

# Interpreting high-throughput expression data of lung cancer subtypes in light of metabolic pathways

A. Rezola Urquia<sup>1</sup>, J. Pey Perez<sup>1</sup>, A. Rubio Díaz-Cordovés<sup>1</sup>, F.J. Planes Pedreño<sup>1</sup>

<sup>1</sup> Bioinformatics Department, CEIT and Tecnun, University of Navarra, San Sebastian, Spain  
{arezola, jpey, arubio, fplanes}@ceit.es

## Abstract

*The analysis of high-throughput molecular data in the context of metabolic pathways is essential to uncover their underlying functional structure. Among different metabolic pathway concepts in Systems Biology, elementary flux modes (EFMs) hold a predominant place, as they naturally capture the complexity and plasticity of cellular metabolism and go beyond pre-defined metabolic maps. However, their full computation is intractable in large metabolic networks and for this reason their application to human metabolism has been so far limited.*

*In a recent work, we determined a subset of EFMs in human metabolism and proposed a new protocol to integrate gene expression data, spotting key “characteristic” EFMs in different scenarios. Our approach was successfully applied to identify metabolic differences among several human healthy tissues.*

*In this article, we evaluated the performance of our approach in a more clinical interesting situation. In particular, we identified key characteristic EFMs in Adenocarcinoma and Squamous Cell Carcinoma Non-Small Cell Lung Cancers based on gene expression data. These major subtypes of lung cancer are well defined in the medical literature and therefore it is suitable as proof of concept for our approach. This work constitutes the starting point to establish a new methodology that allows us to distinguish key metabolic processes among different clinical outcomes.*

## 1. Introduction

With the expansion in the last decade of high-throughput molecular experimental technologies, particularly genomics and transcriptomics, a vast amount of data is available to the scientific community. Their analysis in the context of metabolic pathways is essential to uncover their underlying functional structure and it is a widely extended practice in the field of bioinformatics and systems biology. To this end, different approaches are found in the literature. Some of the existing tools explain high-throughput “omics” data at the light of pre-defined metabolic maps [1]. Despite their wide and successful application, these methods however restrict novel discoveries, as pre-defined maps do not capture the wide variety of complex metabolic states. Instead, the use of unbiased and mathematical concepts of pathways based on genome-scale metabolic networks is suitable for this purpose. Among them, path finding techniques from the field of graph theory have been a recurrent approach. In contrast with their low computational expense, they face important theoretical issues, as recently emphasized in [2]. For this reason, other more general pathway concepts have been introduced to interpret “omics” data, particularly Elementary Flux Modes [3] and recently Elementary Flux Patterns [4].

The EFMs approach was properly formulated in the middle of nineties. In essence, the EFMs approach enables us to decompose the whole metabolic network into minimal modes of behaviour, which naturally links them to the traditional concept of metabolic pathways. Since its definition, it has received much attention, showing that the pathway structure resulting from EFMs is more accurate and varied than paths and pre-defined maps [5]. At the theoretical level, its predictive power has been demonstrated in a number of works, *e.g.* a novel gluconeogenesis pathway from fatty acids with potential applications to obesity. Despite its early definition, the interpretation of “omics” data at the light of EFMs in human metabolism has been so far limited to the pioneering work of [3]. This has been due to the fact that the number of EFMs explodes in a combinatorial fashion as the network size increases and classical approaches fail to compute them in genome-scale networks. That is precisely the case for the human genome-scale metabolic network [6], which involves several thousands of reactions and metabolites.

To face with this issue, we have developed several optimization-based techniques to compute a subset of EFMs in genome-scale metabolic networks [7,8]. Based on them, in a recent work [9], we determined a subset of EFMs in human metabolism and proposed a new protocol to integrate gene expression data, spotting key “characteristic” EFMs in different scenarios. Our approach was successfully applied to identify metabolic differences among several human healthy tissues.

The next natural step is to evaluate the performance of our approach in a more clinical interesting situation. For this purpose, we here focused on lung cancer, as its major subtypes are well defined in the medical literature and therefore it is suitable as proof of concept for our approach.

Currently, lung cancer is divided in two main types of pathologies: Small Cell Lung Cancer (SCLC) and Non Small Cell Lung Cancer (NSCLC). SCLC only represents 15% of detected lung cancers while the remaining 85% are NSCLC. NSCLC are divided in three major subtypes: Adenocarcinoma (AD), Squamous Cell Carcinoma (SQ) and large cell lung carcinomas, which represents 40%, 30% and 10% of the cases, respectively. This classification puts AD and SQ in the most common lung cancer pathologies.

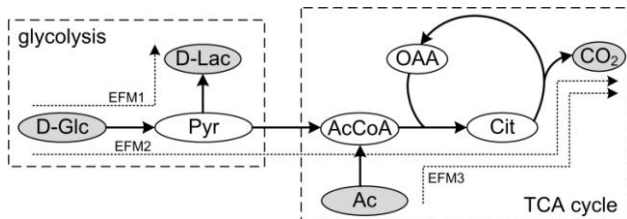
In this article we aim to identify key characteristic EFMs (metabolic pathways) in both AD and SQ NSCLC based

on gene expression data. Ultimately, the objective of our approach is to classify AD and SQ based on their pathways and metabolites. Previous works have directly addressed this question by global gene expression analysis [10]. However, our approach is particularly relevant to define non-invasive means (based on blood or urine samples, for example) able to distinguish between different clinical outcomes. We present below different steps followed to achieve this goal.

## 2. Material and methods

### 2.1. Elementary Flux Modes (EFMs) concept

To illustrate the concept of EFMs, consider Figure 1, which represents a simplified metabolic system involving glycolysis and TCA cycle fed by two possible substrates (glucose and acetate) and excreting lactate and  $\text{CO}_2$ . An EFM is technically a minimal subset of enzymes able to perform in sustained steady state. Steady-state implies that metabolites inside the boundaries of the system (*e.g.* pyruvate) must be in stoichiometric balance, *i.e.* flow in must be equal to flow out. This condition requires the definition of metabolites able to be exchanged outside the system, *i.e.* inputs (*e.g.* glucose) and outputs (*e.g.* lactate). In addition, “minimal” means that the removal of an enzyme leads to disrupting the pathway. In our example in Figure 1, we have 3 EFMs. EFM1 represents anaerobic glycolysis; EFM2 aerobic glycolysis via TCA cycle; EFM3 TCA cycle fed by acetate. It is easy to check that they satisfy the conditions mentioned above. For more technical details, please see [5].



**Figure 1:** Example metabolic network that represent a simplified system involving glycolysis and TCA cycle.

Note that EFMs are minimal modes of behaviour and logically combinations are also possible. However, in different scenarios some of them may prevail over the others. For example, cancer cells produce energy primarily via anaerobic glycolysis (EFM1) even when sufficient oxygen is available (Warburg effect). Note that EFMs typically have different inputs (substrates) and outputs (excreted metabolites). This can be exploited to separate different clinical scenarios.

Finally, the computation of EFMs becomes intractable as the metabolic networks increases. Currently, it is only possible to compute a subset of them via optimization [7,8]. Much research is being accomplished in that direction.

### 2.2. Human metabolic pathway collection

We used a subset of 5875 EFMs previously determined in [9], based on BiGG human metabolic network [6], which involves 2469 reactions and 1587 metabolites. This subset involves a diverse list of metabolic pathways potentially

active in different human physiological conditions. Note here that this subset of EFMs considers not only predefined pathways in the literature and databases, such as KEGG and HumanCyC, but also unreported pathways that may lead to novel metabolic properties.

### 2.3. Gene expression data collection

Gene expression data was extracted from Gene Expression Omnibus (GEO) database. In particular, we considered 58 human NSCLC samples from [10] and 6 healthy human samples from [11]. 40 NSCLC samples were taken from patients clinically diagnosed as Adenocarcinoma (AD), while the other 18 samples from patients with Squamous cell carcinoma (SQ). All these samples were hybridized in an Affymetrix array HGU 133 plus, which contains 54675 probes and 20283 genes.

### 2.4. Absolute expression analysis

We obtained a discrete absolute expression for each gene in our three groups: Adenocarcinoma (AD), Squamous cell carcinoma (SQ) and Control (CN). Absolute expression of genes was classified into three states: high, normal and low. Note that this analysis was conducted for each group separately and no comparison between groups was made, *i.e.* absolute expression of genes is a functional property for each group.

This task has not been widely explored, since the comparison among genes of the same microarray is not direct and straightforward for different technical issues. However, a recent work determined a gene expression barcode for two human microarrays [12], namely HGU 133 A and HGU 133 plus, precisely the one we used in our analysis. Thus, according to this gene expression barcode, we can identify active and inactive genes among the available arrays.

This gene expression barcode outputs two possible states for each probe: active (1) and inactive (0). To obtain a three level classification, as introduced above in our definition of absolute expression, we defined a gene as highly (lowly) expressed if all the probes containing such gene are active (inactive) in all the samples, and moderately expressed otherwise. A less stringent threshold (for example 95% of the probes instead the 100%) can be selected, but the confidence level of gene activity/inactivity will be decreased.

### 2.5. Differential expression analysis

Differential expression analysis focuses on determining which genes have been over-expressed, unchanged or under-expressed between two conditions. From gene expression data considered in subsection 2.3, we analyzed cancer data and determined two sets of differentially expressed genes, namely Adenocarcinoma vs. Squamous (ADvsSQ), and Squamous vs. Adenocarcinoma (SQvsAD). Note that gene expression data from healthy tissues cannot be directly compared with data from cancer tissues as they come from different works and normalization is required, which falls out of the scope of this article.

To do this task, we first assigned an expression value for each gene and sample by calculating the median of the values measured in their associated microarray probes. Then, multiple linear regressions and empirical bayes statistics were applied to determine a p-value indicating the likelihood of a gene not being differentially expressed between both conditions, as standard in the literature. In addition, False Discovery Rate (FDR) approach was applied to correct the effect of multiple hypotheses testing, transforming previous p-values into q-values [13]. Therefore, a unique q-value for each gene is obtained, considering differentially expressed those with a q-value lower than 5%. Note that this threshold is arbitrary. However, the smaller is this threshold, the greater is the confidence level, as in section 2.4. For these tasks, we used limma package in R statistical programme [14]. Once the list of differentially expressed genes is known, determination of over- or under-expression of those genes is straightforward based on linear regression coefficients.

### 2.6. Data transformation: from genes to reactions

Our described methodology leads to a three level classification of the 20283 genes included in the HGU 133 plus array for both differential expression and absolute expression, namely over/highly expressed as '1', unchanged/normally expressed as '0' and under/lowly expressed as '-1'. The next step is to map this gene expression data into the human metabolic network. This can be done using the Boolean laws reported in [6], which describe the set of genes encoding the enzymes (proteins) that catalyze each metabolic reaction. Most recent version of the human metabolic network reconstruction [6] annotates 1496 genes involved in the regulation of 1504 metabolic reactions. From this set of "metabolic genes", 1451 are found in the HGU 133 plus microarray.

Once applied these Boolean rules, we obtain a three level classification of metabolic reactions, similar to the classification of genes. In other words, based on gene expression data, we classify reactions as over/highly expressed (1), unchanged/normally expressed (0) and under/lowly expressed (-1). Note that reactions involving genes not considered in the microarray will have unknown metabolic expression. Similarly, reactions with unknown regulating genes (approx. 965) will be considered with unknown expression value.

We are aware of that this transformation is limited, since from genes to enzymes and fluxes complex regulatory circuits exist and proteomics and metabolomics would be required. However, with the information at hand, *i.e.* gene expression data, this is optimal way to provide a metabolic interpretation.

### 2.7. Selection of characteristic metabolic pathways

Reaction expression data determined in section 2.6 is mapped into our set of 5876 EFMs in each different scenario, namely using separately differential and absolute expression data. We then applied the methodology presented in [9], to determine a list of characteristic EFMs in each scenario. In that work, the concept of characteristic EFMs was introduced based on

absolute expression data, namely EFMs significantly enriched with highly expressed reactions and with the minimum number of lowly expressed reactions. This concept can be easily extended for differential expression data, *i.e.* EFMs significantly enriched with over-expressed reactions with the minimum number of under-expressed reactions.

In [9], using a three level reaction expression classification, as we are doing here, a q-value is assigned to each EFM based on the probability of involving a greater number of highly (over) expressed reactions and smaller number of lowly (under) expressed reactions by chance. In order to obtain such q-values, this approach considers highly (over) and lowly (under) expressed reactions of a metabolic pathway in the same statistical score. In that work, we selected as characteristic EFMs those with q-value lower than 20%. This value is kept in this work.

## 3. Results

Our methodology described above was here applied to find key metabolic differences between our 2 sub-groups of Non Small-Cell Lung Cancer (NSCLC), namely Adenocarcinoma and Squamous cell carcinoma. Table 1 summarizes results for absolute and differential expression analysis described above. In particular, it details gene and reaction expression, as well as characteristic EFMs, in each scenario. We discuss below the results obtained.

	Genes (1/0/-1)	Metabolic Genes (1/0/-1)	Reactions (1/0/-1)	Char. metab. pathways
AD	1096/9265/9922	146/794/508	181/887/421	165
SQ	1317/8145/10821	162/675/611	240/719/530	207
CN	1353/5131/13799	173/452/823	262/494/733	228
ADvsSQ	1725/16163/2395	175/1096/177	333/1017/139	116
SQvsAD	2395/16163/1725	177/1096/175	267/985/237	267

**Table 1:** Gene and reaction expression analysis of adenocarcinoma (AD), squamous cell carcinoma (SQ) and control (CN) tissues.

### 3.1. NSCLC and Control expression data

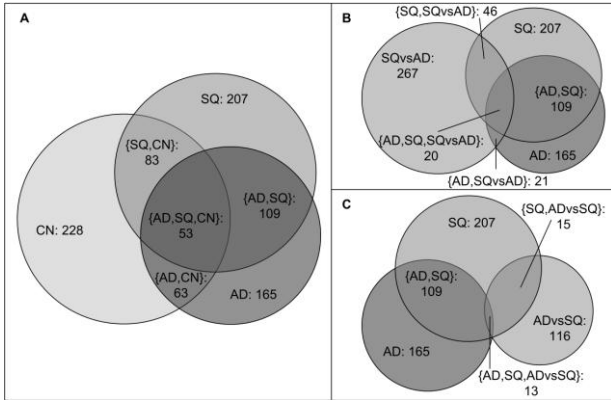
Table 1 details absolute gene and reaction expression in our 3 scenarios: AD NSCLC, SQ NSCLC and CN. Interestingly, the number of lowly expressed (metabolic) genes and reactions in CN is significantly higher than in the other two cancer scenarios. This may suggest that the reprogramming of cancer metabolism enhances systemic robustness, probably activating silenced pathways that guarantee and optimize proliferation. Note however that to a much lesser extent, the number of highly expressed reactions is higher in CN. This may indicate that metabolism in CN is more specific than in cancer and presents less variability across samples in the active set of enzymes.

On the other hand, Table 1 also presents results for differential expression analysis between AD and SQ NSCLC. Although the number of over and under expressed (metabolic) genes is similar, in terms of reactions the metabolism of AD overall presents more activity than SQ.

### 3.2. NSCLC and Control characteristic pathways

Table 1 shows the number of characteristic of EFMs in each of our scenarios. As discussed in [9], the number of characteristic EFMs is not necessarily proportional to the number of highly/over and lowly/under expressed reactions, as they may be more disperse among the network. For this reason, it is not surprising that more characteristic EFMs are found in SQvsAD.

In order to interpret the resulting subsets of characteristic EFMs Figure 2 shows the degree of overlapping found in different scenarios via Venn diagrams.



**Figure 2:** Venn diagrams proportional to the number of characteristic pathways contained in **A)** AD, SQ, CN; **B)** AD, SQ, ADvsSQ; and **C)** AD, SQ and SQvsAD.

In Figure 2A metabolic activity of AD, SQ and CN is clearly distinguished, which illustrates an appropriate performance of our approach. As partially expected, it can be observed that the metabolic activity in cancer tissues (AD and SQ) is more similar between them than to healthy tissue (CN). In addition, a common subset of 56 pathways is found only in AD and SQ, which may represent a core metabolic network of lung cancer.

In order to identify the most relevant EFMs in AD and SQ, we introduced differentially over-expressed EFMs in AD and SQ, as shown in Figures 2B and 2C, respectively. EFMs of key interest are those characteristic of a cancer subtype and additionally differentially over-expressed in that same sub-type. In particular, in Figure 2A, we found 46 EFMs belonging to SQ and SQvsAD. Out of these 46 EFMs, 20 are also active in AD. Note that one clear false positive also arose, *i.e.* an EFM differentially expressed in SQ appears active only in AD. On the other hand, we found 13 EFMs belonging to AD and ADvsSQ (see Figure 2C). These 13 pathways also appear in SQ group. We also detected 2 false positives similarly as above.

### 4. Conclusions

In this article we identified key characteristic EFMs in both AD and SQ NSCLC based on gene expression data. Results are promising and further analysis is required to study 46 EFMs in SQ and 13 EFMs in AD. As noted above, the final objective is to obtain differential inputs and outputs metabolites between these two conditions. This work constitutes the starting point to establish a new methodology that allows us to distinguish key metabolic processes among different clinical outcomes.

### Acknowledgments

AR acknowledges the funding from Asociación de Amigos de la Universidad de Navarra; JP from Basque Government.

### References

- [1] Goffard, N, Weiller, G. PathExpress: a web-based tool to identify relevant pathways in gene expression data. *Nucleic Acids Res.*, vol 35, 2007, pp W176-81.
- [2] Pey, J, Prada, J, Beasley, JE, et al. Path finding methods accounting for stoichiometry in metabolic networks. *Genome Biol.*, vol 12, 2011, pp R49.
- [3] Schwartz, J-M, Gauguier, C, Nacher, JC, et al. Observing metabolic functions at the genome scale. *Genome Biol.*, vol 8, 2007, pp R123.
- [4] Wessely, F, Bartl, M, Guthke, R, et al. Optimal regulatory strategies for metabolic pathways in Escherichia coli depending on protein costs. *Molecular systems biology*, vol 7, 2011, pp 515.
- [5] Schuster, S, Fell, DA, Dandekar, T. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat. Biotech.*, vol 18, 2000, pp 326-32.
- [6] Duarte, NC, Becker, SA, Jamshidi, N, et al. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proceedings of the National Academy of Sciences of the United States of America*, vol 104, 2007, pp 1777-82.
- [7] de Figueiredo, LF, Podhorski, A, Rubio, A, et al. Computing the shortest elementary flux modes in genome-scale metabolic networks. *Bioinformatics*, vol 25, 2009, pp 3158-65.
- [8] Rezola, A, de Figueiredo, LF, Brock, M, et al. Exploring metabolic pathways in genome-scale networks via generating flux modes. *Bioinformatics*, vol 27, 2011, pp 534-40.
- [9] Rezola, A, Pey, J, Figueiredo, LFD, et al. Interpreting high-throughput expression data of human tissues in light of elementary flux modes, *Metabolic engineering* (under review), 2012.
- [10] Kuner, R, Muley, T, Meister, M, et al. Global gene expression analysis reveals specific patterns of cell junctions in non-small cell lung cancer subtypes. *Lung cancer*, vol 63, 2009, pp 32-8.
- [11] Crouser, ED, Culver, DA, Knox, KS, et al. Gene expression profiling identifies MMP-12 and ADAMDEC1 as potential pathogenic mediators of pulmonary sarcoidosis. *American journal of respiratory and critical care medicine*, vol 179, 2009, pp 929-38.
- [12] McCall, MN, Uppal, K, Jaffee, HA, et al. The Gene Expression Barcode: leveraging public data repositories to begin cataloging the human and murine transcriptomes. *Nucleic Acids Res.*, vol 39, 2011, pp D1011-5.
- [13] Storey, JD, Tibshirani, R. Statistical significance for genomewide studies. *Proc Natl Acad Sci USA*, vol 100, 2003, pp 9440-5.
- [14] Smyth, G. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular Biology*, vol 3, 2004.