Author's response to reviews

Title: Large differences in global transcriptional regulatory programs of normal and tumor colon cells

Authors:

David Cordero (dcordero@iconcologia.net)
Xavier Sole (x.sole@iconcologia.net)
Marta Crous-Bou (mcrous@iconcologia.net)
Rebeca Sanz-Pampolina (rebecasanz@iconcologia.net)
Laia Pare-Brunet (lpare@iconcologia.net)
Elisabet Guino (e.quino@iconcologia.net)
David Olivares (dolivares@iconcologia.net)
Antoni Berenguer (toniberenguer@iconcologia.net)
Cristina Santos (csantos@iconcologia.net)
Ramon Salazar (ramonsalazar@iconcologia.net)
Sebastiano Biondo (sbiondo@bellvitgehospital.cat)
Victor Moreno (v.moreno@iconcologia.net)

Version: 4 Date: 6 August 2014

Author's response to reviews: see over
Reviewer’s report #1:

The authors present an interesting finding whereby they highlight major loss of transcriptional regulation in tumour compared to normal colon cells. This paper will be of interests to scientists involved in colon cancer research. The authors present their aims very clearly and also acknowledge some of the limitations of the study. They have tried to validate some of the findings in silico, but to ensure completeness of these findings several computational analyses need to be addressed before the paper can be published.

Major Essential Revisions:

1. The authors observe an interesting trend regarding a large loss of transcriptional interactions, target genes, and transcription factors in the tumour regulatory network, compared to the normal colon regulatory network. This behavior is quite intriguing, but to ensure that this trend is specific to colon cancer, these findings need to be compared to either A) another colon cancer dataset, for example matched normal-tumour pairs for TCGA data or B) against a random dataset. Is the trend observed for GRNs of other normal versus tumour colon datasets? Is the same behavior observed when comparing random GRNs? The key message here is to ensure that much lot of the TFs, target genes, and transcriptional interactions between the tumour and normal samples are truly reflective of the disease, as opposed to being generically lost when looking at any particular non-colon GRN.

We agree with the reviewer in the added value of replicating our main findings using a different dataset. We have tried to identify a suitable dataset to infer regulatory networks with the ARACNe algorithm. The authors of the algorithm stress in their papers the importance of a minimum sample size and they specifically discourage users to apply their algorithm on small datasets [Margolin et al, Nature Protocols 2006; Margolin et al, BMC Bioinformatics 2006]. A hundred samples is the recommended minimum sample size required to infer transcriptional networks with proper accuracy. The TCGA colon dataset has very few pairs of normal-tumor samples: 19 for gene expression arrays and 23 for RNA-Seq. We have revisited again the most comprehensive public gene expression databases (GEO, ArrayExpress) to check for newer potentially useful datasets to replicate our results. Besides the previously mentioned TCGA dataset, the best-identified datasets were the following:

- GSE44861 (47 normal-tumor pairs, Affymetrix HG U133A Array)
- GSE41258 (43 normal-tumor pairs, Affymetrix HG U133A Array)
- GSE20916 (40 normal-tumor pairs, Affymetrix HG U133 Plus 2.0 Array)
- GSE20970 (30 normal-tumor pairs, NCI/ATC Hs-OperonV2 spotted oligonucleotide)
- GSE6988 (25 normal-tumor pairs, Human 17K non-commercial spotted cDNA array)
- GSE35834 (23 normal-tumor pairs, Affymetrix Human Exon 1.0 ST Array)
- E-MTAB-57 (22 normal-tumor pairs, Affymetrix HG U133A Array)
- E-MTAB-833 (21 normal-tumor pairs, Agilent Whole HG Oligo Microarray G4112A)
- GSE38940 (20 normal-tumor pairs, HEEBO Human oligo array custom spotted)
- GSE37182 (15 normal-tumor pairs, Illumina HumanHT-12 V3.0 expression beadchip)

As it can be seen, all datasets lie far away from the minimum requirement of 100 pairs of samples, thus hindering a possible replication of our findings. We finally decided not to pursue this goal because we could not find a suitable dataset with which we could apply the same methodology presented in our study.
Regarding the comparison to a random dataset, the ARACNe algorithm uses a different approach to define a robust estimate of the inferred network: generates a set of 1000 networks using resampling techniques (bootstrap) and only those strong, consistent, connections are finally selected as significant. The problem of generating random datasets is that data permutation completely destroys the correlation structure and the bootstrap step doesn’t identify any significant connection.

Even though we are not able to validate our results in an independent dataset, we believe that our results suggest that our networks are really capturing true transcriptional interactions. This is specifically shown in the results section “In-silico network validation with experimental data”. Moreover, the fact that the list of TFs with increased number of targets in the tumor network shows a strong enrichment in colorectal cancer genes, as it is shown in section “Gain of regulatory activity in tumor cells”, suggests that the differences that we are observing between the both networks may be reflective of true transcriptional regulatory alterations in colon cancer tumor cells.

We thank the reviewer’s comment, and we have added a paragraph in the discussion about the lack of proper datasets to replicate our findings.

2. The authors indicate from their findings [Section entitled Massive loss of regulatory activity in tumour cells, lines 15-24, page 4] that the large loss of edges in the tumour network, compared to the normal network, is suggestive of transcriptional regulation than gene silencing. They then classify the lost edges of the network into four classes based on silencing of the TF and/or targets. While Figure 4 is quite clear in graphically representing these classes, it is important for the reader to be quantitatively aware of the distribution of these classes. How many edges are lost as a result of just TF or target gene silencing? Are there particularly important examples of these cases that are also reflective on the biology of the colon cancer?

We agree with the reviewer that this classification finally has been graphically represented, but we did not show the proportion of edged lost dues to each one of the four situations. In the results section (see ‘Massive loss of regulatory activity in tumor cells’) we previously specified that ~96% of the lost edges in the tumor network do not show expression depletion neither in the TF nor in the target gene. We have now added to the results and also to the figure 4 legend the absolute number and the percentage of lost edges for each one of the 4 classes.

Regarding to the last part of the reviewer’s comment, first we want to thank the reviewer for pointing out this consideration. Our point of view is that the genes involved in the loss of edges due to TFs or target gene silencing (~4%), are more likely to belong to known altered colon cancer pathways due its apparent under-expression. In order to check this idea, we have tested and found that some examples of genes belonging to Wnt signaling pathway are present into three classes based on silencing of the TFs or targets. However, we empathize the underlying idea that the vast majority of lost edges would not be easy to identify just by exploring the expression values of their TFs or targets genes. We think that new and interesting undescribed mechanisms for molecular biology of colon cancer might be related to this gene deregulation without average gene expression change. We have now highlighted this approach to the discussion section.

3. Conversely, exploring the paired expression values of the TF and target genes should be similarly conducted for the 91 TFs and 235 up-regulated genes that were identified in the tumour regulatory network. [Section entitled Gain of Regulatory activity in Tumour cells] What is the correlation of the expression of the new 91 TFs on their target genes? Once again, are some of these TFs and up-
regulated target genes also observed when looking at GRNs of other colon datasets, or when looking at a random network?

Based on the reviewer’s comment, here we include a plot with the distribution of the MI for the 2,224 transcriptional interactions that are present in the tumor network but are missing in the normal network. As it is expected, the MI for the interaction in the tumor network is much higher than the interactions in the normal network. This figure has been added to additional file 2.

Regarding the validation, we point the reviewer to answer of question #1, in which we justify the inability to validate our finding due to a lack of proper expression datasets.

**Minor Essential Revisions:**

1. There are no legends for the Supplementary Data

   We would like to thank the reviewer for pointing this out. We have now added the figure legends in the PDF document uploaded as additional file 2. Other legends and additional descriptions for non-platform-specific file formats were uploaded all together in a PDF document named Additional file 6 (previously Additional file 7), as an index of files.

**Discretionary Revisions:**

1. While the authors present in much detail the numbers of TFs, target genes, and interactions that are lost or gained between the normal and tumour networks, some of the functional enrichment and biological interpretation of these findings, as relates to specifically colon cancer, is lost. The authors stress that there is considerable loss of transcriptional interactions in the tumour network compared to the normal regulatory network. It is not clear however, in the discussion section, which pathways or functionally enriched clusters from the normal network are lost in the tumour network. Are there particular clusters that were lost and which may influence the emergence of new functionality in the tumour cells? It is true that such causal relationships would be hard to verify without experiments, but I think a few sentences that highlight the significance the transcriptional loss are important to
mention for further biological interpretation and use of these findings in future experiments. The discussion section needs to be re-formatted to make some of these concepts clearer.

We agree with the reviewer on this comment. In our work we have generally focused on the gain of activity in the tumor network rather than on the lost interactions. The main reason for this is that since so many thousands of genes and interactions are lost in the tumor regulatory network makes it extremely difficult to detect enriched functions. However, we agree with the reviewer that this is a striking finding and we have added these specific concepts in the discussion to emphasize this result and the potential impact on the future of cancer molecular biology.
Reviewer’s report #2:

In the “Large differences in global transcriptional regulatory programs of normal and tumor colon cells” Cordero et al. aimed to construct global transcriptional networks of normal and tumor colon cells by using reverse engineering approaches and the ARACNe algorithm. The originality of this work consists in the use of colon tumor and normal mucosa deriving from the same patients. This gave a great advantage to them in identifying the dysregulated network specifically present in the tumor cells compared to the adjacent mucosa’s ones.

The most interesting findings of this analysis is the big loss of regulatory activity in the tumor cells compared to the normal cells, that probably could be explained by different mechanisms such as genomic instability of the tumor cells, epigenomic or post-translational modifications as suggested by the authors. One of the limitations of the ARACNe network is that the network constructed is based on TF and gene expression profiles, for this reason would be really interesting to perform high throughput approaches on the normal and tumor colon cells to analyze i.e. methylation and acetylation markers or analyze the mutational status of these cells in order to validate their hypothesis.

Moreover, as major compulsory revision I would like to know if the authors find any major hub corresponding to any transcription factor known to be involved in the colon cancer pathogenesis to demonstrate the soundness of their approach?

In the results section, more specifically at the end of section entitled “Gain of regulatory activity in tumor cells”, we describe an enrichment analysis of KEGG pathways over TFs with increased activity in the tumor network. This analysis revealed a highly remarkable significance (p-value=8.9e-9) of the colorectal cancer KEGG pathway among the set of TFs that increase their regulatory activity in tumor cells. Well-known CRC genes such as FOS, TGFB3 and TGFB1 are present in these results. As an example, TGFB1 increases from 1 to 60 the number of transcriptional targets in tumor cells compared to normal tissue, becoming a prominent hub in the tumor network. Besides these well-known CRC-related genes, and as it is mentioned in the discussion, among the TFs ranked at the top of the list of increased connectivity, our analysis identified other genes also related to colorectal cancer: two oncogenes (MAFB and GLI2); proliferation-related genes (NOTCH3 and TGFB1); a transcriptional repressor involved in CRC epithelial-mesenchymal transition (SNAI2); GREM1, a bone morphogenic protein antagonist family that has been previously related with a locus strongly associated with increased colorectal cancer risk, and whose increased expression GREM1 has also been recently found in colorectal polyps, as well as in the dysplasia to carcinoma transition in colon tumors; and the Wnt signaling genes SFRP4, TWIST1, SMARCA4 and DKK3, potentially involved in CRC angiogenesis. We believe all these results demonstrate the soundness of our approach. However, we thank the reviewer for raising this point, and thus we have emphasized it in the discussion section by the addition of a specific note and the addition of new references pointing to the previously mentioned genes.

In addition, for the validation of their ARACNe predictions the authors uses the ChIP on chip and ChIP Seq data available in ENCODE. I was wondering if they could also use the information about the gene mutations available in the COSMIC database in order to verify if the hubs identified in their colon cancer interactome result mutated and therefore their transcriptional network result perturbed in any other known tumor.
According to the reviewer’s suggestion, we have studied the relationship of somatic mutations in colorectal tumors in the set of relevant genes identified through our network approach. Mutation data were obtained from the COSMIC database using the following parameters: large intestine (tissue), all (subtissue), carcinoma (histology), all (subhistology). Only genes with a mutation frequency greater than 5% were considered for the analysis. We have analyzed the degree distribution (as indicator of regulatory activity) for TFs and target genes, classified as frequently mutated (if present in COSMIC) or not. We have found that regulatory activity is independent of mutations for TFs. However, target genes included in COSMIC database showed a significant larger regulatory control than other non-mutated genes. This is a correlation study using independent data (we do not know if our tumors are actually mutated), but is suggestive that mutated genes trigger a regulatory control in the tumor. The presence of mutations combined with the alteration in their transcriptional regulatory connectivity postulate these genes as strong candidates to be involved in the pathogenesis of colon cancer, and even other type of tumors. We thank the reviewer for this interesting suggestion, and we have added these new findings in the results and discussion sections.

As minor essential revision in the Background, line 26 I would suggest the authors to put more updated references for the application of ARACNe algorithm not only in B cell lymphoma, but also in neuroblastoma [Carro et al., Nature 2010], T-ALL [Della Gatta et al. Nat Med. 2012] and in prostate cancer [Aytes et al., Cancer Cell 2014]

We would like to thank the reviewer for this suggestion. We have added more updated references regarding the application of ARACNe algorithm in other malignancies.