Bayesian Inference of Differential Metabolic Fluxes in Ulcerative Colitis using High-throughput Transcriptomics Data

Guy Karlebach[1]

[1]Department of Computer Science, Fitchburg State University, Pearl Street, Fitchburg, 01420, MA, USA.

Contributing authors: gkarleba@fitchburgstate.edu;

Abstract

Advances in high-throughput technology have enabled the measurement of transcriptomic responses to a broad range of conditions. Our increased understanding of human metabolic reactions has also enabled the construction of large enzymatic network models. With the availability of detailed models and data to fit them to, various methods for analyzing the effects of transcriptional regulation on metabolism have been proposed. While for conditions under which our understanding of the transcriptional response is limited, constructing a quantitative model can prove challenging, knowledge about metabolic constraints can still be useful in interpreting transcriptional changes. In this work, we suggest a novel Bayesian approach that combines systems-level modeling of metabolism with Bayesian inference of the transcriptional response. Our method combines Bayesian reasoning with flux balance constraints that are used in mechanistic models of metabolism, and is thereby able to incorporate both uncertainty and systems-level metabolic knowledge into the model. It models only metabolic differences between two conditions, thereby simplifying the model to capture only effects that are derived from the transcriptomic response to a specific stimulus. We demonstrate the usefulness of our approach by studying the transcriptional regulation of metabolism in two independent ulcerative colitis datasets.

Keywords: Transcriptional Regulation, Bayesian Inference, Differential Metabolic Flux

1 Introduction

Transcriptional regulatory mechanisms are a key factor in determining the metabolic activity of cells under different conditions. For example, cancer cells, which rely on energy produced via glycolysis, modify the expression of glycolytic enzymes [1]. To prevent DNA damage from reactive oxygen species (ROS), they must keep intracellular ROS below a toxic threshold, otherwise cell death will result [2]. Therefore, cancer cell also regulate the pentose phosphate pathway to combat oxidative stress in addition to necessary anabolic processes [3]. While cancer cells differ from healthy ones by other characteristic modifications such as sustained proliferative signaling, changes in metabolism are recognized as a one of the hallmarks of cancer[4]. For a review of metabolic pathway activity in cancer see [5]. Transcriptional regulation of metabolism has also been studied in normal cells. For example, distinct expression profiles corresponding to different gastric sites have been identified, reflecting different metabolic processes [6]. The influence between transcription and metabolism is likely bidirectional. For example, the metabolite alpha-ketoglutarate has been observed to induce transcriptional programs in cancer cells that resemble those of pre-cancer cells [7]. Hence, in order to further our understanding of biological networks and their role in disease it is of high importance to model the interactions between transcription and metabolism.

Metabolomics data are difficult to interpret using traditional approaches. Lee et al. describe the limitations of using metabolomics data in pathway analysis, and call for new algorithms that accommodate the unique properties of metabolites [8]. On the other hand, high-throughout transcriptomics has been used extensively for explaining regulatory changes under various conditions, including those that affect metabolism [9]. GO enrichment analysis identifies significant over-representation of predefined categories of genes in those genes selected as significantly changing in expression[10]. Gene Set Enrichment Analysis (GSEA) identifies high ranking gene pathways using the ranking obtained by testing association with a condition without a cutoff threshold [11]. Both approaches do not evaluate interactions between genes and are therefore applicable to a broad range of datasets, though they do not model metabolic processes directly. Pathway structure has also been used in tandem with gene expression. SPIA scores a pathway based on both overrepresentation of its genes in the differentially expressed set and the significance of their interactions in the pathway graph [12]. BRAS models a pathway using a Bayesian network, a directed acyclic graphical mode, using nodes to represent expression levels and edges to represent relationships between genes [13]. clipper uses graph decomposition of pathways to identify those pathways associated with gene expression differences and the signal paths that have the greatest association with the change [14]. A more metabolism-specific approach, Flux Balance Analysis (FBA) models the metabolic network by assuming flux balance and maximizing an objective such as biomass production, with additional constraints on reaction rates [15, 16]. This requires a detailed knowledge of metabolic reactions at the genome scale, but not of their rates, which are estimated by solving the optimization problem. FBA models have been expanded to account for transcriptional regulation [17, 18]. If an explicit gene regulatory network model is used, both knowledge of the metabolic network and the gene regulatory network are required. Expression data can also be

used to calibrate the metabolic network in FBA, as in [19–21]. The goal is typically to predict a quantitative phenotype rather than to refine the qualitative interpretation of differential expression analysis.
Other data types, like growth rate of cells, can be useful for modeling metabolic changes. While it is possible to measure both the transcriptome and growth rate of single cells [22], such datasets are not common.
In this work we present a novel method that combines Bayesian inference and metabolic constraints to identify differential metabolic fluxes between two conditions, i.e. only those fluxes by which the conditions differ. Our approach does not require a point estimate of reaction rates, and can use posterior error probabilities from Bayesian differential expression analysis to select the most likely fluxes. By combining Bayesian inference with metabolic constraints we are able to eliminate solutions that are not mechanistically sound and at the same time obtain inferential rigor. We do not aim to predict quantitative phenotypes - our goal is to enhance the basic approach to enrichment analysis. We demonstrate the power of the method by modeling differential fluxes in two independent expression datasets with a high degree of agreement.
The next section details how an optimal model for differential fluxes is computed. The Results section then describes application to data from ulcerative colitis patients and healthy controls.

2 Materials and Methods

The input for this analysis are two matrices, S and X. S is a stoichiometric matrix of dimensions $n \times m$, where the n is the number of metabolites considered in the analysis and m is the number of genes/enzymes/reactions that consume and produce metabolites. X is a matrix of gene expression measurements of dimensions $m \times (k_1 + k_2)$, where m is the number of genes/enzymes/reactions and k_1 and k_2 are the number of control samples and case samples, respectively. The matrix X is used for deriving probabilities of differential expression for each individual gene, a problem that has been studied extensively, and so we will only discuss the procedure for doing so briefly and refer to our previous work for full details. The matrix S is used in the optimization process that is the main novelty of this work, as will be described in this section. The output of the analysis is a a subset of the genes, i.e. a set of genes $D \subseteq \{1, 2, ..., m\}$, that represents genes which are associated with metabolic fluxes relevant to the phenotypic difference between cases and controls.
The goal of our analysis is to choose a set of genes/enzymes that together define a set of differential metabolic fluxes between cases and controls. A flux defines the rate of production of a given metabolite. Each gene carries out one or more reactions that consumes some metabolites and produces others. The production and consumption of metabolites by genes can be formally described using a stoichiometric matrix S: each row corresponds to a metabolite and reach column to a gene/reaction. When $S[i, j] < 0$, it represents the number of metabolite molecules of type i that are consumed by reaction j. When $S[i, j] > 0$, it represents the number of metabolites of type i that are produced by reaction j. Consequently, if $\vec{v}$ is a vector of reaction rates, the product $S \cdot v$ results in a vector $\vec{f}$ where the ith entry corresponds to the rate of

production of the metabolites corresponding to the ith row of S. When $\vec{f}$ is the zero vector, we say that flux balance is maintained.
Some of the genes are associated with difference in fluxes between two conditions/groups, cases and controls, that affect the phenotypic differences. As we explain later in this section, probabilities that indicate how likely a gene is to contribute to phenotypic differences can be assigned based in experimental measurements of gene expression in multiple subjects from the two groups. In other words, each gene g is assigned a probability $P(g)$ for producing and consuming metabolites that are relevant to the differences between cases and control. Often, these probabilities are assumed to be completely independent, and the criterion for selecting the relevant genes is then a simply trade-off between true and false positives. Here we add a dependency between the individual gene probabilities, and look instead for the set of genes with the maximal joint probability. The dependency requires that the genes will maintain flux-balance. More formally, if $\vec{G}$ is a vector of indictor variables that represent which genes ate included in the solution, where genes correspond to indices of $\vec{G}$, the probability of the corresponding solution is proportional to:

$$P(\vec{G}) = \begin{cases} \prod_{j:\vec{G}[j]\neq 0} P(j) \cdot \prod_{j:\vec{G}[j]=0} 1 - P(j) & \exists \vec{v}: \ \forall i \ \vec{G}[i] < \vec{v}[i] \leq \vec{G}[i] \cdot \max(\vec{v}) \ \textit{and} \ S \cdot \vec{v} = 0 \\ 0 & \textit{Otherwise} \end{cases} \tag{1}$$

where $max(\vec{v})$ is the largest value in the vector $\vec{v}$. In words, the probability of a solution is the product of the probabilities $P(g)$ of genes that are included in the solution times the product of $1 - P(g)$ for genes that are not included in the solution, if the included genes preserve flux balance, and otherwise 0. The set of genes selected for the solution is then $D = \{i : \vec{G}[i] == 1\}$. The proportionality constant can be seen as an update to the model probabilities obtained through the differential expression analysis, derived from prior knowledge about metabolism and the expression levels of the genes. In the next part of this section, we explain how to obtain the probabilities $P(g)$ for the individual genes, and how to find the solution with the highest probability that satisfies flux balance.
First, we wish to associate each enzyme with a probability of either performing a reaction more often in cases than in controls, or conversely with a probability of performing a reaction less in cases than in controls. Those are the probabilities P in equation 1. We can compare the mean expression level of each enzyme under the two conditions, where a higher mean indicates increased activity. However, we need to associate the difference with a probability. Karlebach and Robinson [23] have previously proposed a method for differential expression and splicing analysis that produced posterior error probabilities (PEP) - the probability of no effect. No effect for gene i in our context means that $\vec{G}[i] = 0$, and so $PEP = 1 - P(\vec{G}[i] = 0)$. To describe the method of [23] briefly, a mixture prior with two components, each representing the distribution of parameter values under one of the two hypotheses (effect and no effect), is defined based on the region of practical equivalence (ROPE), a posterior summary method which separates the possible parameter values into regions of effect and no effect. The novelty of the approach in [23] is to use the same regions to define the prior. One component of the mixture is concentrated in the

ROPE, the region of parameter values that would be interpreted as no-effect, and the other is spread outside the ROPE. After inferring the posterior, the PEP is the density of the posterior that belongs to the region of no effect. The weights of the mixture prior's components are given a uniform prior themselves and are inferred from the data. Hence in a dataset with a small number of true effects the prior will drive the posterior towards the component of no effect, and when the fraction of true effects is larger so will the weight of the prior component of true effects be. Since this method uses gene expression measurements that often come as discrete counts, it can apply a log-transformation to the counts which will then be modeled using a normal likelihood [23] .

While the method described in this work is independent of the way the PEP is inferred, for the sake of completeness we add that the model used in [23] models differences in the fractions of a single gene's multiple isoforms in addition to differences in gene expression. In order to obtain the probabilities of no-effect (PEP) for each enzyme without the need to consider its isoforms, we modified the model of Karlebach and Robinson [25] such that it would only include differential expression parameters, i.e. the parameter β but not the parameter α which modeled changes in isoform proportions in [23]. β models the difference between two distributions of log-transformed counts, one in case and one in controls. The density of the posterior of beta in the region of practical equivalence is the posterior error probability (PEP), the probability of no-effect.

We now turn to finding the optimal solution, i.e. the one that maximizes the probability given in equation 1. This can be formulated as a mixed-integer linear programming (MILP) problem and solved using a solver. The first constraint of the MILP is that the product of the stoichiometric matrix S and the reaction rates vector $\vec{v}$ is equal to the zero vector, i.e $S \cdot \vec{v} = \vec{0}$. Since some of the enzymes that carry out the reactions are expressed at a higher level in cases and some in controls, we reverse the direction of the reactions corresponding to the latter. Formally, if $\vec{x_1}$ and $\vec{x_2}$ are vectors of (log-transformed) measurements of a gene i expression levels in cases and controls (i.e. $\vec{x_1} = X[i, (k_1 + 1)...(k_1 + k_2)]$ and $\vec{x_2} = X[i, 1...k_1]$), respectively, and $\overrightarrow{x_1} < \overrightarrow{x_2}$, we multiply the column in S which corresponds to this gene and its reaction by -1, i.e. $S[, i] \leftarrow -1 \cdot S[, i]$. The reason for this choice is that if an enzyme is chosen by the model as having a significantly decreased activity in cases, then the products of the reaction it carries out are produced less and the metabolites that it consumes are consumed less compared to controls. In order to preserve flux balance other enzymes need to consume more of the metabolites the down-regulated enzyme would have otherwise consumed, and less of the metabolites it would have otherwise produced, and so at least one of these enzymes would need to also be differentially expressed. The probability that the difference between cases and controls reflects a participation of the enzyme in differential fluxes is then given by $P(i)$, which is calculated according to [23]. In this work, reactions whose enzyme is over-expressed in cases are assumed to progress in the forward direction. This assumption simplifies the model, but can be easily removed by adding two mutually exclusive reactions in opposite directions instead of the original bidirectional reaction. One should note that we only model the changes in metabolic fluxes between controls and cases, not the absolute fluxes in

either of these groups.
In order for the solver to choose the enzymes/reactions that preserve flux balance and form the model with the highest probability, we need to ensure that if an enzyme/reaction is not chosen, its reaction rate will be 0. This can be seen in equation 1, where the vector $\vec{v}$ is constrained to 0 where $\vec{G}[i]$ is 0, i.e. when gene i is not selected for the solution. To achieve that, we add additional variables and constraints to the MILP formulation. We define a Boolean variable ρ_i for each reaction/gene(ρ_i is the MILP variable that represents $\vec{G}[i]$ in equation 1), and a continuous variable η_i for each reaction rate(η_i is the MILP variable that represents $\vec{v}[i]$ in equation 1.). If reaction i is chosen as contributing to fluxes that are differential between cases and controls, then its ρ_i variable is set to 1, and the logarithm of $1 - PEP_i$, i.e. the logarithm of the probability that the enzyme is differentially expressed between cases and controls, is added to the objective function(PEP_i is a constant in the MILP formulation that corresponds to $1 - P(j)$ in equation 1). If the reaction is not selected , its ρ_i variable will be set to 0, and the logarithm of its PEP will be added to the objective. The objective for all the genes $1..N$ is thus:

$$_{max} \sum_{i=1}^{N} (1 - \rho_i) \cdot log(PEP_i) + \rho_i \cdot log(1 - PEP_i).$$

and in addition to the flux balance constraints, we add the following two constraints for each reaction:

$$\eta_i \leq \rho_i$$

$$\eta_i \geq \frac{\rho_i}{10^5}$$

Consequently, if ρ_i is set to 0, η_i must also be 0. If ρ_i is set to 1, η_i can vary over five orders of magnitude. Additional constraints on the rates/fluxes, if available in the scientific literature, can further decrease the upper bound. It should be noted that our goal is to find which reactions and metabolites are active due to the conditions that are present in cases but not in controls, or vice versa. Qualitative properties of the inferred metabolic fluxes will then be used to derive insights into the regulation of metabolism under the conditions of the study, as will be demonstrated in the Results section. Since $\vec{v} + \vec{v'}$ preserves flux balance if both vectors do so separately, $\vec{v}$ can be viewed as a differential fluxes vector. If two enzymes can carry out the same reaction, each will have its own ρ and η variables. Maximizing the sum of log-probabilities will maximize the product of probabilities. In this work, we used the Gurobi solver to solve this MILP problem [24].

We refer to the reactions that are selected for the solution ($\rho = 1$) as differential reactions, and to the fluxes that they generate as differential fluxes. The next section describes an application of this method to high-throughout transcriptomic datasets. In order to evaluate the findings of the proposed analysis using Bayesian methods, we define additional Bayesian models.
To evaluate the probability of a chance overlap between two sets, we use a Bayesian

model of drawing balls from an urn without replacement, where white balls represent reactions/metabolites that overlap between the two sets and black balls reactions/metabolites present in only one of the sets/networks. We have the likelihood:
$HG(w|d, m + (n - m) \cdot \theta, n - (m + (n - m) \cdot \theta))$ and the prior:
$\theta \sim beta(1, 1)$
where HG is the hypergeometric PMF, n is the total number of balls in the urn, m is the number of white balls in the urn, d is the total number of balls drawn, w the number of white balls drawn, and $\theta \in [0, 1]$. The parameter θ can inflate the number of white balls in the urn in order to explain the observed number of white balls drawn. Intuitively, larger values of θ indicate that the drawn number of white balls could not have come from the urn if balls were drawn at random. To derive a probability of effect from the posterior, we set the interval $[0, \frac{m}{2 \cdot (n-m)}]$ as a region of practical equivalence (ROPE), i.e. the total density outside this range is the probability of effect. This choice of ROPE means that θ values that scale the number of white balls in the urn to more than one and a half of their actual value indicate non-random sampling. To test for under-representation m is scaled by θ instead of by having a term added to it, so the likelihood becomes:
$HG(w|d, m \cdot \theta, n - m \cdot \theta)$
and the ROPE is set to [0.5,1]. Note that since θ is a real number, rounding is necessary for MCMC [26]. We will refer to this model as the hypergeometric model.
For two sets of Boolean values A_1 and A_2, to compute the probability for an observed difference in the sums of the two sets, i.e. $P(|\sum A_1 - \sum A_2|)$, we model the probabilities of observing these sums using Poisson distributions, i.e. using the following likelihood:
$Pois(c_1|\lambda \cdot f) \cdot Pois(c_2|\lambda + \theta)$
where λ is a rate parameter with an improper flat prior,$c_1 = \sum A_1$ and $c_2 = \sum A_2$, f is a scaling factor computed as the ratio between the sizes of the groups ($f = \frac{|A_1|}{|A_2|}$), and θ is the parameter that models the rate difference between the groups and has the same prior as λ.
The ROPE in this case is [0,1], i.e. P(no effect)=$P(\theta \in [0, 1])$. We will refer to this model as the Poisson model.

3 Results

We obtained two independent transcriptomics datasets from ulcerative colitis patients and healthy controls from the Gene Expression Omnibus [27]: GSE177044 submitted in Wacker et al. [28], and GSE193677 submitted in Argmann et al. [29]. Since GSE177044 utilized whole-blood RNA-Sequencing and GSE193677 obtained samples using biopsy, we expect to see correlated though not identical metabolic changes [29]. The Human-GEM [30] model was used for the differential metabolic reactions between cases and controls, excluding enzymes that were not expressed, separately for each dataset. The product of available kcat values and scaled average RNA-Seq log-counts were used as additional upper bounds on fluxes.
When examining the metabolic reactions that constitute to the differential flux between cases and controls, we expected to see a high overlap between the two

datasets. Indeed, figure 1a shows that the size of the overlap is about half the total number of the differential reactions in GSE193677.
The probability of effect when testing the overlap between the two models was 1 (the posterior is illustrated in figure 2), i.e. the sampled posterior lies wholly outside the ROPE. In the rest of the manuscript when we use Bayesian inference to evaluate significance we will indicate the probability of no-effect (the posterior error probability) in parenthesis, e.g. for the overlap of reactions (p(no effect)=0.001, hypergeometric model).
We also permuted the gene labels for the differential expression results from both datasets, and for each permutation solved the differential flux MILP problem to obtain a set of reactions. The hypergeometric CDF computed using the overlap between differential fluxes in 1,000 simulated dataset pairs was always smaller than the one computed for the ulcerative colitis datasets. Next, for each of the two datasets we chose a set of genes such that the mean PEP in the differential expression analysis is 0.05, which sets the expected fraction of false positives to 0.05. Figure 1b shows the overlap between differentially expressed metabolic genes in the two datasets, compared to the total number of metabolic genes. As can be seen in the figure, the overlap is smaller than the overlap between reactions that were chosen for the model of differential fluxes. This is an indication that the metabolic model captures condition-specific properties more accurately.

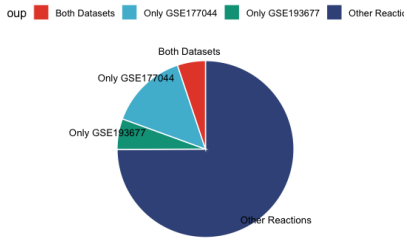
We next turned to derive insights into the differential metabolic reactions under ulcerative colitis by comparing the models generated for the two experimental datasets. The first question the models can address is whether there is a difference in the frequency of metabolite production in differential fluxes compared to the complete metabolic network. In other words, do the models indicate that some metabolites are significantly over- or under-utilized. We computed the fraction of reactions that produce each metabolite in each of the two ulcerative colitis differential flux models and compared it to their fractions in the complete metabolic network. Out of the 10 most under-utilized metabolites in each model/dataset, 4 were common to both models/-datasets(p=0, hypergeometric model), and out of the 10 most over-utilized reactions 5 overlapped (p=0, hypergeometric model). The proportions of reactions producing the four metabolites of the under-utilized compared to the same fractions in the full model are shown in figure 3, and the same information for the over-utilized metabolites is shown in figure 4. The total number of metabolites used by the models was 3,527 and 1,529 for GSE177044 and GSE193677, respectively. Lower-than expected utilization of these metabolites in the differential fluxes could be a result of their lower availability for the modeled reactions in patient samples. Similarly, over-utilization can correspond to over-abundance. We next examined the literature in order to interpret these findings.

PAPS (Phosphoadenosine-5'-phosphosulfate) is an intermediate in the production of hydrogen sulfide that is in turn used in biosynthesis. Hydrogen sulfide has been suggested as contributing to the pathology of ulcerative colitis by interrupting the maintenance of the colonic epithelial barrier [31]. Lower availability of PAPS for cellular metabolism could reflect a metabolic tuning in response to higher-levels of

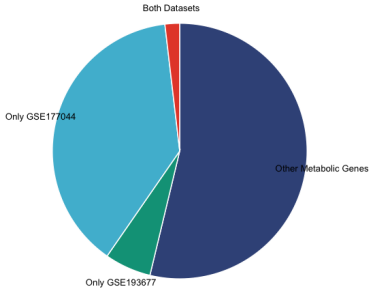
hydrogen sulfide. In randomly generated datasets, the same reduction in PAPS production occurs in only 0.01 of the cases, suggesting that most regulatory tuning will not cause such an extreme reduction. Since the mechanism of pathology involves hydrogen sulfide interrupting beta-oxidation, we tested whether inhibiting beta-oxidation will have an effect of PAP metabolism by down-regulating the ACADM gene [32] *in-silico* in both GSE177044 and GSE193677. Down-regulation can be applied to the model by removing the expression of the gene altogether from the dataset. In both down-regulation models, the fraction of reactions that generate PAPS significantly increased (p=0.001 in GSE177044, p=0.03 in GSE193677, Poisson model). Figure 5 shows the increase in both datasets. This may indicate that the low utilization of PAPS in the differential fluxes is coupled to the transcriptional control of beta-oxidation, and that the regulatory responses in the two models signal an increase in beta-oxidation.

Malonylation of the protein GAPDH (glyceraldehayde-3-phosphate dehydrogenase) leads to dissociation of GAPDH from TNF and therefore promotes inflammation [33]. Malonyl-CoA serves as a donor for this reaction, and therefore its decreased utilization by the metabolic network may reflect a metabolic reprogramming that makes it more available for the malonylation of TNF, a reaction that is not part of the metabolic model. We therefore sought a gene whose inhibition would increase utilization of malonyl-CoA by the metabolic network and whose protein product can be targeted by known compounds. Inhibiting GSTP1 results in an increase in differential metabolic reactions that process malonyl-CoA to a level observed in the complete metabolic network(p=0.001, Poisson model, for both datasets). It is therefore possible that such inhibition will reduce the metabolite's availability for pro-inflammatory activity. The alkaloid piperlongumine is an inhibitor of GSTP1 that has been observed to alleviate inflammation in other disease models [34, 35]. An interesting follow-up experiment would be to measure its effect on a disease model of ulcerative colitis. Currently, we have not developed a method for predicting adverse effects of metabolic changes, and therefore it is not possible to predict if inhibition of GSTP1 has detrimental effects in this context. We can test the metabolic impact of the inhibition by comparing metabolite utilization in GSE177044 before and after inhibition and other randomly generated datasets. Figure 6 shows that after inhibition the network's metabolite utilization is still closer to its state before inhibition than to most states of random datasets. However, the model does not include effects of metabolic changes on gene expression, which may follow more slowly than metabolic network tuning. Additionally, it does not include the reactions that directly regulate inflammation. Nevertheless, the ability to predict changes in uptake or utilization of metabolic compounds can serve as a first step in a series of experiments that test such hypotheses.

The hydroxyeicosatetraenoic acids that are used in the differential fluxes more than expected by chance have also been associated with inflammation and immune response [36, 37]. In order to test whether regulation of inflammation is controlled at the transcriptional level, we repeated the differential expression analysis for all the genes that are not included in Human-GEM, in both datasets, and used the David tool to find enriched GO terms [38]. The 5 top most significant metabolism-related GO terms enriched in both datasets and their false discovery rates in both datasets are shown in Table 1. As can be seen in this table, the enriched categories indicate that differential



(a) Overlap between metabolic reactions in the differential fluxes of the two datasets, GSE177044 and GSE193677. The overlap is about half the number of the all the reactions in GSE193677. The dark blue area denoted as 'Other Reactions' corresponds to reactions that had expressed genes that can carry them out in both datasets, but were not selected for either differential flux.



(b) Overlap between differentially expressed genes in the two datasets, GSE177044 and GSE193677. The dark blue area denoted as 'Other Metabolic Genes' corresponds to genes that are expressed n both datasets, but were not selected for the set of differentially expressed genes.

expression changes that don't directly regulate fluxes also play a role in regulation of metabolism.

Table 1: Top Enriched GO Terms

Term	GSE177044 FDR	GSE193677 FDR
positive regulation of metabolic process	8.57e-49	1.69e-06
positive regulation of macromolecule metabolic process	6.58e-49	4.08e-06
regulation of primary metabolic process	1.82e-85	4.44e-04
regulation of metabolic process	6.20e-31	5.74e-04
negative regulation of nucleobase-containing compound	1.42e-60	6.65e-04

4 Conclusion

This work presented a novel approach for genome-scale metabolic analysis of high-throughput gene expression data. Our method uses Bayesian inference to infer the differential metabolic fluxes between two conditions, obtaining a model that maximizes the probability of the differential fluxes and maintains flux-balance constraints. We showed its application to two datasets of ulcerative colitis patients and healthy controls, obtaining a high degree of agreement and deriving insights about regulation of metabolism under the condition. The method has the potential to improve our understanding of metabolic regulation in a broad range of conditions, since it only requires high-throughout transcriptomics data and a metabolic network model as inputs. Currently, the reactions included in the model are limited to the metabolic level, but the

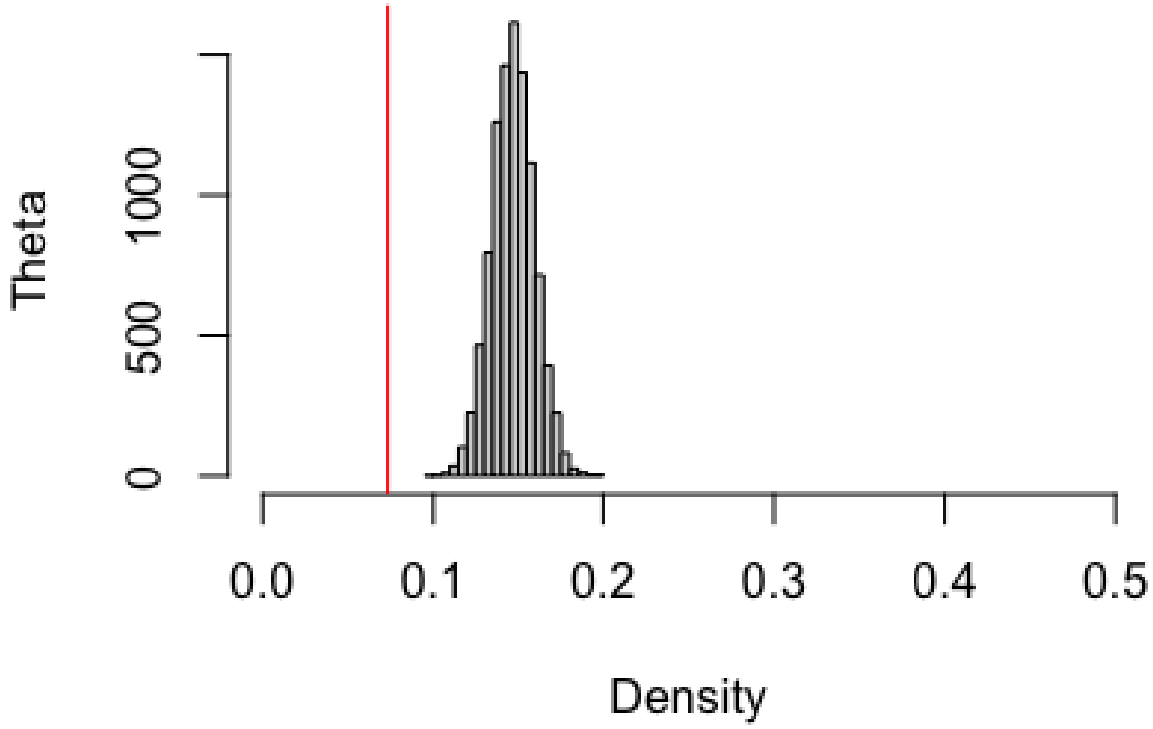


Fig. 2: The posterior of the parameter θ, that models the expected over-representation of reactions that are common to GSE177044 and GSE193677. The red vertical line indicates the upper end of the region of practical equivalence (ROPE), which would contain parameter values corresponding to no-effect (i.e random sampling). As can be seen in the figure, the posterior lies wholly outside the rope.

hypotheses generated using the model linking metabolism to other processes can be tested in the lab. Future work will extend this approach to include a model of the transcriptional network as well.

Declarations

4.1 Competing Interests

The authors declare that they have no competing of interests.

4.2 Data Availability

4.3 Ethics Declaration

Not applicable.

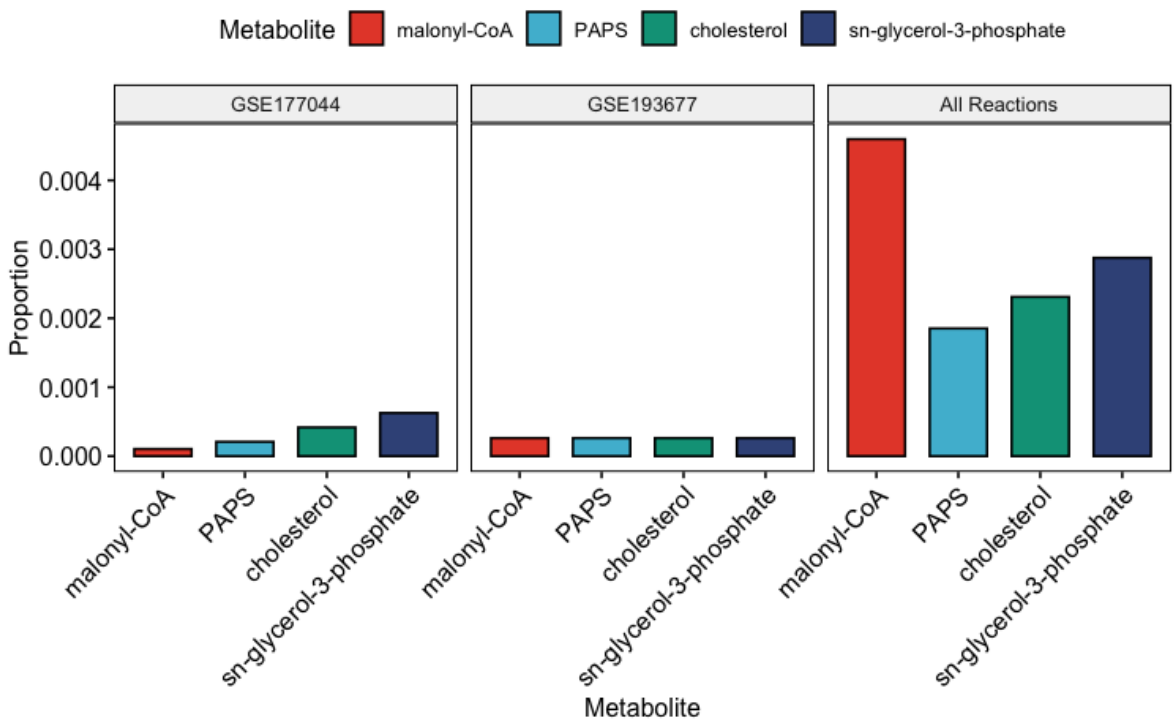


Fig. 3: The fraction of reactions that generate the metabolites malonyl-CoA, PAPS, CTP and cholesterol, in the models of differential fluxes for GSE177044 and GSE193677, and in the complete metabolic network(All Reactions). These four metabolites are among the 10 metabolites that are produced significantly less than expected in both the GSE177044 model and the GSE193677 model.

4.4 Consent to Publish Declaration

Not applicable.

4.5 Consent to Participate Declaration

Not applicable.

4.6 Author Contributions

G.K. performed all the work described in this manuscript.

4.7 Funding

There is no funding to declare. ==========

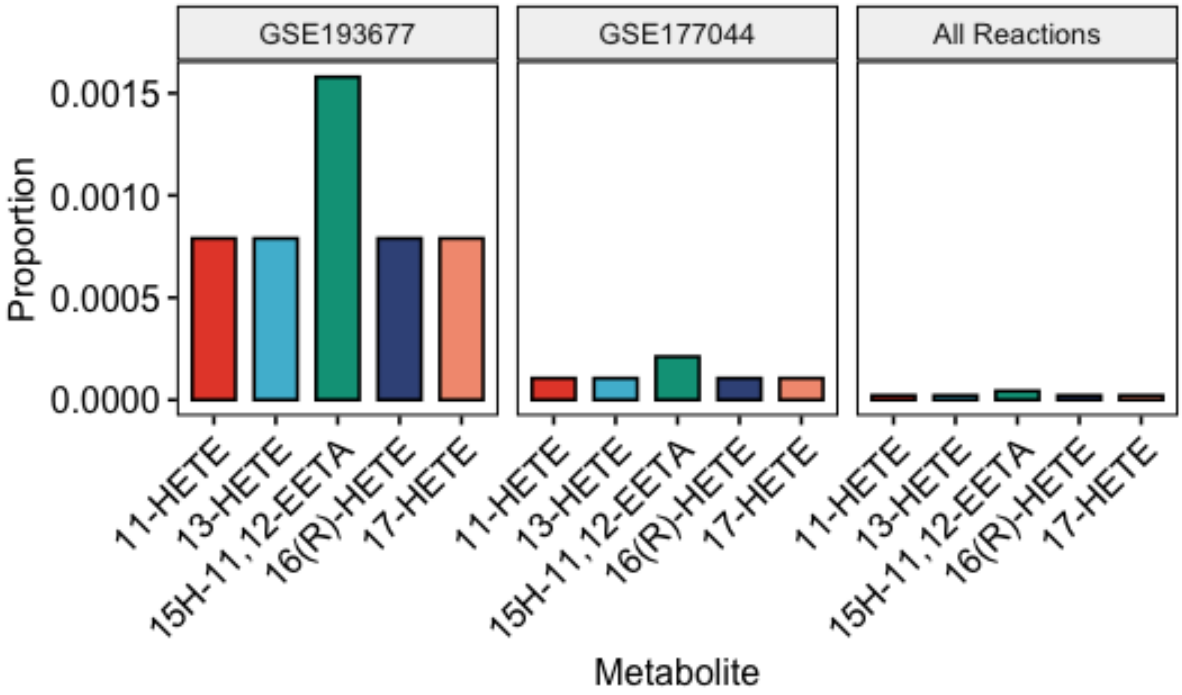


Fig. 4: The fraction of reactions that generate the metabolites 11-HETE, 13-HETE,15H-11,12-EETA, 16(R)-HETE and 17-HETE in the models of differential fluxes for GSE177044 and GSE193677, and in the complete metabolic network(All Reactions). These five metabolites are among the ten metabolites that are produced significantly more than expected in both the GSE177044 model and the GSE193677 model.

References

[1] Paul, S., Ghosh, S., Kumar, S.: Tumor glycolysis, an essential sweet tooth of tumor cells. Seminars in Cancer Biology **86**, 1216–1230 (2022) https://doi.org/10.1016/j.semcancer.2022.09.007

[2] Perera, R.M., Bardeesy, N.: When antioxidants are bad. Nature **475**(7354), 43–44 (2011) https://doi.org/10.1038/475043a

[3] Patra, K.C., Hay, N.: The pentose phosphate pathway and cancer. Trends in Biochemical Sciences **39**(8), 347–354 (2014) https://doi.org/10.1016/j.tibs.2014.06.005

[4] Hanahan, D., Weinberg, R.A.: Hallmarks of cancer: The next generation. Cell **144**(5), 646–674 (2011) https://doi.org/10.1016/j.cell.2011.02.013

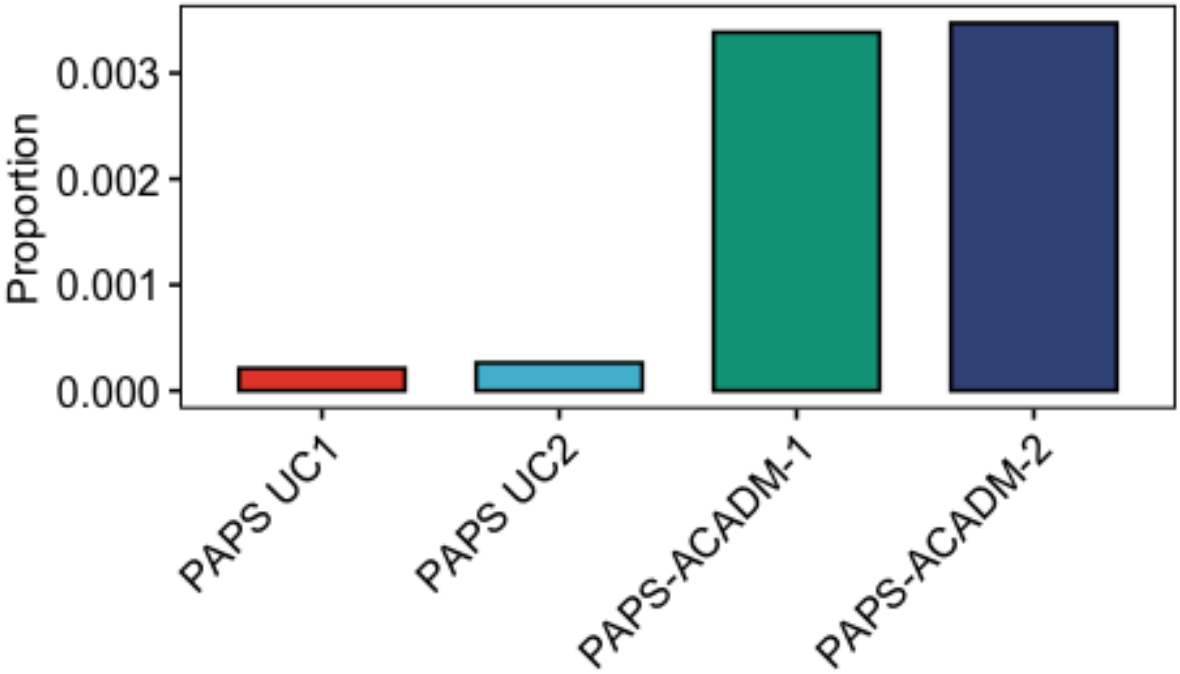


Fig. 5: Reactions that produce PAPS increase significantly in the model of differential fluxes when ADADM has been down-regulated. The bar marked PAPS ACADM-1 shows the proportion of PAPS-producing reactions after ACADM is down-regulated in GSE177044, and bar marked PAPS ACADM-2 shows the same value for the ACADM down-regulation in GSE193677. The bars marked PAPS UC-1 and PAPS-UC2 show the same proportions before the ACADM down-regulation in GSE177044 and GSE193677, respectively.

[5] Park, J.H., Pyun, W.Y., Park, H.W.: Cancer metabolism: Phenotype, signaling and therapeutic targets. Cells **9**(10), 2308 (2020) https://doi.org/10.3390/cells9102308

[6] Koebbe, L.L., Hess, T., Giel, A.-S., Bigge, J., Gehlen, J., Schueller, V., Geppert, M., Dumoulin, F.L., Heller, J., Schepke, M., Plassmann, D., Vieth, M., Venerito, M., Schumacher, J., Maj, C.: The genetic regulation of the gastric transcriptome is associated with metabolic and obesity-related traits and diseases. Physiological Genomics **56**(5), 384–396 (2024) https://doi.org/10.1152/physiolgenomics.00120.2023

[7] Morris, J.P., Yashinskie, J.J., Koche, R., Chandwani, R., Tian, S., Chen, C.-C., Baslan, T., Marinkovic, Z.S., Sánchez-Rivera, F.J., Leach, S.D., Carmona-Fontaine, C., Thompson, C.B., Finley, L.W.S., Lowe, S.W.: alpha-ketoglutarate

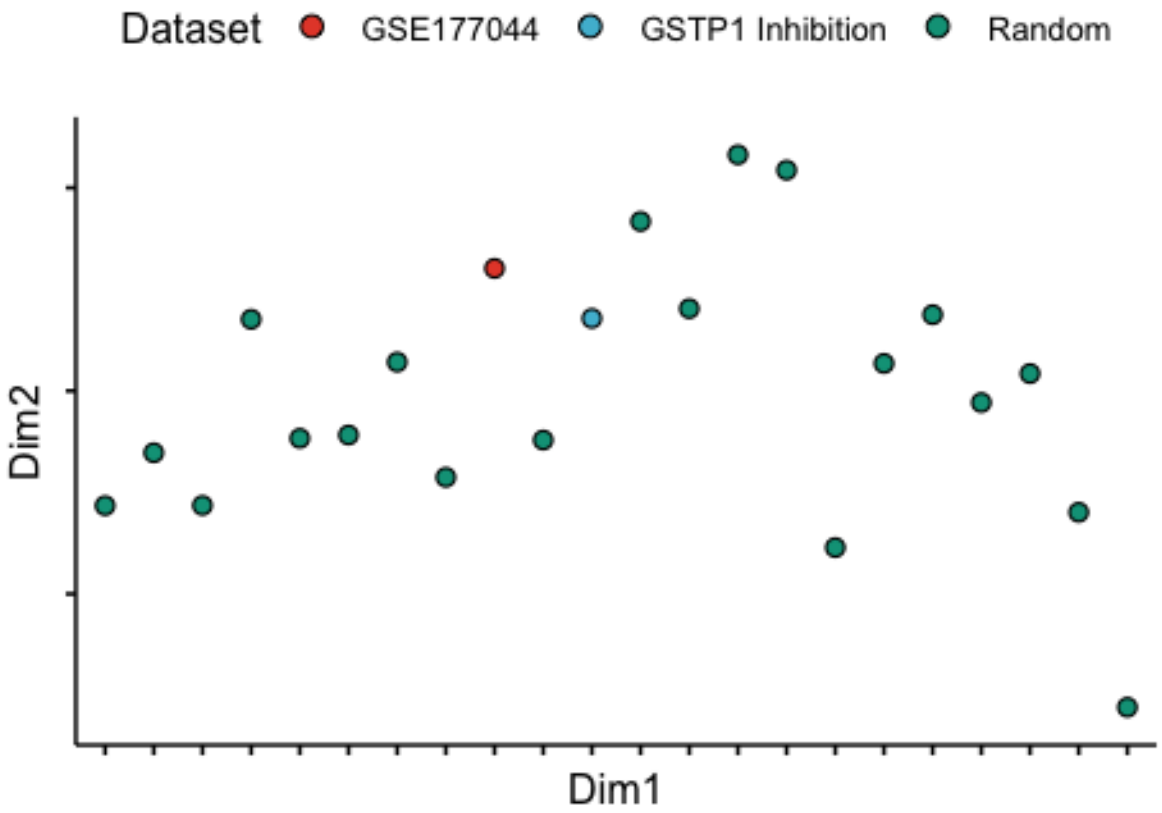


Fig. 6: Similarity in metabolite utilization by the differential flux models constructed for GSE177044, GSE177044 after inhibition of GSTP1 and randomly generated datasets. The fractions of reactions involving each metabolite were mapped to two dimensions using multidimensional scaling. GSE177044 and GSE177044 after GSTP1 inhibition utilize metabolites more similarly than randomly generating datasets, suggesting that inhibition of GSTP1 may have localized effect on the metabolic network.

links p53 to cell fate during tumour suppression. Nature **573**(7775), 595–599 (2019) https://doi.org/10.1038/s41586-019-1577-5

[8] Lee, K.S., Su, X., Huan, T.: Metabolites are not genes — avoiding the misuse of pathway analysis in metabolomics. Nature Metabolism **7**(5), 858–861 (2025) https://doi.org/10.1038/s42255-025-01283-0

[9] Stark, R., Grzelak, M., Hadfield, J.: Rna sequencing: the teenage years. Nature Reviews Genetics **20**(11), 631–656 (2019) https://doi.org/10.1038/s41576-019-0150-2

[10] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.: Gene ontology: tool for the unification of biology. Nature Genetics **25**(1), 25–29 (2000) https://doi.org/10.1038/75556

[11] Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P.: Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences **102**(43), 15545–15550 (2005) https://doi.org/10.1073/pnas.0506580102

[12] Tarca, A.L., Draghici, S., Khatri, P., Hassan, S.S., Mittal, P., Kim, J.-s., Kim, C.J., Kusanovic, J.P., Romero, R.: A novel signaling pathway impact analysis. Bioinformatics **25**(1), 75–82 (2008) https://doi.org/10.1093/bioinformatics/btn577

[13] Isci, S., Ozturk, C., Jones, J., Otu, H.H.: Pathway analysis of high-throughput biological data within a bayesian network framework. Bioinformatics **27**(12), 1667–1674 (2011) https://doi.org/10.1093/bioinformatics/btr269

[14] Martini, P., Sales, G., Massa, M.S., Chiogna, M., Romualdi, C.: Along signal paths: an empirical gene set approach exploiting pathway topology. Nucleic Acids Research **41**(1), 19–19 (2012) https://doi.org/10.1093/nar/gks866

[15] Orth, J.D., Thiele, I., Palsson, B.: What is flux balance analysis? Nature Biotechnology **28**(3), 245–248 (2010) https://doi.org/10.1038/nbt.1614

[16] Ryu, J.Y., Kim, H.U., Lee, S.Y.: Reconstruction of genome-scale human metabolic models using omics data. Integrative Biology **7**(8), 859–868 (2015) https://doi.org/10.1039/c5ib00002e

[17] Imam, S., Schauble, S., Brooks, A.N., Baliga, N.S., Price, N.D.: Data-driven integration of genome-scale regulatory and metabolic network models. Frontiers in Microbiology **6** (2015) https://doi.org/10.3389/fmicb.2015.00409

[18] Blazier, A.S., Papin, J.A.: Integration of expression data in genome-scale metabolic network reconstructions. Frontiers in Physiology **3** (2012) https://doi.org/10.3389/fphys.2012.00299

[19] Shlomi, T., Cabili, M.N., Herrgård, M.J., Palsson, B., Ruppin, E.: Network-based prediction of human tissue-specific metabolism. Nature Biotechnology **26**(9), 1003–1010 (2008) https://doi.org/10.1038/nbt.1487

[20] Tian, M., Reed, J.L.: Integrating proteomic or transcriptomic data into metabolic models using linear bound flux balance analysis. Bioinformatics **34**(22), 3882–3888 (2018) https://doi.org/10.1093/bioinformatics/bty445

[21] Berlo, R.J.P., Ridder, D., Daran, J.-M., Daran-Lapujade, P.A.S., Teusink, B., Reinders, M.J.T.: Predicting metabolic fluxes using gene expression differences as constraints. IEEE/ACM Transactions on Computational Biology and Bioinformatics **8**(1), 206–216 (2011) https://doi.org/10.1109/tcbb.2009.55

[22] Kimmerling, R.J., Prakadan, S.M., Gupta, A.J., Calistri, N.L., Stevens, M.M., Olcum, S., Cermak, N., Drake, R.S., Pelton, K., De Smet, F., Ligon, K.L., Shalek, A.K., Manalis, S.R.: Linking single-cell measurements of mass, growth rate, and gene expression. Genome Biology **19**(1) (2018) https://doi.org/10.1186/s13059-018-1576-0

[23] Karlebach, G., Aronow, B., Baylin, S.B., Butler, D., Foox, J., Levy, S., Meydan, C., Mozsary, C., Saravia-Butler, A.M., Taylor, D.M., Wurtele, E., Mason, C.E., Beheshti, A., Robinson, P.N.: Betacoronavirus-specific alternate splicing. Genomics **114**(2), 110270 (2022) https://doi.org/10.1016/j.ygeno.2022.110270

[24] Gurobi Optimization, LLC: Gurobi Optimizer Reference Manual (2023). https://www.gurobi.com

[25] Karlebach, G., Hansen, P., Veiga, D.F., Steinhaus, R., Danis, D., Li, S., Anczukow, O., Robinson, P.N.: Hba-deals: accurate and simultaneous identification of differential expression and splicing using hierarchical bayesian analysis. Genome Biology **21**(1) (2020) https://doi.org/10.1186/s13059-020-02072-6

[26] Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A.: Stan: A probabilistic programming language. Journal of Statistical Software **76**(1) (2017) https://doi.org/10.18637/jss.v076.i01

[27] Edgar, R.: Gene expression omnibus: Ncbi gene expression and hybridization array data repository. Nucleic Acids Research **30**(1), 207–210 (2002) https://doi.org/10.1093/nar/30.1.207

[28] Wacker, E.M., Uellendahl-Werth, F., Bej, S., Wolkenhauer, O., Vesterhus, M., Lieb, W., Franke, A., Karlsen, T.H., Folseraas, T., Ellinghaus, D.: Whole blood rna sequencing identifies transcriptional differences between primary sclerosing cholangitis and ulcerative colitis. JHEP Reports **6**(2), 100988 (2024) https://doi.org/10.1016/j.jhepr.2023.100988

[29] Argmann, C., Hou, R., Ungaro, R.C., Irizar, H., Al-Taie, Z., Huang, R., Kosoy, R., Venkat, S., Song, W.-M., Di'Narzo, A.F., Losic, B., Hao, K., Peters, L., Comella, P.H., Wei, G., Atreja, A., Mahajan, M., Iuga, A., Desai, P.T., Branigan, P., Stojmirovic, A., Perrigoue, J., Brodmerkel, C., Curran, M., Friedman, J.R., Hart, A., Lamouse-Smith, E., Wehkamp, J., Mehandru, S., Schadt, E.E., Sands, B.E., Dubinsky, M.C., Colombel, J.-F., Kasarskis, A., Suárez-Fariñas, M.: Biopsy and blood-based molecular biomarker of inflammation in ibd. Gut **72**(7), 1271–1287 (2022) https://doi.org/10.1136/gutjnl-2021-326451

[30] Luo, J., Wang, H., Moyer, D., Guo, Z., Robinson, J.L., Gustafsson, J., Anton, M., Chen, Y., Kerkhoven, E.J., Nielsen, J., Li, F.: Reconstruction of human metabolic models with large language models. Proceedings of the National Academy of Sciences **123**(15) (2026) https://doi.org/10.1073/pnas.2516511123

[31] Roediger, W.E.W., Moore, J., Babidge, W.: Colonic sulfide in pathogenesis and treatment of ulcerative colitis. Digestive Diseases and Sciences **42**(8), 1571–1579 (1997) https://doi.org/10.1023/a:1018851723920

[32] Thorpe, C., Kim, J.P.: Structure and mechanism of action of the acyl-coa dehydrogenases ¡sup¿1¡/sup¿. The FASEB Journal **9**(9), 718–725 (1995) https://doi.org/10.1096/fasebj.9.9.7601336

[33] Galván-Peña, S., Carroll, R.G., Newman, C., Hinchy, E.C., Palsson-McDermott, E., Robinson, E.K., Covarrubias, S., Nadin, A., James, A.M., Haneklaus, M., Carpenter, S., Kelly, V.P., Murphy, M.P., Modis, L.K., O'Neill, L.A.: Malonylation of gapdh is an inflammatory signal in macrophages. Nature Communications **10**(1) (2019) https://doi.org/10.1038/s41467-018-08187-6

[34] Thatikonda, S., Pooladanda, V., Sigalapalli, D.K., Godugu, C.: Piperlongumine regulates epigenetic modulation and alleviates psoriasis-like skin inflammation via inhibition of hyperproliferation and inflammation. Cell Death and Disease **11**(1) (2020) https://doi.org/10.1038/s41419-019-2212-y

[35] Harshbarger, W., Gondi, S., Ficarro, S.B., Hunter, J., Udayakumar, D., Gurbani, D., Singer, W.D., Liu, Y., Li, L., Marto, J.A., Westover, K.D.: Structural and biochemical analyses reveal the mechanism of glutathione s-transferase pi 1 inhibition by the anti-cancer compound piperlongumine. Journal of Biological Chemistry **292**(1), 112–120 (2017) https://doi.org/10.1074/jbc.m116.750299

[36] Wollard, P.M., Cunnigham, R.M., Murphy, G.M., Camp, R.D.R., Derm, F.F., Greaves, M.W.: A comparison of the proinflammatory effects of 12(r)- and 12 (s)-hydroxy-5, 8, 10, 14-eicosatetraenoic acid in human skin. Prostaglandins **38**(4), 465–471 (1989) https://doi.org/10.1016/0090-6980(89)90129-9

[37] Nishida, T., Miwa, H., Shigematsu, A., Yamamoto, M., Iida, M., Fujishima, M.: Increased arachidonic acid composition of phospholipids in colonic mucosa from patients with active ulcerative colitis. Gut **28**(8), 1002–1007 (1987) https://doi.org/10.1136/gut.28.8.1002

[38] Sherman, B.T., Hao, M., Qiu, J., Jiao, X., Baseler, M.W., Lane, H.C., Imamichi, T., Chang, W.: David: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). Nucleic Acids Research **50**(W1), 216–221 (2022) https://doi.org/10.1093/nar/gkac194