BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 23 2010, pages 2969–2976
doi:10.1093/bioinformatics/btq567

Systems biology

Advance Access publication October 6, 2010

Pathway Activity Profiling (PAPi): from the metabolite profile to the metabolic pathway activity

Raphael B. M. Aggio, Katya Ruggiero and Silas Granato Villas-Bôas*
School of Biological Sciences, The University of Auckland, 3A Symonds Street, Private Bag 92019, Auckland 1142, New Zealand

Associate Editor: John Quackenbush

ABSTRACT

Motivation: Metabolomics is one of the most recent omics-technologies and uses robust analytical techniques to screen low molecular mass metabolites in biological samples. It has evolved very quickly during the last decade. However, metabolomics datasets are considered highly complex when used to relate metabolite levels to metabolic pathway activity. Despite recent developments in bioinformatics, which have improved the quality of metabolomics data, there is still no straightforward method capable of correlating metabolite level to the activity of different metabolic pathways operating within the cells. Thus, this kind of analysis still depends on extremely laborious and time-consuming processes.

Results: Here, we present a new algorithm Pathway Activity Profiling (PAPi) with which we are able to compare metabolic pathway activities from metabolite profiles. The applicability and potential of PAPi was demonstrated using a previously published data from the yeast *Saccharomyces cerevisiae*. PAPi was able to support the biological interpretations of the previously published observations and, in addition, generated new hypotheses in a straightforward manner. However, PAPi is time consuming to perform manually. Thus, we also present here a new R-software package (PAPi) which implements the PAPi algorithm and facilitates its usage to quickly compare metabolic pathways activities between different experimental conditions. Using the identified metabolites and their respective abundances as input, the PAPi package calculates pathways' Activity Scores, which represents the potential metabolic pathways activities and allows their comparison between conditions. PAPi also performs principal components analysis and analysis of variance or *t*-test to investigate differences in activity level between experimental conditions. In addition, PAPi generates comparative graphs highlighting up- and down-regulated pathway activity.

Availability: These datasets are available in http://www.4shared.com/file/hTWyndYU/extra.html and http://www.4shared.com/file/VbQIIDeu/intra.html. PAPi package is available in: http://www.4shared.com/file/s0uIYWIg/PAPi_10.html

Contact: s.villas-boas@auckland.ac.nz

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on July 20, 2010; revised on September 6, 2010; accepted on October 1, 2010

*To whom correspondence should be addressed.

1 INTRODUCTION

Metabolomics is one of the newest omics technologies and has been evolving rapidly during recent years. Combined with robust analytical methods, metabolomics is capable of screening large numbers of low molecular mass metabolites in biological samples. Metabolites are intermediates of biochemical reactions and are essential in linking different pathways within a biological system. However, metabolites are synthesized and modified by enzymes that are products of gene transcription. Thus, metabolite level is determined by a complex network of reactions in which many regulatory processes involving metabolites, enzymes, mRNAs and genes play an important part. For this reason, metabolomics has been considered essential for the validation of datasets generated by other omics technologies (Çakir *et al.*, 2006) and has been largely applied as a functional genomics tool and as part of systems biology studies (Andersen and Nielsen, 2009; Nielsen and Oliver, 2005; Oliver *et al.*, 1998; Villas-Bôas *et al*, 2004, 2005b, 2008).

However, the convoluted nature of cell metabolism, where the same metabolite can participate in many different pathways, makes the pathway activity analysis, in particular, the most difficult 'omics-data' to interpret (Villas-Bôas *et al.*, 2005a). In addition, as with any other post-genomic technology, metabolomics generates large datasets requiring sophisticated bioinformatics tools for their processing and analysis (Kopka *et al.*, 2005).

Despite the analytical aspects of metabolomics (e.g. quenching of metabolism, metabolite extraction and data acquisition) being well advanced (Dunn, 2008), the correlation between metabolite level and metabolic pathway activity is still considered a complex task to achieve. Consequently, the biological interpretation of metabolomics data remains major bottlenecks in metabolome analysis (Çakir *et al.*, 2006). As a result, to assist with post-genomic data analysis of cell metabolism, a great number of proprietary and open source software packages have been developed during the last 10 years by different companies (e.g. AnalyzerPro® for GC-MS and LC-MS data mining by SpectralWorks Ltd; MarkerView™ for metabolomics and protein/peptide biomarker profiling by Applied Biosystems; Mass Profiler Professional by Agilent Technology) and institutions (e.g. Bioconductor). In addition, many web-based databases are now available and provide important information regarding metabolite diversity, metabolic pathways, biochemical reactions, enzymes and genes (Kopka *et al.*, 2005). Among these, the Kyoto Encyclopedia of Gene and Genomes (KEGG) is one of the most popular databases and it is freely available through http://www.genome.jp/kegg/. In addition, KEGG has application programming interfaces (API) that allow its use by external software. Consequently, several computational tools

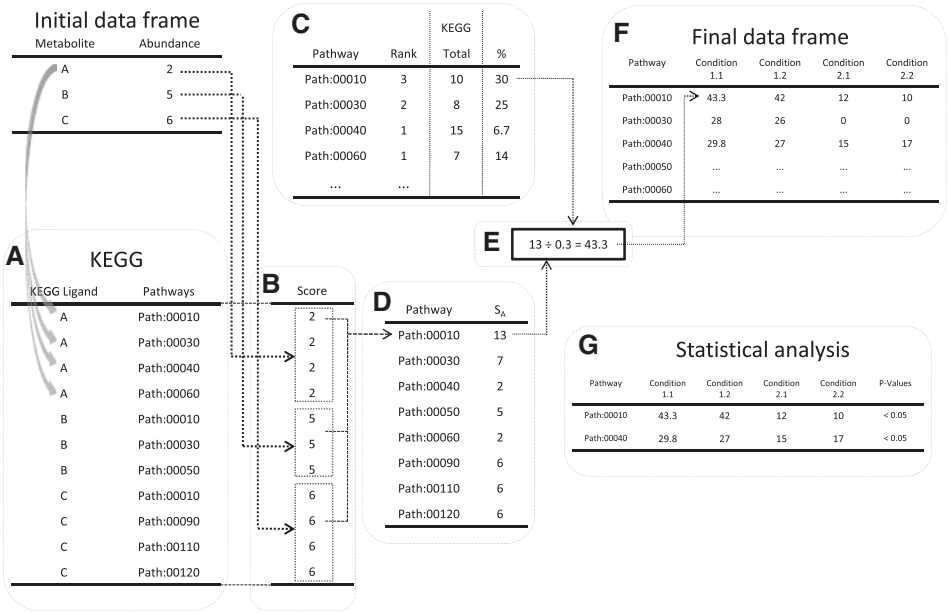

Fig. 1. Description of PAPi algorithm. Starting from a metabolomics dataset (initial data frame), PAPi searchs the KEGG database for potential active metabolic pathways, caculates their AS for different samples and combines the results in a unique frame work. (**A**) All pathways for which each metabolite is known to play a part are collected from the KEGG database. (**B**) Each identified pathway then receives a score based on the abundance/relative abundance of the metabolite to which it is linked. (**C**) The total number of metabolites associated with each pathway is recorded and the pathways are then ranked according to the number of metabolites with which they are associated. The percentage of detected metabolic intermediates is then calculated for each listed pathway. (**D**) Finally, we sum over the scores for each pathway to obtain the total pathway score, S_A, (**E**) and normalize it by dividing by the proportion of metabolites detected from its respective pathway.

have been created to automatically access, extract and manipulate the information contained in these databases (http://www.genome.jp/kegg/soap/; Arita, 2004). R (Ihaka and Gentleman, 1996), an open-source software environment developed for statistical computing (www.r-project.org), is among those with hundreds of available packages developed for different purposes, in particular 'KEGGSOAP' (Zhang and Gentleman, 2009), 'KEGGdb' (Carlson *et al*., 2009) and 'Keggorth' (Carey, 2008) which enable access to and, therefore, use of data from the KEGG database in a flexible way.

These significant advances in bioinformatics tools have improved the quality of both the data generated by omics studies and the subsequent biological interpretations. However, when relating metabolite level to pathway activity, there are only few tools available [e.g. MetPA, Pathway Hunter Tool, Ingenuity Pathway Analysis, Gene Set Enrichment Analysis (GSEA) and Metabolite Set Enrichment Analysis (MSEA)] and most of them require extensive data pre-processing and demand great knowledge about cellular metabolisms, which increases the time-spent and decreases the accessibility to the biological interpretation.

Here, we present a new algorithm Pathway Activity Profiling (PAPi) that, using the metabolite profile and KEGG database, compares the activity of metabolic pathways between different experimental conditions. For this, we defined a new measure for pathway activity, which we called Activity Score (AS). Calculated for each pathway, the AS is based on the number of metabolites identified from each pathway and their relative abundances. As a result, the AS represents the likelihood that a metabolic pathway is active inside the cell and, consequently, allows the comparison of metabolic pathway activities.

However, PAPi is considered time consuming if performed manually. Thus, we developed an R package, PAPi, which implements our new algorithm and facilitates its usage. PAPi uses the data extracted from metabolomics experiments together with the KEGG database to generate relative ASs. PAPi includes functions to perform principal component analysis (PCA) and either a *t*-test or analysis of variance (ANOVA) on the ASs and to generate graphical summaries of the results. The functions also enable the use of (optional) pop-up dialog boxes making them more accessible to new R users.

2 THE ALGORITHM

2.1 Input data

The starting point of PAPi is a data frame containing the KEGG code of the identified metabolites in the first column and their abundances/relative abundances in each sample in the subsequent columns (Supplementary Fig. 1). The KEGG compound code can be found at the KEGG website (http://www.genome.jp/dbget-bin/www_bfind?compound). We assume that data have been normalized (e.g. normalization by internal standard or uncultured medium) and that the data have been appropriately transformed (e.g. log transformation) before being submitted to PAPi.

2.2 Description

To facilitate the algorithm's description, we have divided it into six steps shown in Figure 1. In the first step, the KEGG database is accessed and the pathway(s) associated with each metabolite

are returned. That is, all pathways for which each metabolite is known to play a part are collected from the KEGG database (Fig. 1A). Each identified pathway then receives a score based on the abundance/relative abundance of the metabolite to which it is linked (Fig. 1B). The total number of metabolites associated with each pathway is recorded and the pathways are then ranked according to the number of metabolites with which they are associated. The percentage of detected metabolic intermediates is then calculated for each listed pathway (Fig. 1C). Finally, we sum over the scores for each pathway to obtain the total pathway score, S_A, (Fig. 1D) and normalize it by dividing by the proportion of metabolites detected from its respective pathway (Fig. 1E). The normalized score of each pathway represents the level of its activity inside the cell, where the higher the score the lower the activity. Thus, we define the normalized AS for pathway P as

$$S_A(P) = (r_1 + r_2 + \cdots + r_N)N/k,$$

where

r_i = the relative abundance of metabolite i detected from pathway P,
N = the number of metabolites detected in pathway P, and
k = the total number of metabolites known to play a part in pathway P.

The six operations described above are applied for all samples from each condition studied. Afterwards, the outcome is combined in a final data frame containing the list of all active pathways and their respective normalized scores for each sample. When applied to the analysis of extracellular metabolites (metabolic footprinting), the profile of metabolites should be normalized (subtracted) by the uncultured medium (control sample) before analysis by PAPi. In our example, a two-sample t-test was used to assess pathway differential activity between two conditions, and only those that were statistically significant were retained (Fig. 1G). Note that when three or more conditions are being investigated ANOVA can be applied to test the global hypothesis of a difference between conditions, and then pair-wise comparisons of conditions can be performed for those pathways for which a statistical difference in activity between conditions was declared. Finally, a graph showing all metabolic pathways and their respective normalized scores in both conditions can be generated to represent the data and assist the interpretation (Fig. 2).

If we simply plot the pathways ASs, the resultant graph would be somewhat counterintuitive because AS is inversely related to the predicted pathway activity, i.e. the higher the AS, the lower the predicted pathway activity. Thus, before plotting, we suggest scaling and 'inverting' the pathway ASs. Scaling is performed by setting the AS of one of the conditions to 0, referring to this as the *reference* condition, and then scaling the AS of the other condition relative to the reference, referring to the latter as the *target* condition. We then 'invert' each pathway's AS before plotting, i.e.

$$S'_{A,\text{Reference}}(P) = 0$$

and

$$I_A(P) = \begin{cases} S'_{A,\text{Target}} = -(S'_{A,\text{Target}}/S'_{A,\text{Reference}}) & \text{if } S'_{A,\text{Target}} > S'_{A,\text{Reference}} \\ S'_{A,\text{Target}} = (S'_{A,\text{Reference}}/S'_{A,\text{Target}}) & \text{if } S'_{A,\text{Target}} < S'_{A,\text{Reference}} \end{cases}.$$

For example, if the ASs of the reference and target in pathway P are $S_{A,\text{Reference}}(P) = 3$ and $S_{A,\text{Target}}(P) = 6$, respectively, their scaled scores will be 0 (Reference) and -2 (Target).

While scaling and inverting of pathway ASs is not essential, we recommend their use because they facilitate a more intuitive visualization of the data.

3 RESULTS AND DISCUSSION

Our methodology is based on two main assumptions. We postulate that (i) if a given metabolic pathway is more active in a given condition, a larger number of metabolic intermediates from that pathway is likely to be detected by metabolomics. However, since the same metabolites are usually detected across different conditions, we assume that (ii) the greater the activity of a metabolic pathway, the lower the abundance of the metabolic intermediates from that pathway, because the metabolic flux throughout the pathway is expected to be higher. Higher metabolic flux should result in higher conversion rates of metabolic intermediates inside the cells, reflecting in lower abundances of metabolites produced by that pathway. However, considering that a metabolic pathway presenting lower flux might result in the accumulation of specific intermediates due to lower intracellular conversion rates, the predicted activity of that pathway can be misled by the higher abundances of a few metabolites. Therefore, to minimize this effect, we normalize the pathway ASs by the percentage (%) of metabolites detected from each pathway, because it is assumed (first assumption) that we will detect a higher proportion of metabolic intermediates from a highly active pathway than from pathways of low activity. However, most metabolomics data are based on relative quantification and each metabolite is subject to different response factors depending on the analytical techniques being used. Therefore, unless absolute quantification data is available, PAPi results obtained from relative quantification based metabolomics data can only be used to compare metabolic pathway activity between different data classes. In other words, PAPi results can only be used to predict that pathway A is more active in condition/sample Class I than in condition/sample Class II, but not that pathway A is more active than pathway B. In order to predict that pathway A is more or less active than pathway B, absolute quantification data is required.

Considering the assumptions above, PAPi results in a data frame containing the identified pathways and their respective ASs for each sample. However, this method usually detects over 100 pathways as being potentially active in a cell based on a dataset containing around 50 different identified metabolites and many of these pathways are equally active between different conditions. Thus, statistical analyses are carried out on the ASs to identify pathways that are differentially active between pairs of conditions. A two-sample t-test is performed when only two conditions are being studied, otherwise ANOVA can be performed. The final output of the analysis is a data frame containing the pathways most likely to be defining the differences between data classes and this data frame can be plotted in a line graph.

PAPi also generates unique observations when used to analyze metabolic footprinting data (extracellular metabolites from microbial or cell cultures). For this type of data, we subtract the abundance of metabolites in the uncultured medium from the abundance of metabolites detected in the spent culture medium before applying PAPi (Aggio *et al.*, 2010). This way, some

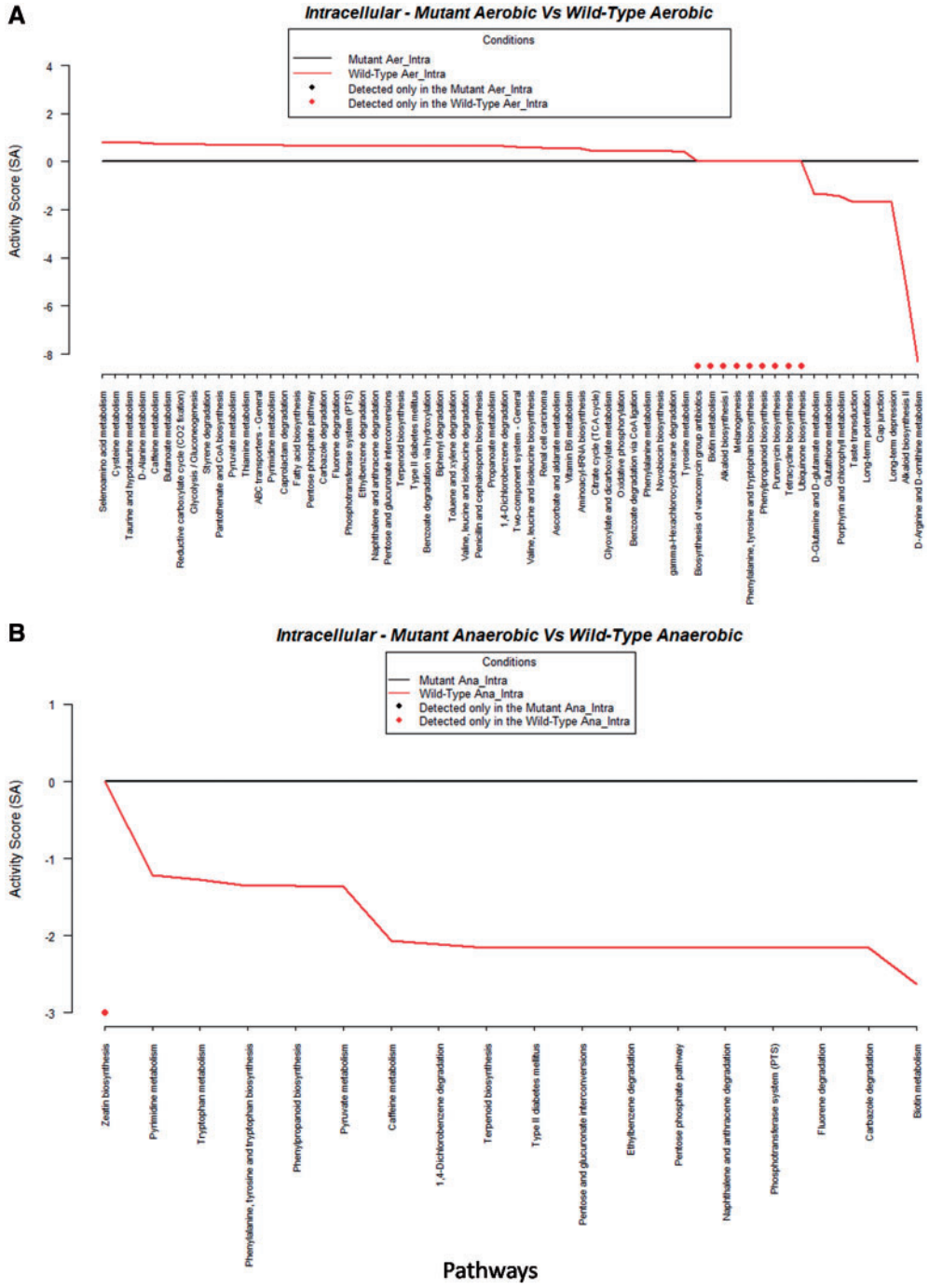

Fig. 2. Comparative metabolic pathway activities of two *S.cerevisiae* strains under different environmental conditions based on intracellular metabolomics data. (**A**) *Saccharomyces cerevisiae* mutant strain (Mutant Aer_Intra) versus wild-type strain (Wild-Type Aer_Intra) under aerobic growth; (**B**) *S.cerevisiae* mutant strain (Mutant Ana_Intra) versus wild-type (Wild-Type Ana_Intra) strain under anaerobic growth. Wild-type (laboratory strain), *S.cerevisiae* CEN.PK.113-7D; Mutant (redox engineered strain), *S.cerevisiae* CEN.MS1.10CT1 (Villas-Bôas *et al*., 2005b).

pathways receive a positive AS while others become negative. A negative pathway AS indicates that metabolites playing part in those pathways were more abundant in the uncultured medium than in the spent culture, suggesting that the activity of those pathways is related to the uptake of metabolites from the medium. A positive score, on the other hand, suggests that metabolic intermediates from those pathways were secreted to the extracellular medium during microbial or cell growth, possibly resulting from a metabolic overflow. This way, important information regarding metabolite uptake and intracellular metabolic overflow is generated, enhancing the biological interpretation of metabolic footprinting data.

Although the assumptions used to build PAPi suit most of the metabolic pathways, the *Glycolysis* pathway seems to work in a distinct way. According to Stephanopoulous and co-workers (Stephanopoulous *et al.* 1998), even when glycolysis is in a high flux state we should expect high abundance of its intermediates, which is pretty reasonable if we consider that Glycolysis is a central metabolic pathway that provides precursors for many essential pathways (e.g. TCA cycle). Thus, we agree that for glycolysis PAPi may not be accurate, but it is still useful in detecting whether glycolysis is operating at different fluxes between experimental conditions.

In addition, it is important to emphasize that by using the non-species-specific KEGG database information we usually observe that pathways not naturally belonging to the organism under study may appear as potentially active by PAPi. This output sounds wrong in principle, but we speculate that it can actually provide important information about possible metabolic interactions between different organisms or species and also about novel metabolic reactions. For instance, extracellular metabolites produced by an organism A can simultaneously play a role in the metabolism of an organism B, which can be a potential metabolic link that allows the interaction between these two organisms. In addition, when a pathway appears as being active it does not mean that the whole pathway is active, but rather specific reactions of that pathway are taking place.

Metabolic pathway activity is directly related to metabolic flux distribution. Thus, pathways presenting lower scores based on intracellular metabolomics data are likely to be operating at high metabolic flux. Thereby, our method not only reduces the time spent on metabolomics data analysis but it may also enable us to compare the metabolic flux of different pathways in different conditions (indirect fluxomics).

3.1 Method validation

To illustrate and validate the use of PAPi, we analyzed a set of yeast metabolomics data published previously by Villas-Bôas *et al*. (2005a) and reanalyzed by Çakir *et al*. (2006).

3.1.1 Villas-Bôas et al. (2005a) The metabolomics data published by Villas-Bôas *et al*. (2005a) consists of intracellular and extracellular metabolite data of two *Saccharomyces cerevisiae* strains: a wild-type laboratorial strain (CEN.PK.113-7D) and a mutant strain (CEN.MS1-10CT1). The mutant was a redox-engineered strain with a deleted NADPH-dependent glutamate dehydrogenase (encoded by *GDH1*) and an over expressed NADH dependent glutamate dehydrogenase (encoded by *GDH2*). The enzyme encoded by GDH1 is considered the major enzyme responsible for nitrogen assimilation during *S.cerevisiae* growth on ammonium as sole nitrogen source, and accounts for a considerable fraction of the NADPH consumed in the cell (Villas-Bôas *et al*., 2005a). Both strains were grown in batch cultures under aerobic and anaerobic conditions using standard minimal mineral medium with glucose as the sole carbon source and ammonium (NH_4^+) as the sole nitrogen source. Due to excellent culture reproducibility, the sample-to-sample variability exceeded flask-to-flask variability; consequently, replicate samples from different shake flasks were treated equivalently. The metabolome dataset analyzed included 15 intracellular and 9 extracellular sample replicates for each experimental condition tested (data classes). In order to deconvolute the peaks of GC-MS spectra and identify the metabolites, we followed the protocol published previously in Aggio *et al*., 2010.

Although the comparison between the wild-type and the mutant strains revealed no differences in growth rates under aerobic (dos Santos *et al*., 2003) and anaerobic (Nissen *et al*., 2000) batch cultivations, PAPi was able to detect the differences between pathway activities due to the differences in the metabolite profile and metabolite abundances obtained by Villas-Bôas *et al.* (2005a). Thus, in order to validate PAPi findings, we compared the pathway profile activity of both strains under the two environmental conditions (aerobic and anaerobic) and correlated the results with observations reported in Villas-Bôas *et al*. (2005a) and Çakir *et al*. (2006) findings.

3.1.2 PAPi results aerobic versus anaerobic cultures: as discussed by Çakir *et al*. (2006), it is expected to find some common effects between the genetic and environmental perturbations. The genetic perturbation (knockout of *GDH1* gene and over expression of *GDH2*) has direct effect on the overall cell balance of $NADPH/NADP^+$ and $NADH/NAD^+$; consequently, the mutation directly affects the cell redox metabolism. On the other hand, the oxygen availability also affects the redox metabolism due to changes in the operation of the TCA cycle and pentose phosphate pathway (PPP). Indeed, Çakir *et al*. (2006) detected reactions commonly changed in both datasets and, and in agreement, PAPi predicted a great number (∼70 %) of metabolic pathways commonly changed in both pair-comparisons (aerobic/anaerobic and wild type/mutant) (Supplementary Table 1).

Furthermore, Villas-Bôas *et al*. (2005a) observed that samples from anaerobic cultivations presented overall higher levels of both intra and extracellular metabolites when compared to aerobic cultivations. In agreement, PAPi also predicted that most metabolic pathways were likely to be operating at lower activity under anaerobic condition (Supplementary Figs 2 and 3), which results in lower biomass biosynthesis and in intracellular accumulation of metabolites, with consequent potential overflow to the extracellular medium.

In addition, despite *S.cerevisiae* not presenting a homologous gene sequence for lactate dehydrogenase, Villas-Bôas *et al*. (2005a) detected high levels of lactate under anaerobic condition for both intra and extracellular samples. High levels of glyoxylate were also detected under anaerobic condition and were subsequently shown to be formed from a novel pathway for glyoxylate biosynthesis in *S.cerevisiae* involving direct deamination of glycine (Villas-Bôas *et al*., 2005b). The results from PAPi (Supplementary Figures 2 and 3) show a significantly lower activity of the *TCA cycle metabolism* (TCA), *pyruvate metabolism* (PM) and *glycine, serine and threonine metabolism* (GSTM) under anaerobic condition. Martins *et al*. (2001) reported that *S.cerevisiae* can synthesize D-lactate via methylglyoxal metabolism that intrinsically linked to PM. Interestingly; methylglyoxal is also a key precursor for the GSTM, which could potentially increase the biosynthesis of glycine. Therefore, we speculate that the methylglyoxal could have been acting as a link between PM and GSTM, where lactate and glyoxylate were the main metabolic products resultant from these pathways. In other words, the predicted low flux in the TCA cycle under anaerobic condition (Çakir *et al*., 2006) resulted in the accumulation of intermediates from PM and consequently increased formation of methylglyoxal. The higher amount of methylglyoxal available in the cell could have increased the formation of D-lactate and also increased the biosynthesis of glycine through GSTM. However, due to reduced incorporation of amino acids into biomass

under anaerobic growth, GSTM flux may have been repressed. Thus, the accumulated levels of free amino acids from GSTM such as glycine could have been preferentially converted to other metabolites such as glyoxylate via glycine deaminase reaction that does not seem to be repressed by glucose and is active under both aerobic and anaerobic conditions (Villas-Bôas *et al*., 2005b) (Supplementary Fig. 4).

Villas-Bôas *et al*. (2005a) also observed the presence of myristic acid in the extracellular samples of anaerobic cultures. In agreement, PAPi predicted a lower activity of both fatty acid biosynthesis (FAB) and fatty acid metabolism (FAM) pathways under anaerobic conditions. Myristic acid is an intermediate metabolite of FAB and is related to biomass formation, since it can react with glycerol to form lipids required for the cellular membrane structure. Myristic acid may have accumulated due to higher availability of precursors for FAB in the central carbon metabolism (e.g. acetyl-CoA) and the reduced requirement of lipids and fatty acids for biomass biosynthesis. Therefore, an accumulation and potential overflow of myristic acid into the extracellular medium is a potential result of the low activity of FAB and FAM pathways predicted by PAPi due to the lower biomass formation under anaerobic conditions.

3.1.3 Wild-type culture versus mutant culture. Villas-Bôas *et al*. (2005a) observed two distinct patterns in metabolite profiles between wild-type and mutant strains. Aerobically, the mutant presented higher levels of many metabolites while the opposite was observed anaerobically. According to our assumptions, a lower pathway activity means low inter-conversion rates of metabolites and potential accumulation of some of its intermediates. On the other hand, a high pathway activity means high inter-conversion rates of metabolites and consequently less abundance of its intermediates. As a result, PAPi predicted lower activity for most pathways in the mutant when grown under aerobic conditions and higher activity was observed under anaerobiosis (Fig. 2).

As discussed by Villas-Bôas *et al*. (2005a) and Çakir *et al*. (2006), due to the deletion of *GDH1*, the mutant strain is expected to present difficulties in assimilating ammonium into glutamate, a key reaction for nitrogen metabolism in the central carbon metabolism. Thus, low intracellular levels of glutamate potentially decrease the activity of pathways derived from glutamate, such as Pyrimidine metabolism and Butanoate metabolism, which was indeed predicted by PAPi.

Çakir *et al*. (2006) also detected a reduced difference in metabolic activity between the mutant and wild type when grown under anaerobic conditions. Accordingly, we observed considerably fewer metabolic pathways presenting significant differences between wild-type and mutant during anaerobic growth (Fig. 2). Thus, the PAPi results also support these predictions made by Çakir *et al*. (2006).

Higher intracellular levels of 2-oxoglutarate was found in the mutant samples during aerobic growth (Villas-Bôas *et al*. 2005a) and Çakir *et al*. (2006) suggested that the activity of alanine aspartate transaminase (AAT) that catalyzes the conversion of oxalacetate to aspartate, could be potentially altered in the mutant under aerobic conditions. Interestingly, PAPi predicted that the *reductive carboxylate cycle* (RCC), a pathway intrinsically connected to Alanine, aspartate and glutamate metabolism through oxalacetate and by the reaction converting pyruvate in L-alanine, was less active in the mutant during aerobic growth (Fig. 2). As suggested by Çakir *et al*. (2006), oxalacetate could have been converted to aspartate. However, we argue that the high level of 2-oxoglutarate and other intermediates of the RCC pathway may favor the formation of pyruvate that can then be converted to L-alanine by alanine dehydrogenase (EC 1.4.1.1). Surprisingly, L-alanine dehydrogenase uses ammonium and NADH as cofactor to convert pyruvate into alanine. Therefore, the deletion of the *GDH1* and overexpression of *GDH2* could have led to an up-regulation of the *reductive carboxylate cycle* pathway with alanine dehydrogenase acting as a secondary pathway for assimilation of nitrogen from ammonium. Although there is still no description of alanine dehydrogenase in *S.cerevisiae*, this enzyme is well-described for bacteria and for which it is known to be strongly related to nitrogen assimilation from ammonium. By using the Blast algorithm through the *Saccharomyces* Genome Database (www.yeastgenome.org) we detected a potential homology sequence of 57% match between the *Haemophilus parasuis* alanine dehydrogenase nucleotide sequence and the region YDR211W (GCD6) of the *S.cerevisiae*, showing that *S.cerevisiae* could possibly transcribe an enzyme with similar amino acid sequence. In addition, Burk *et al*. (2007) demonstrated that there is homology between the saccharopine dehydrogenase of *S.cerevisiae* and the alanine dehydrogenase present in the cyanobacteria *Phormidium lapideum*. It is well known that saccharopine dehydrogenase catalyzes the reversible pyridine nucleotide-dependent oxidative deamination of saccharopine to yield L-lysine and 2-oxoglutarate; however, its homology to alanine dehydrogenase suggests that the nitrogen assimilation from ammonium can potentially also be catalyzed by saccharopine dehydrogenase. Thus, the nitrogen assimilation through the conversion of pyruvate to L-alanine appears to exist in the *S.cerevisiae* metabolism. Consequently, using PAPi we detected a potential secondary nitrogen assimilation pathway not previously described for *S.cerevisiae* and not observed by Villas-Bôas *et al*. (2005a) and Çakir *et al*. (2006). Future experiments are required to confirm this hypothesis.

4 THE PACKAGE

The `PAPi` package comprises four functions: `papi, papi.pca, papi.htest` and `papi.line. papi` is used to calculate the ASs for each sample and stores these, together with the list of all active pathways, in a data frame. [Output generated by the papi function can be demonstrated using `data(demo.results)`.] `papi.pca` generates biplots that can be used to verify sample reproducibility and to explore the pathways that may be responsible for the differences between experimental conditions. `papi.htest` performs either a *t*-test or ANOVA on the ASs and can be used to identify pathways that are differentially active, since these are generally thought to be important in terms of observed differences in metabolism between different conditions. Finally, the results can be summarized using `papi.line` that generates a line graph of average AS plotted against pathway. Pathway activity profiles of all experimental conditions can be superimposed on the same graph. (See Supplementary Fig. 6A, B).

`PAPi` can be applied to the analysis of intra- (i.e. metabolic fingerprinting) and extracellular metabolites (i.e. metabolic footprinting).

4.1 Requirements

`PAPi` was developed under R version 2.10.1 and depends on six other packages, namely KEGGSOAP (Zhang and Gentleman, 2009)

and KEGG.db (Carlson *et al.*, 2009) from Bioconductor, and reshape (Wickham, 2007), gdata (Warnes, 2010), gplots (Warnes, 2009) and plotrix (Lemon, 2010). All of these packages can be installed from the Bioconductor database (http://www.bioconductor.org/) using:

```
>source('http://bioconductor.org/biocLite.R')

>biocLite(c('KEGGSOAP','KEGG.db',
'reshape','gdata','gplots','plotrix'))
```

4.2 Descriptions

We now describe the usage of the four functions in PAPi: `papi`, `papi.pca`, `papi.htest` and `papi.line`.

- **`papi(conditions, data = 'import', out.folder = 'popup')`**

`papi` is applied to a data frame in which the first column contains the identified metabolites' KEGG codes and all subsequent columns contain their abundances in each analyzed sample. (The KEGG codes can be found at http://www.genome.jp/dbget-bin/www_bfind?compound and consist of the letter 'C' followed by a sequence of 5 digits, e.g. Glucose = C00031.) For a sample data frame showing the layout required by `papi`, see `data(demo)`.

`papi` comprises three arguments: `conditions`, `data` and `out.folder`. The `conditions` argument takes a character vector of treatment names, e.g. `conditions = c('cond1', 'cond2')`. The `data = 'import'` (default) argument results in a pop up dialog box, allowing the user to click-and-point to the comma-separated value (CSV) format file from which the data is to be read. Alternatively, `data` can take the name of a data frame containing the samples' metabolite abundances. Similarly, the default behavior of `out.folder` is for a pop-up dialog box to be presented to the user. The user can then select the directory to which the results will be saved. Alternatively, `out.folder` takes a character string naming the path to the directory where the results are to be saved. Since there are no restrictions on the column names of the initial input data frame, during the execution all PAPi functions the user is presented with a dialog box from which he/she must select the columns of the data frame associated with each experimental condition.

`papi` generates a data frame containing the pathways identified across all samples, their KEGG codes and their ASs. In addition, `papi` calculates the average and standard error of the ASs by condition for each pathway, and these are stored in the same data frame (See `data(demo.results)` for an example.). This data frame is automatically saved to a file named `papi_results.csv` in the directory specified by `out.folder`.

- **`papi.htest(conditions, data = 'import', signif.level = 0.05, save = TRUE, out.folder = 'popup')`**

The `conditions`, `data` and `out.folder` arguments behave as described in the `papi` function. If `conditions` is of length 2 (i.e. there are only two experimental conditions) then a *t*-test is used to test for differential pathway activity between conditions, otherwise ANOVA is used. A column of *P*-values resulting from the analyses carried out on the ASs from each pathway is added to the initial input data frame. The level of significance, specified by the `signif.level` argument, is used to create a data frame consisting of only the differentially active pathways. When the argument `save = TRUE` (default) this data frame is saved to a CSV file called `papi_anova.csv`.

- **`papi.pca(graph.name, data = 'import', loadings.code = TRUE, save = TRUE, out.folder = 'popup')`**

The `data` and `out.folder` arguments again behave as described in the papi function, and the save argument as in `papi.htest`. `papi.pca` can be used to apply PCA to the data frames generated by `papi` and/or `papi.htest`. Single or multiple biplots, one per graphical device, can be generated. The number is determined by the length of the character vector supplied to `graph.name`. If `loadings.code = TRUE`, the KEGG code of identified pathways will be used as loadings, otherwise the pathways' names will be used. If `save = TRUE` (default), a file containing a PCA biplot will be generated for each graph with the name format `PCAgraph.name.png`. For example, if `graph.names = c('cond1', 'cond2', 'cond1+cond2')` then three files will be generated, namely `PCAcond1.png`, `PCAcond2.png` and `PCAcond1+cond2.png`. (See Supplementary Fig. 6C and D)

- **`papi.line(conditions, data = 'import', relative = TRUE, save = TRUE, out.folder = 'popup')`**

The `conditions`, `data`, save and `out.folder` arguments behave as described in the `papi.htest`. `papi.line` generates a line graph in which, for each condition, the average normalized total ASs are plotted against the identified pathways (see Supplementary Fig. 6A, B). The `relative = TRUE` (default) argument results in a line graph of pathway activity profiles for each target condition relative to the selected reference condition. By default, the first element of the `conditions` argument is defined as the reference condition. Pathways not identified in the reference condition are excluded from the plot. A horizontal line is drawn at zero to serve as a visual reference (see Supplementary Fig. 6B). Plots of the untransformed ASs can be generated setting `relative = FALSE`. If `save = TRUE`, the line graph is saved to a file called `papi_line_graph.png` in the directory specified by `out.folder`.

5 CONCLUSION

We introduced and validated here a new algorithm (PAPi) and an R package (PAPi) able to correlate metabolomics datasets and metabolic pathways activities in a high-throughput way. Doing so, we are able to compare the activity of metabolic pathways under different conditions, which provides great support for hypothesis generation and facilitates the biological interpretation. Furthermore, as KEGG database also supports enquires using protein and gene transcription levels, PAPi has the potential to combine data from different omics in one unique framework, further simplifying the biological interpretation of the data.

ACKNOWLEDGEMENTS

We thank Gregory Cook, Per Bruheim, Sui Lee, Jeremy van Houtte, Vidya Washington, Nicole Anfang, Xavier Duportet, Kevin Chang, Victor Obolonkin and Liam Fenlay for fruitful discussions and suggestions; and Jens Nielsen for critical reading of the article.

Funding: The Health Research Council of New Zealand (HRC).

Conflict of Interest: none declared.

REFERENCES

Aggio,R. *et al*. (2010) Analytical platform for metabolome analysis of microbial cells using methyl chloroformate derivatization followed by gas chromatography-mass spectrometry. *Nat. Protoc.*, **5**, 1709–1729.

Andersen,M.R. and Nielsen,J. (2009) Current status of systems biology in Aspergilli. *Fungal Genet. Biol.*, **46**, S180–S190.

Arita,M. (2004) Computational resources for metabolomics. *Brief. Funct. Genomics*, **3**, 84–93.

Çakir,T. *et al*. (2006) Integration of metabolome data with metabolic networks reveals reporter reactions. *Mol. Syst. Biol.*, **2**, 50.

Carlson,M. (2009) KEGG.db: A set of annotation maps for KEGG. R package version 2.3.5.

Dunn,W.B. (2008) Current trends and future requirements for the mass spectrometric investigation of microbial, mammalian and plant metabolomes. *Phys. Biol.*, **5**, 11001.

Gentleman,R. *et al*. (2004) Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.

Ihaka,R. and Gentleman,R. (1996) R: a language for data analysis and graphics. *J. Comput. Graph. Stat.*, **5**, 299–314.

Kopka,J. *et al*. (2005) GMD@CSB.DB: the Golm Metabolome Database. *Bioinformatics*, **21**, 1635–1638.

Lemon,J. (2010) Plotrix: a package in the red light district of R. *R-News*, **6**: 8–12.

Nielsen,J. and Oliver,S. (2005) The next wave in metabolome analysis. *Trends Biotechnol.*, **23**, 544–546.

Oliver,S.G. *et al*. (1998) Systematic functional analysis of the yeast genome. *Trends Biotechnol.*, **16**, 373–378.

Stephanopoulos,G.N. *et al*. (1998) *Metabolic engineering: principles and methodologies*. Academic Press, Orlando, Florida.

Villas-Bôas,S.G. *et al*. (2004) Mass spectrometry in metabolome analysis. *Mass Spectrom. Rev.*, **24**, 613–646.

Villas-Bôas,S.G. *et al*. (2005a) High-throughput metabolic state analysis: the missing link in integrated functional genomics of yeasts. *Biochem. J.*, **388**, 669–677.

Villas-Bôas,S.G. *et al*. (2005b) Biosynthesis of glyoxylate from glycine in Saccharomyces cerevisiae. *FEMS Yeast Res.*, **5**, 703–709.

Villas-Bôas,S.G. *et al*. (2008) Phenotypic characterization of transposon-inserted mutants of Clostridium proteoclasticum B316(T) using extracellular metabolomics. *J. Biotechnol.*, **134**, 55–63.

Warnes,G.R. (2009) gplots: various R programming tools for plotting data. R package version 2.7.4. http://CRAN.R-project.org/package=gplots.

Warnes,G.R. (2010) gdata: various R programming tools for data manipulation. R package version 2.7.1. http://CRAN.R-project.org/package=gdata.

Wickham,H. (2007) Reshaping data with the reshape package. *J. Stat. Softw.*, **21**, 1–20.

Zhang,J. and Gentleman,R. (2009) KEGGSOAP: Client-side SOAP access KEGG. R package version 1.20.0.