulation of pathways in various diseases. We were also involved in developing an antigen chip to study the immune system.

The development of colorectal cancer takes decades and requires the accumulation of mutations and other DNA alterations associated with key regulatory genes. Some of these alterations may be due to changes in genome copy number and structure, referred as Chromosomal Instabilities. We use GISTIC (Beroukhim et al., PNAS, 104(50):20007-12), a statistical method that searches for CINons – regions whose DNA copy number is frequently amplified (amplicon) or deleted (deleton), based on both copy number and genotype calls. Next, we focus on regions that contain genes with significant correlation between their DNA copy number and expression level. The CINon map shown on Figure 1 is based on analysis of 122 Affymetrix Xba2 SNP chips, comprising different stages of the disease as well as normal tissue (data collected in the framework of a collaboration with several groups in the US). We find eight amplicons (red) and 67 deletons (blue). Examination of the peak (lowest q-value) in each CINon identifies small regions with potential oncogenes and tumor suppressor genes. Three of these peaks correspond to the known amplified oncogene MYC and deleted tumor suppressors APC and DCC.

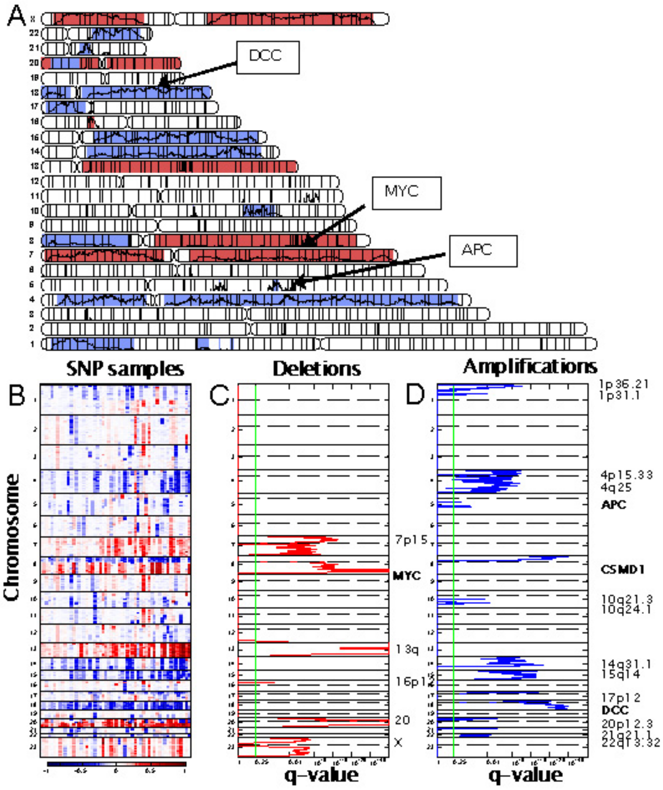

Fig. 1 *CINon map from colon cancer SNP data . (A) Chromosomal view of CINons. The marked red and blue regions correspond to amplifications and deletions that were statistically significant over 56 tumors. The plotted line inside each chromosome corresponds to the –log2(q-values) for each SNP. (B) Colormap of the SNP based DNA copy numbers for the tumors, where each row represents a single SNP, ordered by chromosomal location, and each column represents a tumor. The q-values for amplifications (C) and deletions (D) at each SNP.*

A similar study was performed on glioblastoma, in collaboration with a group at Lausanne. We developed a random model that allows us to use aCGH data to detect significantly amplified and deleted chromosomal regions. We identified known oncogenes (EGFR, PDGFR, CDK4, MDM2) and tumor suppressor genes (PTEN). We applied this model to several other brain tumors, emphasizing the relationships between whole chromosome arm events and local events.

In addition, we developed a method to predict DNA copy number changes from gene expression. Suppose we are given a single sample. We would like to identify whether there are extra chromosomes and which are they. The assumption is that there is a large gene expression dataset of patients that have,

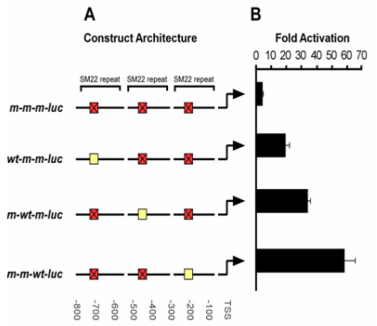

Fig. 2 *Direct experimental demonstration (R. Brosh,Y Buganim, V. Rotter) of the transcriptional effect of the distance of the binding site from the transcription start site. A. We focused on the target gene SM22, whose expression increased significantly in response to Myocardin. We generated a set of four plasmids encoding a luciferase reporter gene (luc) under the regulation of three tandem repeats of a segment from the SM22 promoter, each of which contained either an intact binding site for Myocardin (wt), approximately at its center, or a mutated one (m). B. Fold activation of the luciferase construct, calculated as the ratio of promoter activity in the presence of Myocardin to the promoter activity in the absence of Myocardin. The "m-m-wt-luc" reporter was strongly activated by Myocardin. In contrast, the Myocardin-dependent activation of "m-wt-m-luc" and "wt-m-m-luc" was significantly weaker.*

in the broad sense, the same disease as that of the given sample, to which the single sample can be compared. The basis of the analysis is comparing each gene's expression level with that of the large dataset, by calculating its relative fold change value. When a gene is located on a chromosome which has an extra copy(ies) it generally has an amplified expression level. On the basis of calculating the fold change value for each gene, we can predict which chromosomes (or chromosomal arms) are amplified in a given sample. We applied the method to the St. Jude dataset, that contains 132 Leukemia samples. We have obtained the cytogenetic copy number information for chromosome 21 in each of the 132 samples, and used it to estimate the accuracy of our prediction. The quality of the prediction was measured in a ROC curve that describes the tradeoff between "true positives" and "false positives". The prediction is quite accurate: for false positives rate of 2.9% the true positives rate is 86%.

We have developed a joint copy-number and genotype calling algorithm for SNP chips analysis. SNP chips are a valuable tool becoming widely used to detect variation between different human individuals, yielding new biological insights and genotypes with clinical implications. The SNP chips technology enables the simultaneous measurements of two distinct signals: the genotypes at a given SNP locus and the DNA copy number at the SNPs regions. Previous algorithms and programs for analyzing SNP chips have focused on either one or the other of these signals. We propose a combined, Hidden-Markov-Model based approach, to infer both the genotypes and the copy number of the input samples. We show that our method is more accurate and can detect signals better than previous methods for samples taken from a population of Leukemia and Colon cancer patients. We got a lower error rate in genotype calling, and also were able to detect allele-specific copy number changes and different amplification levels missed by previous methods. We have implemented our algorithm, starting from raw CEL files all the way to graphical outputs generated by the user, implemented in an easy-to-use and user-friendly GUI, enabling browsing, zooming, and exporting images.

We are exploring location bias in functional transcription factor binding sites. Transcription factors regulate expression by binding to specific DNA sequences. A binding event is functional when it affects gene expression. Functionality of a binding site is reflected in conservation of the binding sequence during evolution and in over represented binding in gene groups with coherent biological functions. Functionality is governed by several parameters such as the transcription factor-DNA binding strength, distance of the binding site from the transcription start site and DNA packing. Understanding how these parameters control functionality of different transcription factors in different biological contexts is a must for identifying functional transcription factor binding sites and for understanding regulation of transcription.

We introduced a novel method to screen the promoters of a set of genes with shared biological function against a precompiled library of DNA motifs, and found those motifs for which the registered hits were statistically over-represented in the gene set tested. A hit is registered when the sequence-similarity score of the motif exceeds a threshold; we optimize the value of this threshold independently for every location window, taking into account nucleotide heterogeneity along the promoters of the target genes. The method, combined with binding sequence and location conservation between human and mouse, identifies with high probability functional binding sites for groups of functionally-related genes. We found many location-sensitive functional binding events and showed that they clustered close to the transcription start site. We designed an experiment to measure the transcriptional activity of a transcription factor as a function of the distance between its binding site and the transcription start site (Figure 2). Our analyses indicate that the most prevalent transcriptionally functional mechanisms involve binding in the vicinity of the transcription start site.

By analyzing genome-wide ChIP-on-chip data obtained for several known transcription factors, we discovered that the genes to which they bind split into two groups. In the first group binding has a strong location bias, with a high abundance of binding sites in the first few hundred nucleotides upstream to the transcription start site. The second group contains genes for which the binding sites are uniformly distributed on the first 10,000 base pairs.

Another algorithm we developed is STOP: searching for transcription factor motifs using gene expression. Existing computational methods that identify transcription factor binding sites on a gene's promoter are plagued by significant inaccuracies. Binding of a transcription factor to a particular sequence is assessed by comparing its similarity score, obtained from the transcription factor's known position weight matrix, to a threshold. If the similarity score is above the threshold, the sequence is considered a putative binding site. Determining this threshold is a central part of the problem, for which no satisfactory biologically based solution exists. We developed a method that integrates gene expression data with sequence-based scoring of transcription factor binding sites, for determining a global score threshold for each transcription factor. We validate our method in several ways, including a comparison to mouse genome and gene expression.

In another research project we try to elucidate the molecular background of brain disorders such as Alzheimer's, Parkinson's and Schizophrenia. Our medical collaborators sampled blood and brain tissues from individuals with varying severity of the diseases and from healthy persons. From these samples they extracted data about the expression level of thousands of genes. Analyzing this data presents a two fold challenge; computationally, it is hard to find genes that are over or under expressed among sick individuals because of the overwhelming amount of background noise. The second challenge is to suggest an explanation why are these genes correlated and how can this explanation fit the wider, mostly unknown, picture. Understanding which pathways are affected and play a role in disease onset and progression is a first step towards discovery of effective therapeutic measures. These may lead to slowing disease progression and even to a cure.

We were also involved in development of antigen chips. The immune system is a key player in body maintenance. In order to achieve its tasks, the immune system monitors the body, integrates enormous amounts of information, stores the information in its antibody and lymphocyte repertoires and uses this information to express the type and grade of inflammation needed at each site and at each moment. In short, the immune system functions as the bioinformatic computer, defense force and public works department of the body. To learn about the state of an individual's body, we only need to consult the immune system computer.

In order to reveal the vast information stored in the immune system we have developed a micro-array technology combined with advanced bioinformatic methods which enables us to determine the profile of antibodies present in a drop of blood. We are now studying whether this tool can diagnose diseases such as juvenile diabetes and cancer at an early stage and to predict the outcome of the disease. In another application we study whether this technology may predict organ rejection after lung or kidney transplantation. This is a joint project with Prof. Irun Cohen and Dr. Eli Sahar (ImmunArray Ltd.).

Other collaborative projects at WIS include work with David Givol on cancer stem cells, with Yossi Yarden's group on the regulatory cascades initiated by activating the EGF receptor, with Varda Rotter's group on in-vitro development of cancer and the effects of mutant p53, with Zelig Eshhar on prostate cancer and with Leo Sachs on the interplay between cancer and tissue-specific differentiation.

Selected publications

Tsafrir, D., M. Bacolod, Z. Selvanayagam, I. Tsafrir, J. Shia, Z. Zeng, H. Liu, C. Krier, R. F. Stengel, F. Barany, W.L. Gerald, P.B. Paty, E. Domany and D.A. Notterman, Cancer Research 66, 2129 (2006). Relationship of Gene expression and Chromosomal Abnormalities in Colorectal Cancer.

Gal, H., N. Amariglio, L. Trakhtenbrot, J. Jacob-Hirsh, O. Margalit, A. Avigdor, A. Nagler, S. Tavor, L. Ein-Dor, T. Lapidot, et al. (2006). "Gene expression profiles of AML derived stem cells; similarity to hematopoietic stem cells." Leukemia 20(12): 2147-54.

Ein-Dor, L., O. Zuk and E. Domany (2006). "Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer." Proc Natl Acad Sci U S A 103(15): 5923-8.

Polak, P. and E. Domany (2006). "Alu elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes." BMC Genomics 7: 133.

Quintana, F. J., Y. Merbl, E. Sahar, E. Domany and I. R. Cohen (2006). "Antigen-chip technology for accessing global information about the state of the body." Lupus 15(7): 428-30.

Amit, I., A. Citri, T. Shay, Y. Lu, M. Katz, F. Zhang, G. Tarcic, D. Siwak, J. Lahad, J. Jacob-Hirsch, et al. (2007). "A module of negative feedback regulators defines growth factor signaling." Nat Genet 39(4): 503-12.

Axelsen, J. B., J. Lotem, L. Sachs and E. Domany (2007). "Genes overexpressed in different human solid cancers exhibit different tissue-specific expression profiles." Proc Natl Acad Sci U S A 104(32): 13122-7.

Fainaru, O., T. Shay, S. Hantisteanu, D. Goldenberg, E. Domany and Y. Groner (2007). "TGFbeta-dependent gene expression profile during maturation of dendritic cells." Genes Immun 8(3): 239-44.

Gavert, N., M. Sheffer, S. Raveh, S. Spaderna, M. Shtutman, T. Brabletz, F. Barany, P. Paty, D. Notterman, E. Domany, et al. (2007). "Expression of L1-CAM and ADAM10 in human colon cancer cells induces metastasis." Cancer Res 67(16): 7703-12.

Hertzberg, L., D. R. Betts, S. C. Raimondi, B. W. Schafer, D. A. Notterman, E. Domany and S. Izraeli (2007). "Prediction of chromosomal aneuploidy from gene expression data." Genes Chromosomes Cancer 46(1): 75-86.

Hertzberg, L., S. Izraeli and E. Domany (2007). "STOP: searching for transcription factor motifs using gene expression." Bioinformatics 23(14): 1737-43.

Katzenellenbogen, M., L. Mizrahi, O. Pappo, N. Klopstock, D. Olam, H. Barash, E. Domany, E. Galun and D. Goldenberg (2007). "Molecular mechanisms of the chemopreventive effect on hepatocellular carcinoma development in Mdr2 knockout mice." Mol Cancer Ther 6(4): 1283-91.

Katzenellenbogen, M., L. Mizrahi, O. Pappo, N. Klopstock, D. Olam, J. Jacob-Hirsch, N. Amariglio, G. Rechavi, E. Domany, E. Galun, et al. (2007). "Molecular mechanisms of liver carcinogenesis in the mdr2-knockout mice." Mol Cancer Res 5(11): 1159-70.

Tabach, Y., R. Brosh, Y. Buganim, A. Reiner, O. Zuk, A. Yitzhaky, M. Koudritsky, V. Rotter and E. Domany (2007). "Wide-scale analysis of human functional transcription factor binding reveals a strong bias towards the transcription start site." PLoS ONE 2(8): e807.

Acknowledgements

Eytan Domany is the incumbent of the Henry J Leir Professorial Chair. We acknowledge support from the NIH, grant #5 P01 CA 65930-06; The Ridgefield Foundation; The Israel Academy of Sciences, Bikura Program; The Minerva Foundation; EC FP6 funding; IMOS Tashtiyot; The Wolfson Foundation and the Mitchell Charitable Foundation.

INTERNAL support

Yeda CEO Fund; Mrs Ruth Ziegler Fund for Stem Cell Research