PathQuant tool to compute shortest reactional distances (SRD).

The method was developed using R and delivered as an R package. PathQuant (for Pathway Quantity).

1.1 Input parameters

Gene-metabolite pairs: PathQuant accepts the following input for SRD computation: 1) gene(s); and 2) their associated metabolite(s) in columns, each with their specific KEGG identifiers (IDs). Each row represents a unique gene-metabolite association. Only associations between genes and individual metabolites are taken as entry, hence associations between a gene and a ratio of metabolites must be separated prior to analysis (usually by generating gene-numerator/denominator pairs).

Selected metabolic pathways: PathQuant accepts a list of metabolic pathways, each with its specific KEGG IDs. For this application, we used the map ID: hsa: 01100 in KEGG (referred herein as the 'overview' or 'hsa01100' throughout this article).

1.2 Association classification

PathQuant can classify each association of the input by gene product into four broad categories: enzyme, transporter, other (other proteins, transcription factor, and more) and not classified using KEGG Brite database.

1.3 Metabolic network modelling

The selected KEGG pathway, encoded in KEGG XML file format (KGML), is downloaded using the KEGG API, from the most up-to-date KEGG pathways available, and then moved to a specific version folder in order to keep track, or use a specific version of interest. Users can choose which downloaded version to use or let PathQuant use the up-to-date version. The pathway is then converted into a graph of biochemical reactions (also called compound graph) with metabolites, as nodes and genes, mapped to their corresponding encoded enzymes, as edges. The topology of the pathway is captured in the constructed graph. Genes encoding enzymes catalysing multiple reactions are mapped to multiple edges. The constructed graph

represents exactly the metabolic pathways of KEGG which are built mainly with metabolites that are the main reactants of a reaction, dismissing cofactor metabolites, such as NAD, or common co-substrates/products, such as ATP or H2O. Finally, we use a non-oriented graph as the KEGG standards are not consistent in this matter.

1.4 SRD computation

Our method computes the SRD, which is defined as the shortest reactional distance path between a given gene and a metabolite. The SRD is computed using each metabolic pathway received in input, in which a given pair is mapped. The SRD is computed as described in the workflow figure (Figure1): A distance of 0 is assigned to metabolites, which are the main substrates or products of the reaction catalysed by the enzyme encoded by the selected gene of interest. The SRDs to all other metabolites are obtained using the breadth-first search algorithm. The algorithm is used to find the SRD starting from the substrate and from the product of the mapped gene to the paired metabolite, thereby selecting the smallest SRD between these two. Figure1 depicts an example of SRD computation for a hypothetical reaction The main reactants of this reaction: the substrate and the product are set at an SRD of 0. Everything running deeper than the substrate and the product within the graph is adding a distance of one for each depth: SRD = 1 for the green hypothetical metabolites and SRD = 2 for the blue hypothetical metabolite.

1.5 SRD metric analysis

The utility of the computed SRD metric for the annotation of gene-metabolite associations reported by mGWAS was assessed using different approaches (See STAR Methods).

1.6 Data outputs and visualisation

PathQuant outputs a text file containing gene and metabolite classification, Enzyme Commission number (EC), KEGG Brite, KEGG IDs of used pathways for the SRD computation, and SRD values for all associations. These SRD values can also be visualised in a heatmap and global or multiple distribution plots; a few examples are available at in this manuscript (Results).