FINAL REPORT

Discovery and prioritization of gene regulatory modules driving oncogenesis

This report covers our research activity funded by the Belspo Return grant during the years 2012-2013. During that period, we developed and evaluated an approach to discover the regulatory modules involved in cancer from the expression data.

1. REMINDER OF THE OBJECTIVES

The goal of my project is to discover the regulatory modules whose combined activity best explains how a particular combination of mutations (coding and non-coding) produces the gene expression profiles observed in cancer cells.

The first challenge of our project is to discover perturbed cancer gene modules, together with the upstream transcription factor (TF). A next step to achieve our objectives will be to map the genomic variations (mutations, indels, copy number variations (CNVs)), preferentially identified in the same tumor samples, to the predicted cancer regulatory modules and to prioritize the most affected modules.

2. METHODOLOGY AND RESULTS

iRegulon, a powerful sequence-based method to predict human regulons

In this project, we developed, used and validated a tool called *iRegulon*. This tool is a key step in our approach for the detection of enriched TFs and their regulons. iRegulon is available as a user-friendly Cytoscape plugin (popular software for biological network analysis) and a tutorial is available at http://iregulon.aerts.org. We used thousand of ChIP-seq data corresponding to 115 sequence specific TFs from ENCODE to evaluate the performance of iRegulon. When applied to top scored 200 genes as query, we could recover up to 52% and 62% among the top 1 and top 3 predicted regulators, respectively. We evaluated the impact of the use of large regulatory search space (TSS±10kb) *versus* small regulatory search space (500 upstream TSS). We compared the contribution of the different motif source collections to our performances on ENCODE datasets. We also showed that iRegulon is 40% better at detecting the correct human TFs than 8 other motif discovery approaches (Fig. 1).

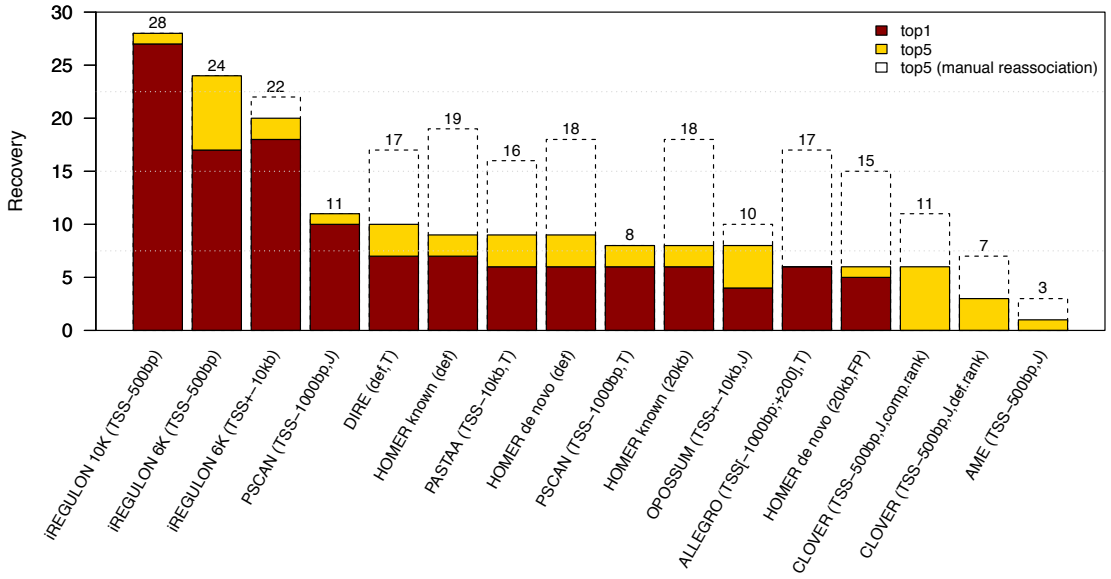

Figure 1. Tool comparison using 30 gene sets constructed as the top 200 target genes based on ChIP peak occurences in the 20kb regulatory region for 30 TFs (selected from FactorBook dataset with the canonical motif found top enriched in the ChIP peaks). Two versions of iRegulon were compared with eight other publicly available motif enrichment tools, namely OPOSSUM (Kwon et al., 2012), DIRE (Gotea and Ovcharenko, 2008), PASTAA (Roider et al., 2009), PSCAN (Zambelli et al., 2009), Clover (Frith et al., 2004), AME (McLeay and Bailey, 2010), Allegro (Halperin et al., 2009) and HOMER (Heinz et al., 2010) (de novo and known motifs tools). The number of times the queried TF was identified in the top1 (red) and top5 (yellow) is recorded. The dashed boxes represent top 5 recoveries if similar motifs are manually re-associated to the query TF.

We are working already on a second version that integrates ChIP-Seq-based rankings in addition to the motifs-based rankings and a larger motif library (9713 instead of 6383 PWMs in the previous version). The evaluation of this second version shows that we can detect 10 more TFs from the ENCODE ChIP-Seq datasets and the performances are improved to up to 70.4% and 75.2% for the top 1 and top 3 predicted regulators, respectively (see Fig. 1). We also developed Galaxy plugins (popular interface for NGS data analysis) for iRegulon and other regulatory motif analysis tools (Jonas Meulemans, bachelor student project under my supervision). The paper of iRegulon is currently under review process in the *Cell Reports* journal.

New approach for the inference of regulatory modules

To detect the gene regulatory modules involved in cancer, we developed a pipeline presented in Figure 2. We first predict cancer modules as sets of tightly co-expressed genes across a gene expression data set using an advanced bi-clustering approach called *GaneSh*, available in *LeMoNe* software package (Joshi et al., 2008,

2009). To assign TFs to each module, we applied our newly developed tool, *iRegulon*, which uses motif discovery to detect the motifs and TFs that have target enrichment in a set of co-expressed genes. These results allow us to infer the direct targets and then to refine the predicted co-expressed modules into regulons (*i.e.* set of TF with its target genes). The candidate regulon for a given TF integrates all the targets that are found as enriched targets for this TF when iRegulon is applied on all different gene clusters. To comfort our TF predictions from iRegulon, we include LeMoNe regulator predictions based on independent expression profile analysis. Next, we use *Gene Set Enrichment Analysis* (GSEA) to score the enrichment of these regulons against different cancer subtypes. The result of this approach is a list of candidate regulons that can be prioritized according to the TF enrichment, their functional enrichment or by other parameters (e.g. Survival association).

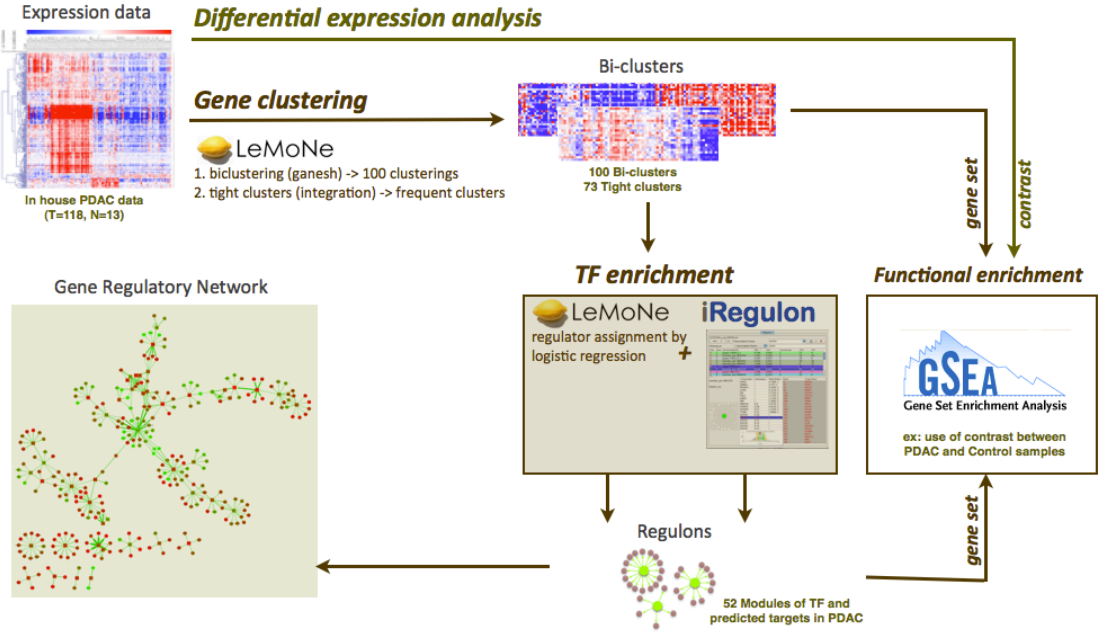

Figure 2. Flowchart with results of its application in PDAC dataset.

Application in pancreatic cancer dataset

We first develop and apply our strategy on an in-house microarray dataset corresponding to 118 samples of different subtypes of pancreatic ductal adenocarcinoma (PDAC) and 13 control samples from a KU Leuven collaborator, Pr. Baki Topal (Abdominal Surgical Oncology Department). These data also include clinical information such as tumor stage that were found to be significantly associated to survival. When applied to these data, we inferred 73 modules and 52 regulons, prioritized them using their enrichment scores on contrasts over supervised and unsupervised sample clusters. None of them show a clear association to survival, but they show a good expression contrast between tumour and control samples or between previously established PDAC subtypes (Collisson et

al., 2011). Using KRAS-dependent signatures, we also predicted that most of our samples are KRAS mutated (top gene mutated in Pancreatic cancer). We used the published somatic mutations from Cancer consortiums (85 donors from Canada and 187 donors from Australia) to get the mutational load of the predicted regulatory modules, but detected only one over-mutated module (4 genes out of 17, including 2 zinc fingers TFs).
Seven top regulons were screened experimentally by Immuno-Histo-Chemistry, but only two TFs were detected consistently with the microarray results (i.e. positive in normal tissues and negative in tumoral tissues or vice versa) and are probably correlated with aggressiveness of the cancer or survival (one of them is found mutated in ICGC PDAC data). These genes have been involved in cancer but not yet in Pancreatic cancer. The paper is under writing process. In the next steps, it will be also possible to assess the microRNA deregulation of our regulatory modules because a microarray dataset has been also performed for the microRNAs on the same PDAC samples.

Application to T-ALL dataset

The application of our approach on T-ALL data, initially planned in the project, was performed by a PhD student (Zeynep Kalender) as she was also working at identifying genomic variations from exome and RNA-seq data in T-ALL in collaboration with Pr. Jan Cools (Center for Human Genetics, KU Leuven). The paper about her transcriptome variation analysis is recently published in *PLoS Genetics* (Kalender Atak et al., 2013). In a next step, we will focus on mapping the mutations in the predicted regulatory modules to get their *mutational load* and identify the master regulators involved in T-ALL. To do this, we will apply the same strategy for the inference of regulatory modules from T-ALL expression datasets (RNA-seq and microarray data).

Pan-cancer regulatory modules

As iRegulon can be applied to any gene signature, we run iRegulon to more than twenty thousand co-expressed gene signatures related to cancer (MSigDB, GeneSigDB and ganesh clusters from inSilicoDB arrays). This results in a scoring matrix of 886 TFs versus 18205 targets showing as a score the number of signatures where a given TF is found enriched with a given candidate target. We call this matrix the *Metaregulons* as it reflects the cancer related TF-targets interactions and it is possible to use the iRegulon plugin to query the targetome of a TF of interest. This metatargetome can be used further to prioritize the "cancer-relatedness" of predicted modules/regulons and is used in the lab to refine target predictions. A next step of this project will be to analyze such network to assess if oncogenes/tumor suppresors or mutations are more often represented in such networks compared to the ChIP-seq-derived targetomes of normal tissue.

Application to microRNA regulatory networks

One of our systematic level findings using iRegulon on 159 annotated microRNA targetomes is that there are significant cross-talks between TF and microRNA regulons. For example, in collaboration with Pierre Lau and Bart de Strooper (KU Leuven), we found that miR-132 is a top down-regulated microRNAs in Alzheimer disease and share 59 targets with FOXO TFs while the messenger RNA of FOXO1a is up-regulated and a key target of miR-132-3p (paper published in *EMBO Molecular Medicine*). On the other hand, in the context of a collaboration with Ashok Venkitaraman (Cambridge, UK), we found that DNA Damage induced miRNAs are p53-dependent indirectly and directly as some of them are enriched in p53 binding (paper to be submitted soon in *PLOS One*). I am co-first author in these two publications. In the near future, we will develop iRegulon to detect enriched TFs from sets of microRNAs. Similarly, I made a version of iRegulon to make TF enrichment analysis from set of lincRNAs. Preliminary tests using our in-house p53 datasets worked well at rediscovering p53 as the top master regulator.

In collaboration with lab members working on the drosophila regulatory networks (Marina Naval and Delphine Potier), we made a small-RNA-seq experiment in order to detect the microRNAs involved in the eye development in two different drosophila species. We found a phenotype for two mutants showing overexpression of top eye-specific microRNAs. As we have previously inferred the transcriptional network from RNA-seq for these species (Naval-Sánchez et al., 2013), we aim at developing an integrative approach similar to iRegulon to predict the miRNA-mRNA-TF regulatory networks.

References

Collisson, E. a, Sadanandam, A., Olson, P., Gibb, W.J., Truitt, M., Gu, S., Cooc, J., Weinkle, J., Kim, G.E., Jakkula, L., et al. (2011). Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy. Nat. Med. *17*, 500–503.

Frith, M.C., Fu, Y., Yu, L., Chen, J.-F., Hansen, U., and Weng, Z. (2004). Detection of functional DNA motifs via statistical over-representation. Nucleic Acids Res. *32*, 1372–1381.

Gotea, V., and Ovcharenko, I. (2008). DiRE: identifying distant regulatory elements of co-expressed genes. Nucleic Acids Res. *36*, W133–9.

Halperin, Y., Linhart, C., Ulitsky, I., and Shamir, R. (2009). Allegro: analyzing expression and sequence in concert to discover regulatory programs. Nucleic Acids Res. *37*, 1566–1579.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell *38*, 576–589.

Joshi, A., Van de Peer, Y., and Michoel, T. (2008). Analysis of a Gibbs sampler method for model-based clustering of gene expression data. Bioinformatics *24*, 176–183.

Joshi, A., De Smet, R., Marchal, K., Van de Peer, Y., and Michoel, T. (2009). Module networks revisited: computational assessment and prioritization of model predictions. Bioinformatics *25*, 490–496.

Kalender Atak, Z., Gianfelici, V., Hulselmans, G., De Keersmaecker, K., Devasia, A.G., Geerdens, E., Mentens, N., Chiaretti, S., Durinck, K., Uyttebroeck, A., et al. (2013). Comprehensive analysis of transcriptome variation uncovers known and novel driver events in T-cell acute lymphoblastic leukemia. PLoS Genet. *9*, e1003997.

Kwon, A.T., Arenillas, D.J., Worsley Hunt, R., and Wasserman, W.W. (2012). oPOSSUM-3: advanced analysis of regulatory motif over-representation across genes or ChIP-Seq datasets. G3 (Bethesda). *2*, 987–1002.

McLeay, R.C., and Bailey, T.L. (2010). Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. BMC Bioinformatics *11*, 165.

Naval-Sánchez, M., Potier, D., Haagen, L., Sánchez, M., Munck, S., Van de Sande, B., Casares, F., Christiaens, V., and Aerts, S. (2013). Comparative motif discovery combined with comparative transcriptomics yields accurate targetome and enhancer predictions. Genome Res. *23*, 74–88.

Roider, H.G., Manke, T., O'Keeffe, S., Vingron, M., and Haas, S. a (2009). PASTAA: identifying transcription factors associated with sets of co-regulated genes. Bioinformatics *25*, 435–442.

Zambelli, F., Pesole, G., and Pavesi, G. (2009). Pscan: finding over-represented transcription factor binding site motifs in sequences from co-regulated or co-expressed genes. Nucleic Acids Res. *37*, W247–52.

3. DIFFUSION AND VALORISATION

PUBLICATIONS

Publications linked to the Belspo project are on going. A copy will be sent once publication is accepted.
** shared authorship*

1. Rekin's Janky*, Annelien Verfaillie*, Hana Imrichová, Bram Van de Sande, Laura Standaert, Valerie Christiaens, Gert Hulselmans, Koen Herten, Marina Naval Sanchez, Delphine Potier, Dmitry Svetlichnyy, Zeynep Kalender Atak, Mark Fiers, Jean-Christophe Marine, and Stein Aerts. Detection of *cis*-regulatory master regulators enables reverse engineering regulons from cancer gene signatures. ***Under review process in Cell Reports.***
2. Hiroyoshi Hattori*, Rekin's Janky*, Wilfried Nietfeld, Stein Aerts, M. Madan Babu, Ashok R. Venkitaraman. p53 shapes genome-wide and cell type-specific changes in microRNA expression during the human DNA damage response. ***Under submission process to PLOS One.***
3. Rekin's Janky*, Maria Mercedes Binda*, Joke Allemeersch, Anke Van den broeck, Olivier Govaere, Johan Swinnen, Tania Roskams, Stein Aerts and Baki Topal. Master regulator analysis across a large PDAC cohort reveals developmental factor as tumor suppressor in PDAC. ***Writing process.***
4. Pierre Lau, Koen Bossers, Rekin's Janky, Evgenia Salta, Carlo Sala Frigerio, Shahar Barbash, Roy Rothman, Annerieke S. R. Sierksma, Amantha Thathiah, David Greenberg, Aikaterini S. Papadopoulou, Tilmann Achsel, Torik Ayoubi, Hermona Soreq, Joost Verhaagen, Dick F. Swaab, Stein Aerts, Bart De Strooper. Alteration of the microRNA network during the progression of Alzheimer's disease. EMBO Mol Med. 2013 Oct;5(10):1613-34. *(Impact Factor: 7.8)*
5. Rekin's Janky. Cancer regulatory modules. 25 September 2012. *KU Leuven F+ Fellowship Report (F+/11/030).*
6. Tamir Chandra, Kristina Kirschner, Jean-Yves Thuret, Benjamin D Pope, Tyrone Ryba, Scott Newman, Kashif Ahmed, Shamith A Samarajiwa, Rafik Salama, Thomas Carroll, Rory Stark, Rekin's Janky, Masako Narita, Lixiang Xue, Agustin Chicas, Sabrina Nũnez, Ralf Janknecht, Yoko Hayashi-Takanaka, Michael D Wilson, Aileen Marshall, Duncan T Odom, M Madan Babu, David P Bazett-Jones, Simon Tavaré, Paul A W Edwards, Scott W Lowe, Hiroshi Kimura, David M Gilbert, Masashi Narita. Independence of repressive histone marks and chromatin compaction during senescent heterochromatic layer formation. *Molecular Cell, Volume 47, Issue 2, 27 July 2012, Pages 203-214. (Impact Factor: 15.28)*

7. Rekin's Janky. Structure and Evolution of prokaryotic transcription factor binding sites. Bacterial Gene Regulation and Transcriptional Networks. Horizon Scientific Press. *Invited book chapter. Bacterial Gene Regulation and Transcriptional Networks, Chapter 3, Caister Academic Press, March 2013, Editors: M. Madan Babu.*

CONFERENCES

1. Rekin's Janky, Annelien Verfaillie, Gert Hulselmans, Stein Aerts**.** Detecting master regulators and cis-regulatory interactions in human cancer related gene networks. Benelux Bioinformatics Conference (BBC) 2013. Brussels, 9-10 December 2013. ***Abstract & Poster & Demo presentation****.*
2. Annelien Verfaillie, Rekin's Janky, Hana Himrichova, Valerie Christiaens, Stein Aerts. Exploring the p53 transcriptional network through integrative genomics reveals new candidate target genes and co-factors. Wellcome Trust Scientific conference "Functional Genomics and Systems Biology 2013". Cambridge, United Kingdom, 21-23 November 2013. ***Poster***.
3. Rekin's Janky, Annelien Verfaillie, Gert Hulselmans, Laura Standaert, Jean-Christophe Marine, Stein Aerts. Detecting master regulators and cis-regulatory interactions in human cancer related gene networks. Cold Spring Harbor Laboratory Genome Informatics 2013. Cold Spring Harbor, United States of America (USA), 30 October - 2 November 2013. ***Abstract & Poster.***
4. Rekin's Janky, Baki Topal, Joke Allemeersch, Maria Mercedes Binda, Anke Van den broeck, Olivier Govaere, Tania Roskams, Stein Aerts. Identification of master transcription factors in cancer. EMBO Conference Series. From Functional Genomics to Systems Biology. EMBL Heidelberg, Germany, 17- 20 November 2012. ***Abstract & Poster.***
5. Rekin's Janky. Prediction of regulatory modules enriched in pancreatic cancer subtypes. Genome Bioinformatics Meetings. Leuven, 8 May 2012. ***Oral presentation.***
6. Rekin's Janky. Identification of master transcription factors in human pancreatic cancer subtypes. UZ Leuven & Leuven International Doctoral School Biomedical Sciences: Cancer Programme. Oncoforum 9, Leuven, 3 May 2012. ***Short oral presentation, abstract and poster.***

4. ADDED VALUE FOR THE RESEARCH PROGRAMME TO WHICH THE RESEARCH IS CONNECTED AND FOR THE HOST UNIT IN GENERAL.

Research in the Laboratory of Computational Biology is focusing on gene and genome regulation, with applications in Drosophila and cancer. The tool, iRegulon, and approaches that I have developed in this research programme are core technologies of our lab and this work is integrated in several projects of the laboratory:

- iRegulon is being applied by Zeynep Kalender (recent PhD) on leukemia data (T-ALL) in a systematic way to identify regulatory modules and map mutations on these modules (see applications discussed in section 1).
- iRegulon has been also developed for applications in Drosophila and is currently used by Delphine Potier (Post-doc) and Marina Naval (PhD student) to detect TFs, co-factors and targets from perturbation experiments and cross-species comparisons.
- The core of the program iRegulon is implemented in order to allow the use of regulatory tracks such as ChIP-seq data but also to query for functional non coding regions. Hana Imrichova (PhD student) is developing and validating extensively this version and Dmitry Schvetlinyy (PhD student) is using this version and our motif collection to develop regulatory models for the detection of functional transcriptional targets.
- One of the recent advance of this approach allows the detection of enriched motif(s)/TF(s) in the promoter region of lincRNAs which would benefit to Annelien Verfaillie (PhD student).
- Several collaborations were also possible thanks to iRegulon within-host department with JC Marine and Bart de Strooper (VIB, Human Genetics department); and within-university with Baki Topal (KUL), and with external labs from other national or international universities.

Since we made the iRegulon plugin accessible to the scientific community, in late September 2013, our website (http://iregulon.aertslab.org/) has been accessed 300 times and the plugin has been downloaded 118 times on the Cytoscape App Store (http://apps.cytoscape.org/apps/iregulon).

5. THE PERSPECTIVES OF INTEGRATION OF THE RESEARCHER IN THE BELGIAN SCIENTIFIC MILIEU

I have been an active research scientist in the Belgian Scientific milieu. I took teaching and supervising responsibilities and give communications in several scientific events.

I supervised, and co-supervised with Bram van de Sande, two bachelor students from HOWEST high-school (Ghent). One of these students got an Innovation award in 2012 (http://www.innovation-awards.be) for the development of the preliminary Cytoscape plugin.

I co-animated with Dr. Delphine Potier and Pr. Stein Aerts a tutorial session resenting iRegulon and other Galaxy analysis tools for RNA-Seq data analysis (Flemish Training Network Life Sciences Workshop - Next Generation Sequencing - Leuven, 17-18-20-21 September 2012). I also give a course to the cancer programme entitled Bioinformatic tools in Next-Generation Sequencing (21 February 2013).

At the national level, I could benefit from the work and interactions of BioMAGNet network (IAP P6/25) as I use RSAT tools from J. van Helden (ULB), and make use of *Ganesh/LeMoNe* approaches from Y. van de Peer (VIB, UGent) and *GENIE3* from P. Geurts (ULG) to make gene modules.

I also presented this work as a poster, talk or software demo in internal and national conferences (Oncoforum, BBC), and international conferences (RECOMB, EMBO and CSHL Genome Informatics).

In my perspectives, I will stay 9 more months in the Laboratory of Computational Biology paid by the FWO, will ask for a renewal of this FWO funding and will apply for independent positions in Belgium.

KU LEUVEN

LABORATORY OF COMPUTATIONAL BIOLOGY
CENTER FOR HUMAN GENETICS
UNIVERSITY OF LEUVEN
HERESTRAAT 49, PO BOX 602
3000 LEUVEN, BELGIUM

Letter of Appreciation for Dr. Rekin's Janky

To Whom It May Concern:

I am extremely pleased that Dr. Rekin's Janky has joined my laboratory as a postdoctoral fellow with the Belgian Science Policy Office (Belspo) funding. Rekin's joined my group in June 2011 after being a Career Development Fellow in Cambridge in Dr. Madan Babu's lab (UK). Rekin's has carried out research in the field of cancer regulatory genomics. During his postdoc he used his expertise in genome regulation to develop and improve a bioinformatics method, called iRegulon, for the detection of master regulators and transcriptional targets in human. When Rekin's started he quickly learned about the iRegulon project in the lab and was able to take the lead on this project, to optimize iRegulon further, and to develop several biological applications. Doing so, he integrated his previous experience on transcriptional regulation, miRNAs and networks, into the iRegulon project. One of his important findings highlights cross talks between miRNA and transcriptional regulation. He made an extensive validation of the iRegulon tool allowing us to compare our performances with other tools and putting us as top in the field. Rekin's main first-author postdoc paper, about iRegulon, is currently under review for the journal Cell Reports. In addition, during his postdoc stay, Rekin's has worked on multiple other projects, using different types of high-throughput data ranging from microarray data (Affymetrix) to next-generation sequencing data (small RNA-seq, RNA-seq and ChIP-seq), and from human to fly. His collaboration with the B. De Strooper lab lead to a co-author position on a publication in EMBO Molecular Medicine (Oct 2013). A second first-author manuscript stemming from another collaboration, with the B. Topal lab on pancreatic cancer, is currently in preparation. Finally, Rekin's presented his work via posters and talks at several international conferences, guided several student internships, and independently writes his own project proposals and reports.

Dr. Rekin's Janky shows excellent communication skills. He is really good at illustrating his ideas and computational concepts with clarity. He developed a very professional website and tutorial for the iRegulon (http://iregulon.aertslab.org). He always finds times to discuss with other lab members, is enthusiastic and a great team player. He takes actively part at lab events and lab meetings, he is very interactive and he is keen to give his feedback and to ask questions if necessary when collaborating and during meetings. Rekin's also submitted projects for rotation students and guided several students during their internship in the lab.

In conclusion, Rekin's Janky is an excellent scientist and I am more than happy to have Dr. Rekin's Janky in my lab. Currently, Rekin's is going to continue his postdoc for an additional year on his FWO postdoc fellowship and I am convinced that he will become an independent and successful scientist in the field of bioinformatics and regulatory genomics.

Sincerely,

Stein Aerts

PROF. DR. IR. STEIN AERTS,
TEL. + 32 16 330710
stein.aerts@med.kuleuven.be
http://www.aertslab.org