Identification of the transcriptional gene network in 2D and 3D human bronchial epithelial cell culture systems

V Belcastro[1], C Mathis[1], C Poussin[1], D Weisensee[2] and J Hoeng[1]

[1]Philip Morris International R&D, Philip Morris Products S.A., Neuchâtel, Switzerland
[2]Philip Morris International R&D, Philip Morris Research Laboratories GmbH, Cologne, Germany

Introduction

Microarray based genome-wide gene expression technology has become standard practice in systems biology to analyze transcriptomic changes at the cell, tissue, and organism level. Currently, the availability of a high number of condition-specific (e.g., disease state, tissue, perturbation) gene expression profiles enable reverse-engineering of transcriptional gene networks associated with a single specific condition such as a given cell type.

It is still not known how, and to what extent, transcriptional gene networks of cell types that are biologically close differ, and how they react when exposed to the same stimulus.

To investigate these aspects further, two cell-type-specific gene networks were reverse-engineered from 2D and 3D human normal bronchial epithelial cell cultures.

Abbreviations

AIR-100: Organotypically differentiated pseudostratified tissue derived from NHBE cells cultured at the air-liquid interface (ALI) for a 2-3 weeks. The reconstituted tissue comprises basal, goblet, and ciliated cells.

Methods

Data collection and analysis

Genome-wide gene expression samples (from cells grown under normal culture conditions) were collected from public repositories and internal studies. All the samples were obtained on Affymetrix HG-U133plus2 Microarray chips.

Culture	# of samples	Type	Ref
2D	(12+16)+(24+36)	Time course	1, PMI(*)
3D	58+36	Time course	PMI(*)

Table 1. Overview of the experiments used to infer the gene networks.
(*) In-house samples (not public yet).

Raw data (CEL files)
↓
Normalization: gcrma, bioconductor

The 2D-culture samples were collected from 2 studies, each including 2 sub-studies (see Table 1 above), and were normalized separately. Principal component analysis (PCA) showed significant variability across studies and across sub-studies within the same study. The sub-study effect was corrected for samples described in Table 1 (see Figure 1). The same approach was applied to remove study effect when samples the two studies were merged.

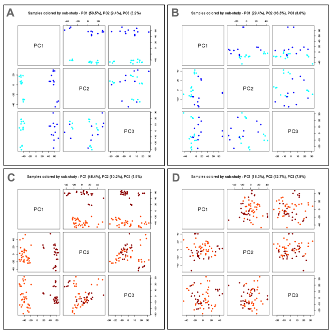

Figure 1. Principal component analysis (PCA) to highlight within study effect. Samples are colored according to the studies (**A** in [1], and **C** in PMI). The sub-study effect (blue and cyan for [1]; dark-red and red for PMI), which is clearly observable both in **A** and in **C**, was considered a batch effect and removed (**B** and **D**) using ComBat R package [2].

Once the sub-study effects were removed, normalized samples were merged (Figure 2).

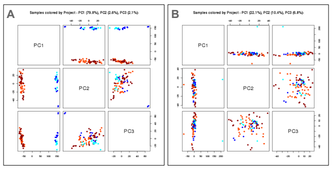

Figure 2. Samples are colored according to the study where they were generated (blue gradations for [1] and red gradations for PMI). The study effect observed in **A** was considered a batch effect and removed (**B**) using ComBat R package.

Samples from the 3D-culture showed similar variability across the 2 sub-studies generated at PMI. Thus, a similar procedure was applied to remove the sub-study effect.

Gene expressions data was filtered to remove probe sets with very low signals across samples. Out of 54,675 probe sets, 14,531 passed the filtering (12,855 for the 3D-culture). Probe sets were then mapped on gene symbols by selecting the most expressed probe set in the case of multiple associations (8,692 genes for 2D-culture and 9,846 genes for 3D-culture).

Gene network inference

Transcriptional gene networks from 2D- and 3D-culture samples were inferred by correlating gene expression profiles separately for each cell culture. Pairs of genes (edges), that had a correlation value with an associated FDR below 5E-06 (t-statistic [3] followed by the Benjamini-Hochberg correction) were retained for further analysis.

These selected pairs of genes were then used to build the 2D and 3D-culture transcriptional gene networks (Table 2).

culture	# of genes	# of edges
2D	8,692	3,777,374
3D	9,846	2,423,273

Table 2. Results of the inference process (correlation FDR<5.0E-05).

culture	# of communities	# of rich-clubs
2D	262	24
3D	249	34

Table 3. Results of the community finding process.

The adjacency matrices derived from the two gene networks were used to compute the Jaccard [4] distances between genes. The distances between genes were then used to discover highly connected communities of genes. To group genes into communities, the community finding affinity propagation clustering algorithm [5] was used.

Results

The gene sub-networks derived from the top scored edges are highly connected. The application of the clustering algorithm allowed identification of groups of highly connected genes (communities), as shown in Figure 3.

The community finding algorithm [5] grouped genes in communities based on the number of neighbors they share. 2D and 3D-culture gene networks account 262 and 249 communities, respectively.

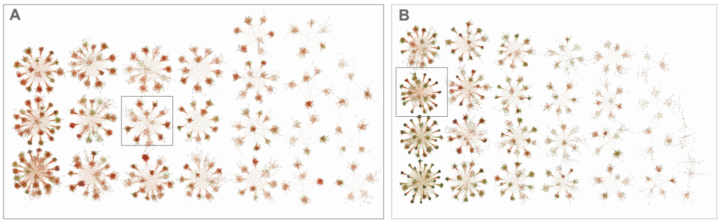

Figure 3. Overview of the communities and rich-clubs (cluster of communities) for 2D- (**A**) and 3D-culture (**B**). Intra-community and intra-rich-club significant correlations were retained. Negative correlations are reported in green, positive correlations in red. Distances between genes and between communities reflect the absolute values of the correlations [6].

In total 48% (2D-culture) and 52% (3D-culture) of the communities were observed to be associated with known functions based on gene ontological enrichments (FDR<0.05). Among the most significantly enriched communities *ribosomal genes* (Figure 4A, 2D-culture) and genes involved in *cell cycle* (Figure 4B, 3D-culture) were found.

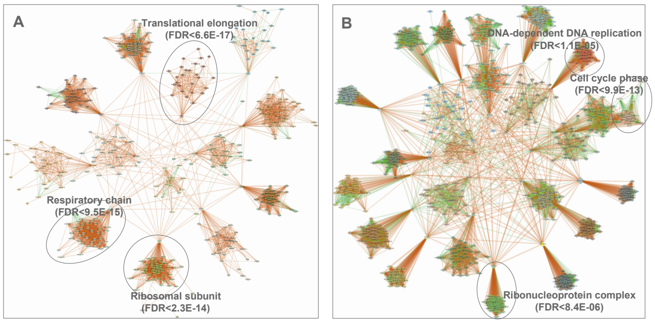

Figure 4. Example of rich-clubs, and communities of genes part of a rich-club, that were identified to be significantly associated to biological functions for the 2D- (**A**) and 3D-culture (**B**).

Conservation enrichment analysis allowed identification of those communities that were conserved between 2D and 3D-culture-derived networks. The analysis showed that 111 communities found in the 2D-culture, and 115 communities found in the 3D-culture were conserved across the gene networks (FDR<0.01). Among the top conserved community pairs, those significantly (FDR<6E-10) associated with the *cadmium ion binding* (yellow nodes in Figure 5) were found.

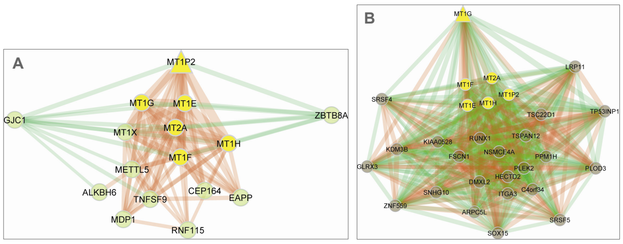

Figure 5. Example of community conservation across 2D- and 3D-culture gene networks. The 2 communities from 2D- (**A**) and 3D-culture (**B**) are enriched for *cadmium ion binding* (FDR<8.3E-13 and 1.8E-08, respectively). Nodes that are present in both the communities are shown in yellow.

Conclusions

Transcriptional gene networks were reverse-engineered from gene expression profiles of 2D and 3D human bronchial epithelial cells grown under normal (non-stimulated) culture conditions. The gene community analysis revealed that 50% of the gene communities were conserved across the transcriptional gene networks of NHBE (2D) and AIR100 (3D) cellular models when applying a FDR<0.01. The 16% of the genes that are only in the AIR100 gene network were enriched for genes involved in "*immune response*" (FDR<8.7E-04), which could be due to the maturity of the 3D tissue-like culture. The networks may serve as reference *in silico* models to investigate condition-specific interactions and to evaluate the response of differential perturbations collected in these two *in vitro* lung assays, which are essential for respiratory research.

References

[1] Jorgensen E et al, Cigarette smoke induces endoplasmic reticulum stress and the unfolded protein response in normal and malignant human lung cells. BMC Cancer 2008; 11(8):229

[2] Li C and Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 2007, 8:118-127

[3] Benjamini Y and Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society 1995, 57(1):289-300

[4] Tan et al. Introduction to data mining. Addison Wesley 2005, ISBN 0-321-32136-7

[5] Frey BJ and Dueck D. Clustering by passing messages between data points. Science 2007, 315:972-976

[6] Smoot ME et al, Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics 2011, 431-432

PMI RESEARCH & DEVELOPMENT

Philip Morris International Research & Development, Quai Jeanrenaud
5, 2000 Neuchâtel, Switzerland
T: +41 58 242 21 11, **F**: +41 58 242 28 11, **W**:
www.philipmorrisinternational.com

ISMB/ECCB 2011, Vienna, July 17-19