Recon 3D: A Resource Enabling A Three-Dimensional View of Gene Variation in Human Metabolism

Supplementary Information

Elizabeth Brunk[a,b], Swagatika Sahoo[c,d], Daniel C. Zielinski[a], Ali Altunkaya[e], Andreas Dräger[f], Nathan Mih[a], Francesco Gatto[a,g], Avlant Nilsson[g], German Andres Preciat Gonzalez[c], Maike Kathrin Aurich[c], Andreas Prlić[e], Anand Sastry[a], Anna D. Danielsdottir[c], Almut Heinken[c], Alberto Noronha[c], , Peter W. Rose[e], Stephen K. Burley[e,h], Ronan M.T. Fleming[c], Jens Nielsen[b,g], Ines Thiele*[c], Bernhard O. Palsson*[a,b]

* correspondence should be addressed to: I.T. (ines.thiele@gmail.com) and B.O.P (palsson@eng.ucsd.edu)

[a] Department of Bioengineering, University of California San Diego CA 92093

[b] The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, 2800 Lyngby, Denmark

[c] Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Campus Belval, Esch-Sur-Alzette, Luxembourg

[d] Current address: Department of Chemical Engineering, Indian Institute of Technology Madras, India 600036

[e] RCSB Protein Data Bank, San Diego Supercomputer Center, University of California, San Diego, La Jolla, CA 92093, USA

[f] Applied Bioinformatics Group, Center for Bioinformatics Tübingen (ZBIT), University of Tübingen, 72076 Tübingen, Germany

[g] Department of Biology and Biological Engineering, Chalmers University of Technology, Sweden

[h] Department of Chemistry and Chemical Biology, Center for Integrative Proteomics Research, Institute for Quantitative Biomedicine, and Rutgers Cancer Institute of New Jersey, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA

Supplementary Notes

Supplementary Note 1: Content and statistics of Recon 3D

Addition of reactions

The overall approach to the assembly of Recon 3D and the different resources considered are shown in Supplementary Figure 1. Overall, 6163 new reactions, 1589 new metabolites, and 1654 new genes were added to the reconstruction (Supplementary Data File 1&2). The major proportion of the 6163 newly added reactions belong to transport (32%) and lipid metabolism (24%) reactions, followed by exchange (19%), xenobiotic (11%), and amino acid (7%) metabolism (Supplementary Figure 2B-C). The additions for transport reactions include, novel membrane transporters, e.g., ATP-binding cassette family of transporters (ABCs), aquaporins, and organic anion transporters (OATPs), as well as intracellular transport proteins, e.g., ATP synthase, H+ transporting mitochondrial proteins (*ATP5S*, GeneID: 27109, *ATP5O*, GeneID: 539, *ATP5L2*, GeneID: 267020), mitochondrial pyruvate carrier 1 (*MPC1*, GeneID: 51660), and mitochondrial pyruvate carrier 2 (*MPC2*, GeneID: 25874). In the following, we will discuss details on newly added reactions for lipid and amino acid subsystems.

New reactions from lipid metabolism

The major added/expanded lipid metabolic pathways include, (i) phospholipid metabolism (427 reactions), (ii) cholesterol metabolism (403 reactions), (ii) fatty acid oxidation (301 reactions), (iii) fatty acid synthesis (214 reactions), eicosanoid metabolism (134 reactions), and (iv) triglyceride metabolism (four reactions).

For the phospholipid metabolism, reactions were added to capture sphingomyelins, monoacyl-glyceride, lysophopshatidyl-choline, lysophosphatidyl-inositol, lysophosphatidyl-ethanolamine, glycerophosphate, and endocannabinoids. These phospholipid species have been detected in the blood metabolomics data (http://www.metabolomicscentre.nl/). Sphingomyelines play an important role in myelin formation in neurons[4]. Sphingomyelins (SM) are synthesized from ceramide. This reaction is catalyzed by sphingomyelin synthase (E.C. 2.7.8.27), where in phosphatidyl choline transfers its phospho-choline to ceramide forming SM, and releasing a molecule of diacyl-glycerol[4]. Most of the new reactions captured ceramides that have a fatty acyl chain length from 14 to 25 carbon atoms in the 'R-group'. While saturated fatty acyl molecules could be easily added, only few of the unsaturated fatty acyl molecules could be added, for which the exact position of the unsaturation has been specified. Therefore, a provision of defined information of lipid structure is necessary to include the unsaturated fatty acyl molecules. The monoacyl-glycerides are synthesized from the diacylglycerides via triacylglycerol lipase (E.C. 3.1.1.3) in a hydrolysis reaction releasing fatty acyl CoAs as byproduct. This reaction mainly occurs at the plasma membrane or extracellular space in liver, adipocytes, skeletal muscle, and testis[5]. Monoacyl-glycerides were added with fatty acyl chain length of 16 to 18 carbon atoms. Lysophopshatidyl-choline, lysophosphatidyl-inositol, and lysophosphatidyl-ethanolamine are synthesized by the action of phospholipases (E.C. 3.1.1.4) from membrane-derived phospholipids. Additionally, they can be produced from acyltransferases (E.C. 2.3.1.43), wherein the fatty acyl-CoAs are transferred from phospholipids, mostly phosphatidylcholine to cholesterol forming cholesterol-ester and releasing lysophosphatidyl-choline (in the case that the reactant is phosphatidylcholine) as a byproduct[6]. The added lyso-phospholipids contained fatty acyl-CoAs of chain length ranging from 14 to 26 carbon atoms. The glycerol-phosphate metabolism was extended to include a novel intermediary metabolite, i.e., glycerol-2-phosphate, which is particularly active in brain[7,8]. Furthermore, the mitochondrial glycerol-3-phosphate-dehydrogenase reaction was added to Recon 3, since this reaction is active in muscle and has been implicated to be associated with type-2 diabetes[9]. Endocannabinoids are family of small molecules present in brain and act as cannabinoid receptor (G-protein-coupled receptor superfamily) agonist[10]. Their synthesis is mediated via phospholipase D (E.C. 3.1.4.4), involving the hydrolysis of phosphatidyl-ethanolamine and the release of endocannabinoids[11]. Phospatidic acid is released as a byproduct. Endocannabinoids of acyl-CoAs chain length from 12 to 20 carbon atoms were added.

For the cholesterol metabolism, cholesterol-esters were included. While the cholesterol-ester synthesis occurs via acyltransferase (described above, see E.C. 2.3.1.43), which is mostly extracellular. The degradation of cholesterol-ester occurs in the lysosome, via the sterol esterase (E.C. 3.1.1.13), releasing free cholesterol and the respective fatty acids. Cholesterol ester degradation involves esterification with fatty acids with chain length between 18 to 22 carbon atoms. In order to enable mapping of biomarkers of organic aciduria and dicarboxylic aciduria (OMIM: 222730), a novel fatty acid oxidation pathway, i.e., omega-one oxidation, was added. This pathway involves hydroxylation of the fatty acid at the omega one position[12]. Prominent products of the pathway include 7-hydroxy-octanoate and 5-hydroxy-hexanoate.

New reactions from amino acid metabolism

In the case of amino acids, the reaction addition include, (i) peptide metabolism (242 reactions), (ii) protein metabolism (43 reactions), (iii) phenylalanine metabolism (40 reactions), (iv) tyrosine metabolism (30 reactions), and metabolism of other important amino acids, .e.g., lysine, methionine, arginine, branched-chain amino acids, etc. (87 reactions.)The peptides were taken from the study published by Chen et al that described genomic, transcriptomic, proteomic, metabolomic, and autoantibody profiles from a single patient[13]. All the di-and tri-peptides were assumed originate from the diet[14]. These peptides were included with their hydrolysis into individual amino acids as well as their re-formation from the respective amino acids. For phenylalanine metabolism, reactions capturing human-gut microbial co-metabolism were added. This included metabolism of phenylacetate (phenylalanine derivative). Phenylacetate ligation with coenzyme A generates phenylacetyl-CoA, which further undergoes conjugation with glycine via glycine-N-acyltransferase (E.C. 2.3.1.13.), producing the acylglycine, i.e., phenylacetylglycine[15]. The ligation reaction is ATP dependent. Phenylacetylglycine has been detected in urine and serves a biomarker for phospholipidosis[16]. In the case of tyrosine metabolism, reactions were added to capture dopamine formation from phenylethylamine[17], and other important diagnostic biomarkers used in the newborn screening panel in Luxembourg[18], e.g., N-acetyl-tyrosine, succinylacetone, vanilpyruvic acid, and N-acetylvanilalanine. These tyrosine intermediates are detected in the urine of the newborn, when aromatic L-aminoacid decarboxylase deficiency (OMIM: 608643) or hereditary tyrosinemia (OMIM: 276700) is suspected[19,20]. While acetylation of tyrosine produces N-acetyl-tyrosine, succinylacetone is produced via an alternate route from fumarylacetoacetate, which includes reduction followed by decarboxylation. Vanilpyruvic acid is formed from vanilalanine via alanine transaminase (E.C. 2.6.1.2), and N-acetylvanilalanine is formed via acetylation of 3-methoxy-tyrosine (see Supplementary Data File 1 for details). In the case of lysine metabolism, reactions were added to capture novel human metabolic functions, e.g., acetylation of lysine, hydrolysis of acetyl-lysine (acyl-lysine deacylase, E.C. 3.5.1.17), and homocitrulline (lysine carbamoyltransferase, E.C. 2.1.3.8)[21]. Additionally, formation of 3-hydroxy-glutarate from glutaryl-CoA and glutaconate from glutaconyl-CoA were included, as these serve as important biomarkers for the diagnosis of organic acidurias[22] . For methionine metabolism, an alternative route for methionine transamination, which is active in skeletal muscle, was added (see below for details). In order to capture biomarkers of maple syrup urine disease (OMIM: 248600) and beta-ketothiolase deficiency (OMIM: 203750), i.e., 2-hydroxy-3-methyl-valerate and 2-methyl-3-hydroxy-butyrate[22], respectively, the isoleucine degradation pathway was extended to account the synthesis of these biomarker metabolites. This included adding oxidation of 3-methyl-2-oxopentanoate to 2-hydroxy-3-methyl-valerate via lactate dehydrogenase (E.C. 1.1.1.27) and hydrolysis of hydroxy-methyl-butanoyl CoA to 2-methyl-3-hydroxy-butyrate (Supplementary Data File 1). As discussed above, acylglycines are important biomarkers for metabolic disorders. Therefore, reactions were added to capture, conjugation of glycine with hexanoyl CoA and suberyl CoA, which form hexanoyl glycine and suberyl glycine, respectively. These metabolites serve as biomarkers for hereditary medium-chain acyl-CoA dehydrogenase deficiency (OMIM: 201450)[23]. Histidine metabolism was expanded to include histidine derivatives, i.e., homocarnosine, via carnosine synthase (E.C. 6.3.2.11) and methyl-histidine[21]. Similarly, for the aforementioned amino acids, reactions were added to capture the respective diagnostic relevant biomarker. Supplementary Data File 1 should be referred to for further details.

Addition of metabolites

Similar to the metabolic sub-system for added reaction, major fraction for new metabolites belonged to (i) lipids (42%), and (ii) amino acids (19%) class, followed by miscellaneous and xenobiotics (18% each), carbohydrates (2%), vitamins (1.4%), and nucleotide (0.3%) (Supplementary Figure 2C). The new metabolites belonging to the lipid class included (i) bile acid metabolite (e.g., ursodeoxycholyl CoA), (ii) lyso-glycerophosphocholine (e.g., 1-myristoyl-glycerophosphocholine), (iii) lysophosphatdyl-ethanolamine (e.g., 1-oleoylglycerophospho-ethanolamine), (iv) lysophosphatidyl-inostiol (e.g., 1- palmitoylglycerophosphoinositol), (v) phosphatidylglycerol (e.g., 1-oleoylglycerol), (vi) fatty acids (e.g., caproic acid), (vii) dicarboxylic acid (e.g., 3-hydroxy-adipate), (viii) fatty acid derivative (e.g., ethylmalonic acid), (ix) cholesterol ester (e.g., 1-gamma-Linolenoyl-cholesterol), (x) sphingomyelin (e.g., SM (d18:1/14:0)), (xi) endocannabinoids (e.g., docosahexaenoyl ethanolamide), (xii) hormone derivative (e.g., 19-hydroxyandrost-4-ene-3,17-dione), (xiii) prostaglandins (e.g., 15-keto-prostaglandin F2a), (xiv) thromboxane (e.g., thromboxane B2). The majority of the metabolites categorized as amino acids were peptides, which represent 79% of the newly added amino acid metabolites, and include dipeptides (e.g., arginyl-arginine) and tripeptides (e.g., cystyl-asparaginyl-methionine). Other important added amino acid include amino acid derivatives that serve as biomarkers for inherited metabolic disorders, e.g., vanilpyruvic acid[22] or metabolites mediating host-microbe interaction, e.g., 3-hydroxy-glutarate (described above). Further important additions include vitamin derivatives (e.g., menaquinones), dietary sugar alcohol (e.g., arabitol), and phenolic compounds (e.g., cresol and cresol-sulphate).

Exo-metabolomic data measures metabolites in blood, urine, plasma, cerebrospinal fluid (CSF), or other biofluid[24], and represent metabolic signatures of human metabolism. A major proportion of the reaction additions that forms the expanded Recon 3 (Supplementary Figure 3) comes from the mapping of multi-metabolomics data sets (i.e., 1852/5175, 36 %). These data sets have been taken from human blood metabolome[25], KORA study[26–28], blood metabolomics data set from Hankemier, T. et al (http://www.metabolomicscentre.nl/), IEM data[29,30], human breast milk metabolomics[31], the integrative personal omics profile[13], the human metabolome database (HMDB)[24], and exo-metabolome measurements of cancer cell lines[32]. Taking all the data sets together, 542 unique metabolites mapped onto Recon 2, while 391 unique metabolites were newly added. However, 1814 metabolites could be not be mapped, hence, could not be included in Recon 3. This occurred, since, the exact biochemical pathway information is scarce/absent, or the metabolomic data did not provide adequate information on the specific chemistry of the chemical entities, e.g., specific position of double bonds in case of lipid molecules (Supplementary Data File 4). We also considered 34 reactions (Supplementary Data File 1) suggested in[21], and chose those that were provided as acceptable solutions and experimentally validated for occurrence in human cells.

From the 542 mapped metabolomics derived metabolites, 348 were also listed in HMDB as being detected in human biofluids. While, 232/348, i.e., 67% of HMDB metabolites already existed in the extracellular space of Recon. The remaining 116/348, i.e., 33% of metabolites were part of intracellular metabolism of Recon. Therefore, 149 transport reactions were added to enable these metabolites to be connected to the extracellular space of Recon. E.g., coproporphyrin I and III were connected to the extracellular space by addition of three transport reactions each via, i.e., (i) a passive diffusion process, (ii) active transport mediated by ATP-binding cassettes, and (iii) solute carrier organic anion transporters (Supplementary Data File 1).

Addition of genes and modification of GPR rules

The newly added genes were mostly those, whose encoded proteins participate in (i) miscellaneous processes (77%), (ii) lipid metabolism (10%), (iii) carbohydrate metabolism (5%), (iv) transport processes (5%), (v) amino acid (3%), followed by nucleotide metabolism and vitamin metabolism (1%), Supplementary Figure 2D. The genes grouped under 'miscellaneous' were majorly obtained from HMR 2.0[24], i.e., 99% of the miscellaneous category genes. These genes comprised of those encoding receptor proteins, calcium/calmodulin protein kinases, cathepsin, polymerase family proteins, cell cycle proteins, ubiquitin system, etc. The majority of the added genes encoding for enzymes of the lipid metabolism belonged to phospholipid metabolism , e.g., sphingomyelin synthase (GeneID: 166929, *SGMS2*),

phospholipase A2 (GeneID; 123745, *PLA2G4E*), and glycerophosphocholine phosphodiesterase (GeneID: 56261, *GPCPD1*). Additionally, existing reactions were modified for the GPR rules as per the most recent literature evidence (Supplementary Data File 7). Of all the 2201 reactions with modified for the GPRs, 63% were those incorporated into Recon 2 originally from Hepatonet 1[33], followed by Recon 1 reactions (31%)[34], and Edinburgh human metabolic network (2%)[35]. In line with the modified GPRs, new genes were also added. The newly added genes encoding enzymes of the amino acid metabolism, belonged to protein assembly, e.g., albumin (GeneID: 213, *ALB*), plasminogen (GeneID: 5340, *PLG*), and coagulation factor II (GeneID: 2147, *F2*). The other important additions include indoleamine 2, 3-dioxygenase (GeneID: 169355, *IDO2*) that encodes enzyme participating in tryptophan metabolism, and gamma-glutamyltransferases (GeneID: 92086, 124975, 102724197, *GGTLC1*, *GGT6*, *LOC102724197*). Protein products of these genes participate in glutathione metabolism (Supplementary Data File 1). The published Recon 2[36] contained certain genes that could not be mapped to unique GeneIDs within EntrezGene[37]. These ambiguous genes (111 genes) were removed from Recon 3 (Supplementary Table 1B). Taken together, the genetic component of Recon 3 underwent addition of novel genes, modification of GPR associations, as well as removal of ambiguous genes.

Formulation of metabolic tasks

During and after the construction and debugging process (Fig. S1), Recon 3 was validated against 431 metabolic objective functions (Supplementary Data File 3). For metabolic functions that the model should be active for. Of the various functions tested, 100 metabolic objectives were newly formulated (Supplementary Data File 3) and include mostly organ-specific metabolic tasks, which should also be fulfilled by the global metabolic model. Of the 100 metabolic tasks, 81 metabolic tasks are unique for a specific organ: (i) skeletal muscle (36 metabolic tasks), (ii) kidney (22 metabolic tasks), (iii) brain (11 metabolic tasks), (iv) retina (seven metabolic tasks), (v) lung (four metabolic tasks), and (vi) heart (one metabolic tasks). The remaining 19 metabolic tasks are shared between organs: (i) heart, skeletal muscle, and kidney (five metabolic tasks), (ii) brain and kidney (four metabolic tasks), (iii) heart and skeletal muscle (three metabolic tasks), (iv) brain and heart (three metabolic tasks), (v) brain, skeletal muscle, and heart (two metabolic tasks), (vi) brain, heart, and kidney (one metabolic tasks), and (vii) kidney and skeletal muscle (one metabolic tasks).

Supplementary Note 2: Building a curated and high quality Recon 3

Mapping multi-metabolomic data sets

The mapping of the multi-metabolomic data sets was done manually, checking one metabolite at a time (Supplementary Data File 4). The metabolite description was used as a search criterion to map it onto the Recon metabolite names. Additionally, metabolite databases, e.g., HMDB[24], ChEBI[38], PubChem[39], and ChemSpider (http://www.chemspider.com/) were used to extract the corresponding metabolite identifier and mapped onto the human metabolic reconstruction using the metabolite identifier. Furthermore, the chemical structures of the mapped metabolites were checked to identify the true matches. Such a mapping process is time consuming and involves extensive manual efforts. Therefore, we suggest that metabolomics data sets should be provided with extensive database dependent as well as database independent metabolite identifiers (e.g., HMDB ID, ChEBI ID, Inchi strings) that could assist in automating and accelerating the mapping procedure. Furthermore, this ensures reusability of the metabolomics data sets by a broader community.

A point worth mentioning is that metabolites, which could not be mapped onto Recon 3, and arose either when the metabolite was (i) unknown (i.e., the precise pathway for synthesis and degradation is currently unknown), or (ii) of xenobiotic origin, hence, excluded from inclusion into a global network that aims to model metabolism of endogenous metabolites, or (iii) precise chemical information on the metabolite was lacking. The latter scenario is particularly true for lipid molecules, where in the exact position of the double bond is not specified or the exact lipid composition of a phospholipid or triglyceride or sphingomyelin species is missing. To enable accurate mapping of metabolomics data sets

onto a metabolic network, it is essential to be provided with all the required details including the pathway information, i.e., the exact lipid composition and position of double bonds within the lipid molecule.

Issues with mapping lipidomic data onto Recon

Lipidomic data currently do not measure, to the best of our knowledge,: (i) the exact position of the double bond in case of unsaturated fatty acids (e.g., FA (24:2), FA (22:2), FA (19:1)), (ii) the exact lipid composition of a glycerophospholipid (e.g., PC (O-44:5), PC (40:5), PE (36:4)) or triglyceride (e.g., TG (52:6), TG (52:3), TG (62:3)) or sphingomyelin (e.g., SM (d18:1/18:2), SM (d18:1/ 23:1), SM (d18:0/24:2)), and (iii) the precise pathway for synthesis and degradation of the lipid molecule (e.g., 12-hydroxy-5Z,8Z,10E,14Z-eicosatetraenoic acid, 8S-hydroxy-5Z,9E,11Z,14Z-eicosatetraenoic acid, 12-oxo-5Z,8Z,10E,14Z-eicosatetraenoic acid). However, to enable the accurate mapping of lipidomic data sets onto a metabolic network, such information needs to be available.

When the above information is not available, this leads to ambiguity in adding the precise metabolite to the metabolic network, and if added, the charge and precise elemental formula of such metabolites is noted as 'X' or with generic formula, e.g., 'R' for unknown fatty acid composition. This creates mass and charge unbalanced reactions within the network, resulting in inaccurate predictions by the final model. Additionally, enzymes have substrate specificity, with regards to the position of the unsaturation[13,36,40]. Due to lack of information the precise enzyme and gene-protein-information cannot be added to the metabolic network.

Addition of organ-specific reactions

Recon 3 is global human metabolic network, hence, should be able to capture metabolic transformations across all major cell types and organs in the human body. A total of 155 new reactions and 11 new metabolites were added to capture organ-specific metabolism of organs, i.e., colon (42 reactions, 6 metabolites), skeletal muscle (six reactions, 2 metabolites), brain (one reaction), kidney (one reaction), and for both liver and adipocyte (one reaction) (Supplementary Data File 1). These new reactions include an alternative route of methionine degradation reported to be occurring in skeletal muscle and liver[41], transport of metabolites from colonic cells to the lumen facilitating gut microbioata utilization, and synthesis of nicotinate ribonucleotide from quinolinate in kidney[42]. Furthermore, reactions from previously published metabolic models for red blood cell[43] and adipocyte[44] that could not be mapped onto Recon 2 reaction content, were added, i.e., red blood cells (45 reactions, 3 metabolites) and adipocytes (22 reactions). From the adipocyte model, 539 reactions mapped onto Recon 3 and for RBC model, 318 reactions mapped onto Recon 3 (Supplementary Data File 5).

Mapping of the reaction content onto Recon 3 was done manually using the reaction abbreviation, reaction description, and reaction formula. In the case of adipocytes, the blood compartment was replaced with extracellular compartment to find the correct matches with Recon reactions. Additionally, the published adipocyte model[44] contained lumped version of the fatty acid oxidation reactions, hence, the corresponding un-lumped versions were included into Recon 3, following an in-depth manual curation of the newly added reactions and un-lumped versions. Supplementary Data File 5 gives details on mapping of reactions from published reconstructions onto Recon 3.

Maintenance of certain metabolite pools and metabolite storage as reserve for energy demands within the cells has been reported to be crucial for maintaining the organ specific functions. Typically these are glycogen storage in liver and skeletal muscle[4], or fatty acid storage in the adipocytes[45]. During periods of fasting, liver glycogen serves to maintain the blood glucose levels. Additionally, triglyceride stores in the adipocytes are broken down to supply fatty acids to skeletal muscle and heart to be utilized as energy resource[46]. A thorough manual search was performed to obtain the list for storage capacity of dietary nutrients. This was represented by formulating specific demand reactions (Supplementary Data File 1). The nutrient storage capacity of organs, such as liver, adipocyte, brain, retina, muscle, adipocytes, and kidney was reconstructed in the form of 25 reactions (Supplementary Data File 1). Typical examples include vitamin

storage by liver[47], free fatty acids storage in adipocyte[48], requirement of docosahexanoic acid in brain and retina[49], and uptake and requirement of choline by brain, heart, kidney, muscle, and liver[50,51] (Supplementary Data File 6).

The bile salts aid in the physiological digestion and absorption processes. Bile is formed in the liver and drained into the gall bladder, from where it is released into the duodenum (first part of the small intestine) for efficient digestion and absorption of food. Hence, the storage functions for 26 bile salts for gall bladder were captured by adding the respective demand reactions (Supplementary Data File 1).

Adding reactions from transport module

The previously published review on membrane transport proteins and reactions[40] was used as a compendium of transporters. This review identified the missing transport reactions that should be part of Recon, as per the most current knowledge on transport proteins. Additionally, it reported the substrate specificity of the missing transport proteins as well as transport mechanism and directionality of the transport reactions in the form of 51 reactions (Supplementary Data File 1). The remaining 42 reactions pre-existed in Recon 2 and were updated only with their gene-protein-reaction association (Supplementary Tables 7-8). These were incorporated into Recon 3.

Adding reactions to capture host-microbe interactions

The human gut harbors more number of microorganisms than its own cells. These microbes have been implicated in a number of human diseases, such as vascular disease, inflammatory bowel disease, and cancer[52]. These reactions were identified in a previously published study that involved studying interactions between the human metabolic network, Recon 2[36], and a microbial community of eleven microbes[53]. A set of 24 such reactions were formulated following manual curation of the relevant scientific literature (Supplementary Data File 1) and were incorporated into Recon 3. These reactions involved metabolism of diagnostically relevant microbial metabolites, e.g., para-cresol, indoxyl-sulfate, phenylacetylglycine, and para-cresol sulfate that have been associated with a number of metabolic disorders ranging from renal disorders[54], phospholipidosis[55], endothelial dysfunction, and cardiovascular disease[56].

Additions of reactions to enable modeling of diet formulations

We formulated three dietary patterns in previous studies, i.e., an average American diet, balanced diet, and vegetarian diets[57,58]. All these dietary metabolites should be, in principle, part of the exchange medium of the global reconstruction. However, lipoamide and dextrin were present in Recon 2 only as intracellular metabolite. Therefore, four reactions were added to enable transport and exchange of these metabolites between extracellular and intracellular compartments (Supplementary Data File 1). Additionally, five new fatty acids were added along with their metabolic and transport pathways in the form of 20 reactions (Supplementary Data File 1). These fatty acids include 7, 10 hexadecadienoic acid, 7, 10, 13-hexadecatrienoic acid, 2-heptadecanoic acid, 2, 11, 14-eicosatrienoic acid, and eicosapentenoic acid. The later fatty acid belongs to the omega-3 fatty acids, and has been shown to impart cardio protective effects[59].

Addition of reactions to capture lipoprotein metabolism

The transport of lipid metabolites, e.g., cholesterol and triglyceride within the blood circulation is enabled within lipoprotein particles. These are carrier particles that have a nonpolar lipid core, mainly triacylglycerol and cholesterol-ester[4]. The outer layer is composed of free cholesterol, phospholipids, and apo-proteins[4]. While the core lipid and peripheral apo-protein composition changes depending on the type of lipoproteins, the integral apo-protein is ApoB[4]. The chylomicron and very density lipoprotein (VLDL) majorly carry triacylglycerol, while the intermediate low density lipoprotein (IDL), low density lipoprotein (LDL), and high density lipoprotein (HDL) majorly carry cholesterol-ester. Chylomicron is synthesized by the small intestine followed by its release into the blood circulation, whereby, the capillary lipoprotein lipase (E.C. 3.1.1.3) hydrolyzes the triacylglycerol to release free fatty acids and glycerol to be utilized by adipocytes, skeletal muscle, and other extra-hepatic tissues. The hydrolyzed chylomicron, i.e., remnant chylomicron enters the liver for further processing. VLDL is synthesized by liver, which when enters the systemic circulation is hydrolyzed via lipoprotein lipase and converted to IDL, and finally to LDL. Tissues expressing LDL

receptors, e.g., liver uptakes LDL for further modification. The HDL is synthesized by liver and small intestine in the nascent discoidal form, and upon acquisition of cholesterol-esters from blood and other peripheral tissues become fully matured HDL and taken up by liver for utilization. Capillary enzyme lecithin cholesterol esterase (E.C. 2.3.1.43) aids in transfer of free cholesterol to HDL by converting it into cholesterol-ester to be utilized by the liver. Compared to HDL, LDL contains more cholesterol-ester and is considered toxic to the vascular endothelium, owing to their oxidation leading to atherosclerosis. The HDL renders a cardio-protective effect, since it transfers cholesterol from the peripheral tissues to the liver. Additionally, these lipoprotein ratios, i.e., total cholesterol/HDL and LDL/HDL hold potential as an index for cardiovascular disease[60]. Therefore, 44 reactions and five new metabolites were included in Recon 3 to capture the lipoprotein metabolism (Supplementary Data File 1).

Addition of reactions from HMR

A total of 2478 reactions were added to Recon 3 from HMR 2.0[24]. Therefore, we manually mapped the metabolite and reaction onto the human metabolic reconstruction to identify overlaps. For HMR reactions that contained metabolites already present in the human metabolic reconstruction, the Recon 3 metabolite abbreviations were used. Otherwise the HMR nomenclature was preserved. Not all HMR metabolites have information on their composition. Furthermore, the HMR metabolites are only provided as uncharged compounds. Where possible we mass- and charge balanced the reactions prior to their inclusion to Recon 3. We did not further curate the GPRs of HMR reactions, i.e., only 'or' rules are provided. To ensure that protein complexes are correctly captured, further manual curation and literature review will be necessary in a future effort. However, obsolete gene entries were removed from the GPRs. Note that we also included flux inconsistent reaction from HMR, as we assumed that those reactions were included into HMR only after careful evaluation of existing evidence in the biochemical literature by the authors.

Addition of drug module

The recently published drug module [57], was considered for addition to Recon 3D to expand the xenobiotic metabolism. This module captures the absorption, distribution, metabolism, and excretion of five highly prescribed drug groups. These include eight drugs from statin group, three from anti-hypertensives, two each from immunosuppressants and analgesic groups. Combining the reactions, from the specific drug metabolism, i.e., lovastatin (C10AA02), simvastatin (C10AA01), atorvastatin (C10AA05), pravastatin (C10AA03), cerivastatin, (C10AA06), pitavastatin (C10AA08), rosuvastatin (C10AA07) , losartan (C09CA01), torasemide (C03CA04, C03CA01), nifedipine (C08CA05), cyclosporine A (L04AD01, S01XA18), tacrolimus (D11AX14, L04AA05), ibuprofen (C01EB16), acetaminophen (N02BE01), gliclazide (A10BB09), midazolam (N05CD08), and allopurinol (M04AA01), a total of 721 reactions were added.

Addition of bile acid module

Some of the secondary bile acids produced by the gut microbiota, e.g., lithocholic acid, are cytotoxic for human and need to be detoxified. Host mechanisms to facilitate the elimination of cytotoxic bile acids into feces include sulfation[61] and glucuronidation[62]. Glucuronidation of bile acids is carried out by UDP glucuronosyltransferase (UGT), namely UGT1A3 in the liver[63] , UGT2B7 in the small intestine and the liver[64], UGT2A1 in the small intestine[65], and UGT2A3 in the colon[65]. In total, 22 glucuronidation reactions for primary and secondary bile acids were created. Sulfation at the 3-hydroxy positions of bile acids is carried out by sulfotransferase SULT2A1[66]. In total, 12 sulfation reactions for primary and secondary bile acids were created. Moreover, CYP3A4 performs hydroxylation and oxidation of several bile acids[67,68] and seven reactions were included accordingly. In the hepatocyte, the transport of bile acids is carried out by the Na^+-taurocholate co-transporting polypeptide (NTCP; SLC10A1) and the organic anion transporting polypeptide (OATP) family (OATP1B1 and OATP1B3) on the sinusoidal membrane and the bile salt export pump (BSEP; ABCB11) on the canalicular membrane[66]. The major transporters in the ileocyte are the dependent bile acid transporter (ASBT; SLC10A2) on the apical side and the heteromeric organic solute transporter alpha-beta (OSTα-OSTβ ; SLC51A, SLC51B) on the basolateral membrane[66]. OSTα-OSTβ are also present in the colonocyte[66]. Moreover, the ATP-dependent transporters MRP3 transports bile acids on the basolateral side of the ileocyte and colonocyte. On the apical side, this function is

performed by MRP2 or ABCG2[66]. Uptake reactions for primary and secondary bile acids and secretion reactions for glucuronidated and sulfated bile acids were formulated accordingly. In total, the bile acid module added 39 unique metabolites, 41 metabolic reactions, 79 exchange and demand reactions, and 96 transport reactions to Recon3.

Debugging of blocked reactions for the Recon 3

To ensure that all added reactions could carry flux (with the exception of the HMR reactions, see previous paragraph), we performed flux variability analysis (FVA[69]) while having all exchange reactions unconstrained. For certain metabolites, transport and exchange reactions were added to enable a non-zero flux through the newly added reactions, leading to the addition of 251/5175 (5 %) reactions (Supplementary Data File 1). Typical examples include the debugging for the bile acid metabolites and their metabolic reactions. In total, we added 60 new bile acid reactions to capture the metabolism and transport of bile acids, e.g., 7-Ketolithocholate, ursodeoxycholate, lithocholate, chenodeoxycholate. The newly added phosphatidycholine, phosphatidylethanolamine, and phosphatidylinositol metabolites were dead-end metabolites (i.e., metabolites that are either only consumed or produced within the network). This occurred because these metabolites were generated only in the extracellular space by the lecithin:cholesterol acyltransferase (E.C. 2.3.1.43)[70] and hence, were not connected to the intracellular metabolism. Consequently, 48 new reactions were added to transport these metabolites to the cytosol, involving their generation from the membrane phospholipids by phospholipase A2 (E.C. 3.1.1.4)[71]. Therefore, the debugging process not only added transport reactions, but also, novel metabolic routes. The described additions were done by an extensive manual curation of the scientific literature (Supplementary Data File 1).

Refinements performed for Recon 3

Not only the global human metabolic network, Recon 3, expanded but also refined. The model refinement was done in two stages: (i) removal of duplicate metabolites and associated reactions suggested by Quek et al, where in Recon 2 was reduced for metabolic flux analysis[72], (ii) modification in gene-protein-reaction associations, (iii) adjustment of metabolite formulae to pH 7.2, and (iv) further mass and charge balancing of reactions based on updated metabolite formulae.

Removal of duplicate metabolites

Of the 95 metabolites that were mentioned with their corresponding duplicate entries by Quek et al[72], we manually checked the metabolite identifiers and chemical structures, and 71 of these were considered for replacement (Supplementary Data File 10). The remaining 3/95 metabolites were wrongly annotated, and 21/95 metabolites were maintained as distinct liver and uterine homologs, consistent with the original reconstructions[34,36]. Additionally, we identified and removed further duplicate metabolites in Recon 3D, with the aim to eliminate as many as possible of such instances. We would like to highlight that also metabolic reconstructions should have standardized information, i.e., database dependent and independent metabolite identifiers (e.g., HMDB IDs[24], ChEBI IDs[38], INCHI[73] and SMILES[74] descriptors) to facilitate and accelerate the process of cross-mapping between reconstructions. This would also reduce the occurrence of doubled metabolite and reaction entries, which remains a manual and very time consuming process. The need for such standards has been previously formulated[75].

Modification of gene-protein-reaction associations

The previously published hepatocyte-specific metabolic network, Hepotanet1[33], defined only ‘or’ GPR associations GPRs, which were consequently included as such in Recon 2. In Recon 3, we updated the GPRs of these original Hepotanet1 reactions. We also updated the GPRs of the existing Recon 1[34] and EHMN[76] reactions as per the updated information in EntrezGene[37] and Uniprot[77] databases. Typical cases include (1) reaction ‘2AMACSULT’ involving sulfate group transfer existed with no GPR and was assigned with sulfotransferase (GeneID: 6818, *SULT1A3*). (2) reaction ‘3NTD7l’ involving hydrolysis of nucleosides existed with one gene encoding phosphatases (GeneID: 53, *ACP2*), and was additionally assigned with newly identified phosphatases (*ACP2*, GeneID: 54, *ACP5*, GeneID: 51205, *ACP6*, GeneID: 55, *ACPP*). (3) reaction ‘r0309’ involving fatty acid elongation was incorrectly assigned to gene encoding amine oxidase and was now

correctly assigned with enoyl-CoA reductases (GeneID: 55825, *PECR*, GeneID: 51102, *MECR*). All the GPR updates and refinements were done manually, after careful evaluation of gene annotation data from EntrezGene[37], OMIM[78], and Uniprot[77] databases. Additionally, the associated scientific literature was extensively used to assign the correct 'and' or 'or' GPRs. The full list of these updates for 2180 reactions is provided in Supplementary Data FIles 7-8. Note that the update does not only include the new gene ID but also Boolean rules. Further for the details of all the gene content of Recon 3, see Supplementary Data File 8. Additionally, we simplified the Boolean rules of the GPRs by removing unnecessary parentheses.

Adjustment of metabolite formulae to pH 7.2

The metabolites in genome-scale metabolic reconstructions are generally adjusted to an internal pH of 7.2[79]. Metabolite formulae for metabolites from HMR 2.0 were adjusted to this pH value, additionally, we also checked and corrected the metabolite formulae of the remaining metabolites. Briefly, we used the MOL files and the getMetabolitepKa.m function in the COBRA toolbox[80] as well as ChemAxon (via https://chemicalize.com/) to calculate the pKa values for each metabolite, and select the most predominant microspecies at ph 7.2. The molecular formulae and the charge were assigned accordingly.

Mass and charge balancing of reactions

To ensure the mass- and charge balancing of the reactions in Recon 3D, we identified all mass- and charge unbalanced reactions, using the updated metabolite formulae. For reactions that were imbalanced in protons, protons were either added or removed from the reactions to achieve balancing. All other instances were inspected manually and corrected to the best of our knowledge based on current literature. Nonetheless, mass- and charge imbalanced reactions remain in Recon 3D, which are either i) due to the presence of metabolites that contain an R group, thus, that represent an unspecified metabolite, ii) metabolites without any further information provided by the original sources (i.e., HMR 2.0 [81], EHMN [76], and HepatoNet [33]), or iii) the correct reaction mechanism is unknown (e.g., nature of cofactor).

Biomass objective functions

Recon 3 contains three different versions of the biomass function reaction. These are (i) biomass_reaction, (ii) biomass_maintenance, and (iii) biomass_maintenance_noTrTr. The biomass_reaction is the general biomass reaction as in Recon 3, biomass_ maintenance is same as biomass_reaction except for the nuclear deoxynucleotides, and biomass_maintenance_noTrTr is devoid of amino acids, nuclear deoxynucleotides, and cellular deoxynucleotides except for adenosine-triphosphate. The biomass_reaction can be used only for tissues known to possess re-generative capacity, i.e., liver[82], heart[83], and kidney[84]. For others, biomass_ maintenance was added, indicating the maintenance of cellular metabolic profiles, i.e., the model's capability to synthesize all the biomass components excepting the nuclear deoxynucleotides. The biomass_maintenance_noTrTr reaction can be used to model specific conditions, e.g., fasting condition. Such a modification was done as the human body has no store for amino acids[46]. Amino acids if stored intracellularly, increase the osmotic pressure, necessitating their rapid catabolism[46]. Such catabolic processes mainly occur for those that are not required for protein synthesis.

Formulating metabolic objectives for Recon 3

During the literature search for the organ-specific metabolic pathways, each organ was noted with its chief metabolic functions, e.g., Cori's cycle between liver and skeletal muscle, arginine synthesis in kidney, citrulline synthesis by small intestine, cholesterol synthesis by spleen, and vitamin D synthesis by skin. Glucose from liver enters skeletal muscle, where it is converted to lactate via anaerobic glycolysis. The muscle then releases lactate back into the circulation to be utilized for gluconeogenesis by the liver, contributing to the muscle-liver-Cori's cycle[4]. Kidney is the major organ for synthesis of arginine from citrulline[85]. Citrulline synthesized in the small intestine reaches kidney for further metabolism by urea cycle reactions, thereby, contributing to inter-organ amino acid metabolism. Spleen is one of the important haematopoietic organ, and synthesis of dolichol and cholesterol from acetate are important indicators of this process[86].

The human skin is mainly responsible for synthesis of vitamin D from 7-dehydrocholesterol in multiple reaction step[87]. These physiological functions and their representative biochemical reactions were set as metabolic objectives for each organ. While most of the objectives were same as the ones, for which Recon 2 was tested for[36], new ones were formulated for skeletal muscle (50 metabolic objectives), kidney (34 metabolic objectives), brain (21 metabolic objectives), heart (18 metabolic objectives), retina (7 metabolic objectives), and lung (4 metabolic objectives). The total number of objectives are shown in Supplementary Data File 3.

These metabolic objectives not only captured the overall metabolism and physiological functions of the organ, but also, reactions/objectives contributing to the inter-organ metabolism.

Formulation of a flux balance model

Leak testing

The reconstruction was tested for secretion or production of metabolites from nothing. A chief reason for leakage in metabolic networks is occurrence of mass imbalanced reactions. Therefore, all exchange and sink reactions are constrained to zero for the lower bound. Then, all exchange reactions were optimized individually for a non-zero flux. Additionally, a demand reaction for each compartment-specific metabolite in the model was created and maximized. The leak test in the following refers to the test of non-zero fluxes through all these reactions.

Additionally, we tested that the model does not i) produce energy from water, ii) produce energy from water and oxygen, iii) produce matter when atp demand (DM_atp_c_), and v) carry flux through an h[m] demand (max, min) and through an h[c] demand (max, min). Furthermore, the model passed all tests defined by Agren et al[88]. Finally, we confirmed that the atp yield from different carbon sources was consistent with reported literature values (Supplementary Table 1C). Note that the predicted atp yield was higher than the calculated for the longer chain fatty acids. This is due to transport of protons across intracellular membranes, against the proton gradient present in cells. We limited this effect by introducing an inner mitochondrial compartment as suggested by Swainston et al. [89]. Briefly, the following reactions were reformulated:

Recon 2		**Recon 3D**	
Reaction abbreviation	**Reaction formula**	**Reaction abbreviation**	**Reaction formula**
PIt2m	h[c] + pi[c] -> h[m] + pi[m]	PIt2mi	h[i] + pi[i] -> h[m] + pi[m]
ATPS4m	4 h[c] + adp[m] + pi[m] -> h2o[m] + 3 h[m] + atp[m]	ATPS4mi	adp[m] + pi[m] + 4 h[i] -> atp[m] + h2o[m] + 3 h[m]
CYOR_u10m	2 h[m] + 2 ficytC[m] + q10h2[m] -> 4 h[c] + q10[m] + 2 focytC[m]	CYOR_u10mi	2 ficytC[m] + 2 h[m] + q10h2[m] -> 2 focytC[m] + q10[m] + 4 h[i]
Htm	h[c] -> h[m]	Htmi	h[i] -> h[m]
NADH2_u10m	5 h[m] + nadh[m] + q10[m] -> 4 h[c] + nad[m] + q10h2[m]	NADH2_u10mi	5 h[m] + nadh[m] + q10[m] -> nad[m] + q10h2[m] + 4 h[i]
CYOOm3	o2[m] + 7.92 h[m] + 4 focytC[m] -> 1.96 h2o[m] + 4 h[c] + 4 ficytC[m] + 0.02 o2s[m]	CYOOm3i	4 focytC[m] + 7.92 h[m] + o2[m] -> 4 ficytC[m] + 1.96 h2o[m] + 4 h[i] + 0.02 o2s[m]
CYOOm2	4.0 focytC[m] + 8.0 h[m] + o2[m] -> 4.0 ficytC[m] + 4.0 h[c] + 2.0 h2o[m]	CYOOm2i	4.0 focytC[m] + 8.0 h[m] + o2[m] -> 4.0 ficytC[m] + 4.0 h[i] + 2.0 h2o[m]

Alternative approaches have been proposed in the literature (e.g., [81]), however, we opted against the introduction of artificial “metabolites” representing the proton gradients across all intracellular membranes and the extracellular membrane because it leads to mass imbalanced reactions. We believe that the correct way to deal with this issue inherent to mass balance models is the use of thermodynamic constraints. Firstly, to adjustment of the metabolites to

the correct pH present in the different organelles, as described by Haraldsdottir et al. [90] and secondly to add constraints representing energy conservation and the second law of thermodynamics. However, it is an open problem to develop robust and efficient modelling techniques for application of the latter constraints, as well as biochemically derived reaction bounds, to genome-scale models in general, which beyond the scope of the current work.

In the case that one or more leaking reactions were identified, the flux vector associated with the maximization of the leaking reaction was inspected manually for reversible reactions that contributed maximally to the defined objective. We then reviewed the literature and thermodynamic data to identify those reactions that have mostly likest a net irreversible flux. In cases of mass- and charge imbalanced reactions, we also investigated the option of closing upper and lower bounds, thus, avoiding any flux through these reactions. We ensured that closing those reactions did not impeded flux through the biomass reactions and through the metabolic functions (see below). Generally, we updated the reversibility of one reaction at the time and repeated the leak test. Repeating this procedure in a greedily manner allowed us to identify those reactions that need to be unidirectional or closed to obtain a leak free model of Recon 3, while having the best possible literature and thermodynamic support. However, we acknowledge that our choices may not be agreed upon by every user, hence, we provide Recon 3 with the original set of constraints (which yields a leaking model) as well as the leak free model. The differences in reaction bounds are listed in Supplementary Data File 11. The user may use different constraints than provided in the model, as a model is by nature condition-specific. In any case, the user should always ensure that his/her condition-specific model is leak free, e.g., using the appropriate function in The COBRA Toolbox[80].

Stoichiometric and flux consistency

A set of metabolic reactions is termed stoichiometrically consistent, when it is possible to assign a positive molecular mass to each metabolite, without violating mass balance for any reaction. Given chemical formulae for each metabolite, we checked for elemental balance of each reaction, but there is no guarantee that a set of supposedly elementally balanced reactions is stoichiometrically consistent because it is possible to mistakenly specify the chemical formula for a metabolite. In addition, although it might not be possible to determine if a reaction is elementally balanced, e.g., due to missing formulae, a reaction may still be stoichiometrically consistent with a complementary set of mass balanced reactions. Therefore, to identify the largest set of stoichiometrically consistent internal reactions in Recon 3D, we employed a novel algorithm (Fleming et. al., *submitted*) that minimises the number of reactions that violate a mass conservation constraint, subject to a strictly positive molecular mass for each metabolite. With stoichiometrically inconsistent internal reactions omitted, we computed the flux consistent subset of internal stoichiometrically consistent and external reactions to derive a flux balance model with 10,600 reactions involving 5,835 metabolites. This algorithmic approach reinforces the standards described in[79] for conversion of a reconstruction into a mathematical model for flux balance analysis. It also provides an means of confirming mass balance that is complementary to atom mapping or checking of elemental balance. The general purpose source code for conversion of a reconstruction into a model for flux balance analysis, as well as tutorial examples of the procedure, are made available within The COBRA Toolbox[80], documented here: https://opencobra.github.io/cobratoolbox/.

Metabolic function test

The leak free model was then used to test for non-zero fluxes through 431 metabolic objective functions (Supplementary Data File 3), which represent the typical biochemical functions that could be performed by any cell within the human body.

Model Recon 3D content

At its end, the model of Recon 3D is stoichiometric and flux consistent and contains 10,600 reactions, 5,835 metabolites, and 2246 transcripts. A comparison of its content is provided in Supplementary Table 1(A-C).

Comparison of published Recon 2.2 and Recon 3

The reaction content of the recently published Recon 2.2 by Swainston et al[89], was assessed for its inclusion into Recon 3. Recon 2.2 was assembled from the published (i) Recon 2[36] (6936 reactions), (ii) drug module[57] (664 reactions), and (iii) transport module[40] (68 reactions). The transport module is already included within Recon 3, and drug module consists of drug metabolic and transport reactions. Since Recon 3 is a global reconstruction that captures the metabolic repertoire under normal healthy condition, the drug reactions were excluded. E.g., the published drug module contains metabolic and transport reactions of statins group of drugs, which is prescribed to consume only under hypercholesterolemia to maintain a controlled blood cholesterol levels. Of the remaining 117 reactions, twelve reactions already existed in Recon 3. When the final 105 (i.e., 117-12 = 105 final reactions) reactions were analyzed, 77% of these (i.e., 81/105) were fatty acid oxidation reactions, occurring either in mitochondria or peroxisome. These reactions either existed in Recon 3 with their corresponding counterparts, or, the same reaction existed with a modified metabolite abbreviation. E.g., (i) 'FAOXC10080m' in Recon 2.2 has its Recon 2/3 counterpart as 'FAOXC10C8m', (ii) 'FAOXC10080x' in Recon 2.2 is same as 'FAOXC10C8x' in Recon 2/3, (iii) 'FAOXC182_9E_12Em' in Recon 2.2 exists in Recon 2/3 as 'FAOXTC182TC162m' with a modified metabolite abbreviation for trans-7,10-hexadecadienoyl CoA. The leftover ones belonged to (i) carnitine shuttle and biomass (six reactions each), (ii) transport and fatty acid activation (three reactions each), (iii) NAD metabolism and exchange (two reactions each), and (iv) bile acid and nucleotide metabolism (one reaction each). Moreover, in Recon 2.2, 471 reactions present in Recon 2 and Recon 3 were removed without detailing evidence for removal. Consequently, we did not remove these reactions from Recon 3. In summary, Recon 2.2 did not provided any new reactions that should be considered for inclusion in Recon 3.

Model simulations and validation

Infant Growth Simulations

To validate the simulation capacity of Recon3D, we carried out simulations of infant growth on human breast milk using the Simulation Toolbox for Infant Growth with focus on metabolism (STIG-met)[91]. The STIG-met facilitates genome scale metabolic simulations of infant growth, accounting for age dependent changes in metabolism. It was devised using the model HMR 2.00[92], here we replace HMR 2.00 with Recon3D and replicate the simulations. Metabolite identifiers for the constituents of breastmilk and infant biomass were mapped to their Recon3D counterparts using CHEBI identifiers. The biomass equation of infant was then transferred to Recon 3D. The model dependent conversion factor for energy expenditure in kcal to mol ATP was recalculated and found to be 22.5 kcal/ATP for glucose (28.13 in HMR 2.00) and 20.42 for palmitate (27.34). The simulated growth curves were in good agreement with the original simulations (Supplementary Figure 4). STIG-met simulations are made available through: https://github.com/SBRG/Recon3D/

Supplementary Note 3: GEM-PRO reconstruction, QC/QA and analysis

We followed the previously described procedure[93] to map, assess, and refine PDB or homology models for integration into the genome-scale model (Supplementary Data File 11). In creating a GEM-PRO for the *H. sapiens* model, Recon 3D, there were a number of additional challenges which led to supplements in our ID mapping workflow. Mainly, isoforms of genes led to inconsistencies between database entries and difficulty linking to available homology models. Additional QC/QA steps were taken in order to ensure the correct sequence was being retrieved (Supplementary Figure 5). In total, out of the 20,266 human proteins documented in UniProt[2] (queried July 2016), 19,213 are functionally annotated (i.e., not hypothetical) and 17% of this subset is metabolic, well-characterized and included in Recon3D. This is explained by the following points: (1) As this is a metabolic model, we only annotate genes/proteins that participate in metabolism. The rest (83%) represents, to a large extent, a population of proteins that extends beyond metabolism. Many may include transcription factors, protein degradation machinery, protein synthesis machinery, etc. (2) We rely heavily on biochemical assays and literature that have characterized the proteins (such that we can assign stoichiometric

coefficients to the metabolites participating in these reactions). (3) Network reconstructions undergo additions every couple of years, with new subsystems being added to them (as more information is discovered about the proteins in those subsystems); a great visual for this can be found in ref[94].

Identifier mapping

Mapping to UniProt accession numbers

For a given gene in Recon 3, there are a number of associated isoforms, annotated as the gene and a isoform number, separated by a decimal (e.g., "314.2"). We take the gene identifier, which represents an Entrez gene ID[37], and directly map this to its corresponding UniProt accession number (UAC) utilizing the UniProt ID mapping service. Then, we directly map isoform numbers to available isoforms in the UAC entry. These are annotated with reviewed isoform-specific sequences, allowing us to filter for the correct experimental PDB structure in later stages.

Mapping to RefSeq, Ensembl, and PDB identifiers

In some cases, the number of isoform sequences annotated in Recon 3 does not match the number of isoforms available in UniProt. For these, we generated a separate mapping pipeline to the RefSeq and Ensembl databases[95]. The Bioservices Python package[96] and Ensembl Biomart tables[97] were used in order to first map the gene IDs without their isoform identifier to their corresponding entries, and then back to isoform IDs according to the transcript name as listed in Ensembl (Supplementary Figure 4). In the final master data frame, the source of these sequences is noted in the RefSeq and Ensembl columns. The information here was also utilized in order to cross-reference what was successfully mapped with the UniProt mapping service. Once the correct isoform entry was found, available PDB mappings were found using the entry ID (UniProt, RefSeq, or Ensembl Protein ID), or by a sequence BLAST to the PDB. We note that the difficulty in mapping isoforms and inconsistencies between databases points to a larger need of consistency and standardization for this biological property (Supplementary Data File 12).

Quality assessment and model structure refinement

To ensure that the highest quality structures were chosen for integration, we set cutoffs for sequence identity at 70% and resolution at 2.8 Å. These cutoffs filter out experimental structures which are found mostly in the subsystems of ROS detoxification, chondroitin sulfate degradation, and nucleotide sugar metabolism. Overall, 70% of all available structures are retained as suitable experimental structures and subject to further minimal refinements, and 30% are marked “lower/medium quality” and flagged for homology modeling (i.e. primarily to fill in missing residues in the protein). In total, 255 experimental structures from the PDB were subject to our model refinement pipeline[93], and 241 were successfully modeled to represent the correct wild-type sequence. The remaining 14 were manually inspected and adjusted accordingly for input to the model refinement pipeline.

Homology modeling

For PDB structures with missing residues, we have filled in the gaps by querying previously generated databases of I-TASSER homology models[98,99], and manually generating homology models for genes that were not part of these databases using a previously defined protocol[100]. In the final master GEM-PRO data frame (Supplementary Data File 11), we note where available homology models have been mapped to their respective genes. For most homology modeling procedures, the amino acid sequence of a protein is all that is required to generate a homology model of a protein. It is important to note that certain PDB structures with unresolved residues or gaps in the structure can also be homology modeled to enhance the structural coverage of the amino acid sequence. Any sequences longer than 600 amino acids long were not homology modeled.

The databases that we map to are contributed by Skolnick, et al., who have previously generated homology models for the *H. sapiens* genome in their SUNPRO database, excluding sequences above 600 amino acids in size[99]. These models were produced by a newer version of the I-TASSER workflow, TASSERVMT-lite. This information was organized by RefSeq ID. As an additional QC/QA step, we ensured the proper homology model was linked to a given gene by conducting pairwise sequence alignments for each gene's sequence and the sequence corresponding to the homology model. Only models with a sequence identity of 90% or higher were considered a match, and were subjected to additional QC checks[93].

We assessed the overall quality of the information coming from homologous templates in terms of (i) which organism the protein was crystallized from; (ii) the resolution of the PDB template and (iii) the deposition date. We used these properties to compare the templates that were used to construct homology models in the previous GEM-PRO models with those of the recently updated versions (Supplementary Tables 2-4). To obtain organism information, we use the PDB 4-letter identifier to query the PDB database for a numeric taxonomy identifier, which is then used to query the UniProt taxonomy url (http://www.uniprot.org/taxonomy). For information regarding the deposition date and the resolution, we parse the PDB header using the Biopython PDB module[101]. A C-score and root mean squared deviation (RMSD) from the original template provide an estimate that indicates how close the model is to the native structure. In general, C-score is in the range [0,1] and a value greater than 0.5 indicates higher confidence. The models are also ranked-ordered based on the structure density during I-TASSER refinement simulations (for more information, please see the original protocol[102]). The average C-score of the mapped models is 0.68 ± 0.16. For the homology template-based quality checks, we find that the homologous templates are derived from 162 different organisms, with the top 60% coming from *H. sapiens* (35%), *E. coli* (7%), *M. musculus* (5.2%), *R. norvegicus* (4.9%) and *S. cerevisiae* (3.4%). The average resolution of the templates is 2.2 ±0.74 Å. The most recent templates date back to 2011 and there are 131 templates that have been deposited as of 2008. To assign a final quality to all models based on energetic and geometric parameters, we utilize the PSQS[103] and PROCHECK programs[104]. The average total PSQS score for all homology models is -0.12 ± 0.12 (Supplementary Figure 6). PROCHECK reports that the average percentage of residues in favored positions is 85.7% ± 11%, and an average overall G-factor of -0.96 ± 2.7.

Dissemination and use of the Recon 3D minimal GEM-PRO model

To facilitate the future use of the Recon 3D GEM-PRO model, the procedure to collect sequence and structure information as described above has been consolidated into a shareable JSON file, which we call the "minimal" GEM-PRO needed to start structural analyses. This model assigns a single representative structure per gene in the reconstructed metabolic model, and is available at https://github.com/SBRG/Recon3D. The accompanying software package required for reading and working with the GEM-PRO JSON is available at https://github.com/SBRG/ssbio. This entire repository can be cloned to a user's computer and contains Jupyter notebooks in the root directory to guide a user through the content available in the Recon 3D GEM-PRO model (Recon3D_GP - Loading and Exploring the GEM-PRO.ipynb) as well as to update the model with revised sequence information or newly deposited structures in the PDB (Recon3D_GP - Updating the GEM-PRO.ipynb). This repository also includes all sequence and structure files mapped per gene, metadata downloaded through UniProt and the PDB, as well as the ability to rerun the QC/QA pipeline with different parameters such as sequence identity and resolution cutoffs. These notebooks also include basic visualization features enabled with the NGL viewer package[105].

Mapping and alignment of PDBs to their representative protein domains

To obtain a direct mapping between a PDB file and its representative domain, we use the PDB's RESTful web service tool (http://www.rcsb.org/pdb/software/rest.do). Querying the entire metabolic proteome for Recon 3D, we obtain representative domain template structures for all experimentally determined protein structures (Supplementary Data File 23). Once a protein's representative template structure has been determined, we perform pairwise 3D structural alignments[106,107] between the protein and the representative domain using the RCSB PDB jCE/jFATCAT Structure Alignment server and stand alone software (http://source.rcsb.org/jfatcatserver/). We also used the RCSB PDB Protein comparison tool for domain-domain comparisons or PDB-PDB comparisons (http://www.rcsb.org/pdb/workbench/workbench.do). The representative domains are either assigned as SCOP domains (i.e., where domain IDs begin with 'd') or PDB-annotated domains, in which the representative template structure is taken from a given PDB file (e.g., PDP:4RFZAb is taken from PDB entry 4rfz, chain A).

Addition of molecular structures of metabolites in Recon 3D

To identify structures for the given set of metabolites in Recon 3D, we evaluated a number of databases where metabolite structures are publicly available, such as PDB (ligand-expo: http://ligand-expo.rcsb.org/, http://ligand-expo.rcsb.org/ld-search.html), PubChem[39] Url (https://pubchem.ncbi.nlm.nih.gov/), and ChEBI Url (http://www.ebi.ac.uk/chebi/). We downloaded structures in various formats: 2D structure in .mol format (ChEBI), 3D structure in .sdf format (PubChem[39]) and in .pdb/.xyz format (RCSB). Supplementary Data File 15 provides all the information content processed for metabolites in Recon 3D, which includes SMILEs and INCHI descriptors, Kyoto Encyclopedia of Genes and Genomes (KEGG)[108] IDs, CID IDs, CID file names, ChEBI file names, ChEBI IDs, and experimental coordinate file URL locations and the ideal coordinate file name. The ChEBI mapping procedure contained the following steps: (i) identification of the particular metabolite (e.g., 11-deoxycorticosterone) from ChEBI using the ChEBI source link. The metabolite name will be the starting point of search which is taken from the metabolite names in the Supplementary Data File S1.xlsx; (ii) checking the molecular formula and charge (neutral or charged) of the metabolite in the ChEBI database; (iii) capturing the ChEBI link, ChEBI ID, SMILES and INCHI into the respective fields in the dataset spreadsheet. (iv) 2D-structure is downloaded in .mol format. The same overall search was conducted in Pubchem and PDB (Ligand expo) with slight variations as to the initial search inputs and file type outputs. In Pubchem, we identified a given metabolite using the Pubchem source link whereas, for PDB, we used SMILES/INCHI descriptors as inputs for the initial search. Finally, additional structure files were downloaded from other metabolic databases, including the Lipid Mass Structure Database (LMSD)[109], BioPath database[110], ChemSpider database[111], and the Human Metabolome DataBase (HMDB)[24]. Furthermore, if no metabolite structures could be identified in any database, we drew the structures manually using ChemAxon (www.chemaxon.com) (Preciat et al. 2017 *in press*).

Atom mapping data from reactions in Recon 3D

An atom mapping is a mechanistic description of a chemical reaction where each substrate atom maps to exactly one product atom, and vice-versa. The set of atom mappings for a reaction reveals key aspects of reaction mechanism, e.g., chemical bond breakage and formation. Two reactions, with identical stoichiometry may occur by different reaction mechanisms, corresponding to distinct sets of atom mappings. As such, atom mapping information opens the door to new applications beyond the level of reaction stoichiometry alone. For example, when reaction stoichiometry is combined with information on atom mappings for a whole metabolic network, one can identify the set of linearly independent conserved moieties for network, each of which corresponds to a particular identifiable molecular substructure[112]. The corresponding set of conserved moiety vectors forms a sparse basis for the left nullspace of the stoichiometric matrix[113]. Constraints derived from this left nullspace basis are a fundamental part of kinetic modelling because the amount of each conserved moiety is a time invariant. Moreover, since conserved moiety consists of a set of

invariant atoms that follow the same path through a metabolic network, in principle, it is sufficient to label a single atom within a moiety in order to isotopically detect the possible paths of an entire moiety through a metabolic network [114].

Generation of atom mapping data requires chemical structures, reaction stoichiometry and an atom mapping algorithm. We provide metabolite chemical structures, for 2369/2797 (85%) of the unique metabolites in the Recon3D derived model. No chemical structures are provided for the remaining 428 unique metabolites due to insufficient information about the precise chemical structure of some metabolite species (e.g., eumelanin), or because some Recon3D reactions do not specify the nature of the reactant sufficiently, e.g., in lipid metabolism, a substrate may correspond to a family of compounds, which may differ slightly in structure, due to the number and position of double bonds.

Reaction atom mappings, are provided for 89% (7,804/8,791) of the internal reactions in the Recon 3D derived model, based on a prior comparison of the predictive accuracy of six different atom mapping algorithms (Preciat et al. 2017 *in press*). Atom mappings were predicted using the Reaction Decoder Tool[115], and the DREAM algorithm[116] for 7,535 (86%) mass balanced reactions with implicit and explicit hydrogens respectively, while Reaction Decoder Tool, and the CLCA algorithm[117] were used to predict atom mappings for a further 269 reactions with incompletely specified metabolites (e.g., R group) with implicit and explicit hydrogens respectively. We compared these predictions for internal reactions to a set of 512 reactions with atom mappings that we and others manually curated (172 new and 340 pre-existing from the BioPath database[110]). This reaction set is representative of all six top level EC numbers. Based on this comparison, we observed that the predicted atom mappings are highly accurate for most of the reaction types (Supplementary Figure 7). No predictions are provided for the remainder of the reactions from Recon 3D, as no chemical structure was available for one or more metabolites. Metabolite structures and atom mappings are provided as chemical table files (MOL, RXN and SMILES formats). Atom mapping data, that contain information about the atoms, bonds, connectivity, and coordinates of a molecule or set of molecules involved in a reaction, are presented in two standard chemical formats, SMILES and RXN files, which are based on standard chemical formats for metabolites, SMILES and MOL files, respectively, and they are disseminated via the Virtual Metabolic Human database (VMH, https://vmh.life/)

Mapping to human variation and pharmacogenomics datasets

The dataset of human single nucleotide polymorphisms (SNPs) and single nucleotide variants (SNVs) was collected from UniProt from a subset of protein altering variants from the 1000 Genomes Project (available in via FTP download: homo_sapiens_variation.txt.gz). Furthermore, all SNPs/SNVs for model genes were downloaded directly from dbSNP (http://www.ncbi.nlm.nih.gov/SNP/) via the Ensembl BioMart interface[97]. We then selected all variants that were characterized to be “damaging” or “possibly damaging” as a predicted functional impact using the PolyPhen2 bioinformatics tool[118]. Functional annotations of the missense mutations were also annotated using SIFT (http://sift.jcvi.org/). In total, we selected 3536 known or potentially deleterious missense variants that mapped to the human metabolic network (Recon 3D). We find 1385 genes in metabolism that have variants characterized within this dataset (43% of genes in the metabolic network). In addition, we linked the missense variants to their gene-drug associations (clinically relevant pharmacogenomics interactions) using the PharmGKB pharmacogenomics database (https://www.pharmgkb.org/). All annotated gene-drug pairs contain information such as: (i) dosing guidelines; (ii) drug label annotations and each pair is generally specified in more than 1 type of annotation (dosing guideline, drug label, clinical annotation, variant annotation, VIP, or pathway). These selected pharmacogenomic associations allow us to understand whether certain missense variants have functional effects on drug therapies. All selected missense variants and their drug associations have been provided as Supplementary Data Files 15 and 16.

Supplementary Note 4: Metabolic network visualization

A planar visualization of human metabolism, ReconMap 2.0[119], which was previously manually drawn using CellDesigner[120] and saved as an SBML Level 2 Version 4 format[120,121]. This map included 4,030 chemical species and 5,535 reactions, whose graphical appearance was encoded in a CellDesigner-specific annotation of the SBML file. In order to improve the software interoperability of this map and to facilitate subsequent post-processing of the graphical information, this SBML file was upconverted to Level 3 Version 1[122,123] with the extension packages Layout[124] and Render.

Conversion from CellDesigner's SBML Level 2 Version 4 to SBML Level 3 Version 1 with Layout and Render package

The first step was conducted using CellDesigner-Parser (https://github.com/funasoul/celldesigner-parser/). CellDesigner-Parser was developed as part of a Google Summer of Code project in 2016 with the aim to convert SBML files with CellDesigner-specific graphics information into SBML with Layout extension, which is a community format for storing size and position of graphical components. However, the CellDesigner map identified secondary metabolites using a specific color-coding scheme. Such rendering information is, however, out of the scope of the Layout extension for SBML. In order to include this information, the program needed to be extended so that it was able to export color information in form of the Render extension for SBML.

Postprocessing

For the second step, a postprocessing algorithm was developed and implemented in Java™ based on JSBML[125,126]. This algorithm performs the following operations:

1. Detection of SBML elements with undeclared identifier and automatic completion
2. Update of MIRIAM (Minimal Information Required in the Annotation of Models)[127] URNs to identifiers.org URIs[128].
3. Detection and removal of unconnected metabolites (species) and graphical objects (nodes)
4. Replacement of pseudo-metabolites that were used to label pathway headlines with text glyphs
5. Removing orphan styles and colors that were no longer used because of deletions of components the previous steps
6. Shaving away all CellDesigner annotation because this information is now encoded in a combination of SBML with Layout and Render extension packages.
7. Determination of SBO terms (Systems Biology Ontology)[128,129] for model components based on color codes. These are required for the correct assignment of participant roles in Escher's data model.
8. Merging of MIRIAM annotations with identical qualifier, thereby replacement of model qualifiers with biological qualifiers where necessary and alphabetic sorting of cross-references
9. Trimming the names of all metabolites (species) with leading/trailing blanks
10. Updating cryptic identifiers such as "s7126" that were internally used by CellDesigner to human-recognizable BiGG[130] ids such as "h_n".
11. Identification of duplicate metabolite definitions and solving contradicting annotation before deleting duplicates

12. Assignment of correct compartments to metabolites (species), such as e (external), n (nucleus), c (cytosol), g (Golgi apparatus), etc.[130,131]

13. Pulling of additional database cross references from BiGG Models database and assignment of them to model components (compartments, species, and reactions) using the stand-alone ModelPolisher application[132] (https://github.com/SBRG/ModelPolisher/)

14. Removal of empty XHTML head statements in notes of all SBML components

As a result, an SBML Level 3 Version 2 Release 2 file with Layout and Render package was obtained.This format of ReconMap 2.2 is available to download at https://vmh.life/#downloadview. The conversion of ReconMap 2.0 into a standard format enables its use with a wider variety of visualization tools than was previously possible with the CellDesigner format. An overview of the resulting map is displayed in Supplementary Figure 8.

Conversion into Escher JSON format

The aforementioned SBML file was converted to an Escher[133] specific JSON format using the stand-alone application EscherConverter[132,133] (https://github.com/SBRG/EscherConverter/). Originally, EscherConverter was designed as a converter from Escher's JSON format to SBML Level 3 Version 1. During the Google Summer of Code 2016 EscherConverter was extended to a bi-directional converter. As a result, an Escher representation of the manually drawn map was obtained that needed one additional post-processing step in order to scale the distances between all metabolites. This stretching of the map was required in order to adjust the differences between the original SBGN-based[134] CellDesigner layout and the Escher representation. The Escher compatible JSON format of ReconMap 2.2 is available to download at https://vmh.life/#downloadview. Supplementary Figure 9 shows an overview of this map.

In addition, Escher maps for the human red blood cell and the human platelets were manually drawn in Escher based on the models and published network maps of iAB_RBC_283[43] (Supplementary Figure 10) and iAT_PLT_636[43,135] (Supplementary Figure 11) in BiGG[130] Models Database.

Supplementary Note 5: 3D mutation hot spot analysis

We further filtered the set of mutations (whose genes are associated with experimental protein structures) based on whether the location of the mutated residue itself was resolved (e.g., certain protein domains are unresolved due to flexibility or unstructured regions of the protein being challenging to crystallize). Once the subset of mutations were established to (i) be linked to genes with experimental protein structures and (ii) be located within regions of the protein that were experimentally determined, we carried out 3D structure alignments between all proteins and their representative domains (mapping to representative protein domains is described previously in section: "mapping and alignment of PDBs to their representative domains"). In contrast to sequence alignments, 3D structure alignments find a best fit in terms of the three-dimensional shape or geometry of two proteins. Therefore, any two proteins that have different sequences but share a common domain architecture cauren be successfully aligned in 3D space. Similar to sequence alignments, the 3D structural alignment provides a direct residue-to-residue mapping for residues that share structurally equivalent positions in a common/shared domain motif. Once this residue-to-residue mapping was established for all proteins in our dataset, we located 3D "hotspot" mutations by tallying all residues in the representative domains that map to mutated residues in a given protein of interest. To this end, certain residues in a representative domain may have multiple hits if more than one gene is linked to that representative domain and the same structurally equivalent residue is mutated across various genes. Supplementary Table 17 provides the mapping between the residue number of the Uniprot missense variant > the PDB residue number > the PDB chain where the residue is located > the representative domain ID linked to a given PDB chain > the structurally equivalent residue within that representative domain.

Once all the mutated residues with structurally-equivalent positions in all representative domains were found, we performed statistical analyses on the number of near neighbors (within 5 or 10Å sphere) of a given mutation position (Supplementary Table 5). We also tallied the number of mutations within +/- 5 amino acids (in sequence) from a given mutation position. In most cases, we find a higher number of mutations in the 3D vicinity of a mutation than by sequence-position alone. We used Uniprot to obtain information about the SNP/SNV variant, disease associations and domain motifs for all mutations in our dataset (Supplementary Files 19 and 20).

We performed a literature search to explore disease relevance in all the most prominent domains (Supplementary Figure 12(a)). We found a number of supporting articles that suggest these domains do play a prevalent role in various disease phenotypes: kinase domains[136–138]; phosphatase domains[139,140]; ABC transporters[141]; and peptidase s53 family[142].

Somatic cancer mutation mapping and data preprocessing

3D hotspot analysis

We used the TCGA level 3 variant data in the cBioPortal (http://www.cbioportal.org/). For this study, we used high level (processed) data from a subset of pre-analyzed mutations from 178 tumour–normal pairs of lung squamous cell carcinoma[143]. When the MutSig1.0 approach was applied on this dataset[144], it identified 450 genes as significantly mutated. Starting from this set of genes, we identified a subset of 86 genes that have Uniprot accession numbers and protein structural information. Within this set of genes, we found that 889 somatic cancer mutations map to residues that have been successfully resolved in the crystallographic structures of proteins. We used the list of 86 genes to query the cBioportal web-based dataset and downloaded various information including: somatic cancer mutations, cancer study sample IDs, amino acid mutations, annotations (coming from various sources, such as http://oncokb.org/ and https://www.mycancergenome.org/), type of mutation, copy number changes, overlapping mutations in COSMIC, the predicted functional impact score (from Mutation Assessor), variant allele frequency in the tumor sample, and total number of nonsynonomous mutations in the sample. A summary of cancer data sets used in this study is given in Supplementary Data File 21 and a detailed summary of all somatic mutations for this set of genes is provided in Supplementary Data Files 22-23. The 3D hotspot analysis was carried out as detailed above and mutations were rank-ordered on the basis of how many mutations fell within a 5Å sphere (i.e., number of nearest neighbors). From this ordering, we took the top 25% of the dataset and compared the mutation annotations (e.g., whether the gene is a known oncogene, the mutation is known to cause tumorigenesis, the mutation is found in other cancer types, the mutation is linked to drug-associations, etc.) to the entire dataset. We linked this analysis to other disease-relevance information to create disease networks (Supplementary Figure 12(b)).

Statistics. We performed a sensitivity analysis to understand whether the selection of data points had an effect on the significance of these results. We find that the 3D hotspot analysis is more likely to select somatic mutations compared to a random selection:

Data points selected	Percentage of total data	Pval (3D)	Pval (random)
50	0.065	0.021	0.241
100	0.131	0.049	0.112
200	0.262	0.029	0.182
500	0.655	0.017	0.182

700	0.917	0.027	0.182

For annotations of mutations that are known oncogenes (KO) and known hotspots (HS), selection of the data based on 3D hotspot analysis is significant, regardless the number of data (or % of data) selected (pval < 0.05). Compared to a random selection, our computed (using a two-tailed t-test) pval is > 0.1.

Data points selected	Percentage of total data	% of KO (3D)	% of KO (random)	% of HS (3D)	% of HS (random)
50	0.065	0.370	0.046	0.725	0.098
100	0.131	0.546	0.081	0.843	0.196
200	0.262	0.825	0.220	0.882	0.470
500	0.655	0.825	0.430	0.882	0.640

The above 3D hotspot analysis approach was then applied to 22 genes from which cancer mutations have already been analyzed[145] and 92 genes involved in cholesterol metabolism, owing to the fact that cholesterol biosynthesis plays an important role in GBM[146]. All genes are given in Supplementary Data File 23.

Gene deletion simulations in GBM

In silico single gene deletion (SGD) simulations were performed as previously described[147]. Given a certain GEM, the simulation of a SGD was performed by formulating the linear program problem (1) for each gene *g* in the GEM:

$$(1)\ max\ v_{obj} \text{ subject to:}$$

$$(2)\ 0 < v_{obj} < \gamma$$

$$(3)\ S \cdot v = 0$$

$$(4) -1000 \leq v_j \leq +1000\ \forall j \in \{Exchange\ reaction\ indexes\ for\ medium\ metabolites\}$$

$$(5)\ v_r = 0\ where\ r \in \{Reaction\ indexes\ univocally\ encoded\ by\ gene\ g\}$$

where v_{obj} is the flux through the biomass equation, γ is an arbitrary number set to 1, ***S*** is the stoichiometric matrix of the GEM (that is a *m* x *n* matrix where *m* is the number of metabolites and *n* is the number of reactions and each (*i,j*) entry is the stoichiometric coefficient of the metabolite corresponding to row *i* in the reaction corresponding to column *j*), ***v*** is the vector containing the values of the fluxes through each reaction in the GEM, and *j* indexes each exchange reaction known to be present in a rich mammalian medium (Ham's medium, HAM) as defined in Agren et al. (2014)[88]. The simulation was carried out for the following GEMs: Recon3D, HMR2.00, and 22 personalised GEMs for glioblastoma multiforme (GBM) previously reconstructed using HMR2.00 as a template from as many GBM expression profiles retrieved at The Cancer Genome Atlas[148]. (Supplementary Figure 13)

For the purpose of the study, the optimization part was not relevant as we were solely interested on whether a feasible region existed upon the constraint imposed by the gene knockout (5), i.e. whether the fact that the encoded reactions could not carry flux implied no flux in the biomass equation. A gene was deemed essential *in silico* when there was no

solution to (1), i.e. there could not be found a flux distribution such that the biomass equation carried non-zero flux. The problem was formulated using native functions in the RAVEN Toolbox[149] and solved using MOSEK v.7. SGD simulations and all the native functions used can be found in the files made available through: https://github.com/SBRG/Recon3D/

Supplementary Note 6: Metabolic response phenotypes across drugs

To compute metabolic pathways with gene expression perturbed by drugs, the human metabolic network model was first converted into an irreversible network. Then, the MetChange algorithm[150] was run using gene expression presence/absence p-values from the Connectivity Map (Cmap) database[151] build 02. When multiple controls were present, a standard score was generated. When a treated sample was from a batch with a single control, the mean and standard deviation of all control samples was used instead. Cell line standard scores were then generated in the following manner. First, for each cell line, the median scores of all samples for each drug were found and used as the cell line-specific response. Then, to simplify compartment-specific scores to a general metabolite response, cytosolic metabolite scores were taken when available. If no cytosolic metabolite existed, the median of scores across all compartments was taken as the metabolite score. Finally, a standard score across all drugs was calculated for each cell line. Consensus drug perturbations across cell lines were calculated by averaging cell-specific MetChange scores and standardizing across all drugs.

Determination and analysis of drug indication signatures

Drug indications were taken from Side Effect Resource (SIDER) database[152] for all available drugs overlapping with the Cmap database. Synonyms were aggregated when present as with side effects. A minimum of 10 drugs for each indication were required for the inclusion in the analysis, corresponding for a much greater number of expression sets for each indication. A total of 48 drug indications were analyzed for 1459 expression sets corresponding to 334 drugs. A genetic algorithm (Supplementary Figure 15) was then implemented as follows. The matrix of MetChange scores for the 1459 expression sets was input as well as the corresponding presence/absence calls for a particular indication for each expression set. A maximum number of predictor metabolites was set to 20 metabolites. A set of 125 individuals was generated with length of the number of metabolites in the MetChange scores and values of -1, 0, or 1 to indicate negative, zero, or positive prediction of a high MetChange metabolite score on likelihood of indication. Each expression set was then scored for the indication for each individual by multiplying the predictor set for the individual by the MetChange scores for the expression set. These indication scores were then ranked, and an ROC curve was generated by comparison of scores with the drug indication presence/absence for each expression set at increasing thresholds. The AUC of this curve was then used as the objective function to maximize in the genetic algorithm (Supplementary Figure 15(b)). Importantly, signature metabolites were on average no more perturbed than other metabolites (mean MetChange score=1.008 for signature metabolites and 1.018 for all metabolites), so it does not appear that this association is a result of the previously-determined association of high scoring metabolites with corresponding drugs. Genetic algorithm creation, mutation, and crossover parameters were used as implemented in the OptGene function of the COBRA Toolbox 2.0[80]. The genetic algorithm was solved using the Global Optimization Toolbox in MATLAB (MathWorks). Statistical analysis of literature co-associations of drugs with indication metabolite signatures was performed with (1) a Wilcoxon rank sum test on the drug indication metabolite signatures against 1000 permutations of the signatures for the AUC of predicting presence/absence of literature drug/metabolite co-association, and; (2) a hypergeometric test for the enrichment of literature association among drug/metabolite pairs in predicted signatures.

Identification of metabolic signatures in metabolism and disease

We then examined particular gene indication signatures and found that conserved these metabolic signatures have implications with respect to known clinical effects of classes of drugs. Schizophrenia had a signature that was highly

conserved (median AUC of 0.80 for cross validation gene signatures), even though dopamine and serotonin receptors are not expressed in the cell types examined (Supplementary Data File 24). Antipsychotics have been known to activate a number of other signaling pathways, including ubiquitous histamine receptors, suggesting non-canonical pathways may be facilitating drug effects of potential clinical importance. Interestingly, histamine receptors via phospholipase C are known to cause epigenetic modulation through histone acetylation.

Of particular interest was the antipsychotic-induced upregulation of SCD1, stearoyl-CoA desaturase, which was contrary to the findings of a recent study[10]. SCD1 -/- mice have been found to have reduced levels of VLDL and TG[11]. In support of a link between clinical outcome and these non-dopamine dependent metabolic effects, antipsychotic responders were found to have higher VLDL and TG levels compared with non-responders[12]. The observed upregulation of SCD1 has further implications for the known increased risk of cardiovascular disease among antipsychotic-treated schizophrenia patients[13], compared with antipsychotic-naïve patients in which risk factors have been found to not be present[14]. Inhibition of SCD1 has been proposed as treatment for metabolic syndrome[15]. Supplementation with poly-unsaturated ω-3 fatty acids has been proposed as a treatment for schizophrenia, with encouraging but not fully validated results[16, 17]. ω-3 fatty acids have been shown to decrease expression of SCD1[18]. Interestingly, simultaneous treatment with antipsychotics and ω-3 fatty acids negates the well-known benefit of the latter on cardiovascular health[19]. The findings in this analysis suggest that the decreased cardiovascular risk of ω-3 fatty acids and increased cardiovascular risk of antipsychotics may be linked through SCD1. Concomitant treatment with antipsychotics and SCD1 inhibitors may serve to reduce the risk of cardiovascular disease in schizophrenic patients. This analysis shows that metabolic signatures can be used to identify conserved effects between superficially disparate classes of drugs (Supplementary Tables 8-9).

Supplementary Tables

Supplementary Table 1. Comparison of the components of Recon 1, Recon 2, Recon 2.2, HMR2.0, and Recon 3D, and the gain in knowledge.

Supplementary Table 1A.Overview of key features of Recon 3D and comparison with its predecessors. [a] As defined in the present work.

		Recon 1[34]	**Recon 2**[36]	**Recon 2.04**[36]	**HMR 2.0**[81]	**Recon 2.2**[89]	**Recon 3D**[a]
Incorporated published metabolic reconstruction	Recon 1[34]	NA	√		√		
	Recon 2[36]		√	√	√	√	√
	EHMN[76]		√	√	√	√	√
	HMR[88]				√		√
	Drug module[57]					√	√
	Transport module[40]					√	√
	Host-microbe relevant reactions[8]						√
New subsystem extension	Dopamine metabolism[a]						√
	Bile acid metabolism[a]						√
	Sphingolipid metabolism[a]						√

	Lipid metabolism[a]						√
	Metabolic and transport reaction addition based on 10 metabolic data sources[a]						√
	Peptide metabolism[a]						√
	Organ-specific reactions[a]						√
	Diet-specific reactions[a]						√
Gene ID's		EntrezGene	EntrezGene	EntrezGene	Ensemble	HGNC	EntrezGene (HGNC, Ensemble)
GPRs		√	√	√		√ (refined)	√ (refined and expanded)
Browsable website			√	√			√
Mass- and charge balance to pH 7.2		√	√	√			√
Protein structures							√
Protein Domain mapping							√
Mol structures							√
Atom-mapping							√

Supplementary Table 1B. Comparison of major properties of Recon 3 and its predecessors presented here. *Exchange reactions were added for all metabolites present in the 's' compartment to allow the computation of blocked reactions. The compartments present across all the models include, cytosol [c], extracellular space [e], Golgi apparatus [g], lysosome [l], mitochondria [m], nucleus [n], endoplasmic reticulum [r], and peroxisome [x]. Recon 3 and Recon 3 - models have an additional inner mitochondrial compartment [i].

	Recon 1	**Recon 2**	**Recon 2 - model**	**HMR 2.0***	**Recon 2.2**	**Recon 3**	**Recon 3 - model**
Reactions	3,742	7,440	7,440	8,793	7,785	13,543	10,600
Metabolites	2,766	5,063	5,063	6,006	5,324	8,399	5,835
Metabolites (unique)	1,509	2,626	2,626	3,161	2,652	4,140	2,797
Compartments (unique)	8	8	8	9	9	9	9
Genes (unique)	1,496	1,789	1,789	3,765	1,675	3,288	1,882
Subsystems	100	100	100	137	100	110	102
Deadends	3,40 (12%)	1,175 (23%)	1,212 (24%)	926 (15%)	1,213 (23%)	882 (11%)	0 (0%)
Blocked Reactions	1,273 (34%)	1,606 (22%)	2,123 (29%)	1,996 (23%)	1,873 (24%)	1,582 (12%)	0 (0%)
Size of S (m; n)	2766;3742	5063;7440	5063;7440	6006;8793	5324;7785	8399;13543	5835;10600

Rank of S	2,674	4,666	4,666	5,877	4,945	8,121	5,739
Sparsity (% of nonzero entries in S)	0.1382	0.0837	0.0837	0.0713	0.0781	0.0499	0.0654

Supplementary Table 1C. *Comparison of predicted and theoretical ATP yield on different carbon sources.* [a]*Taken from* [89]

	Recon 2.2: ATP yield [a]	**Recon3: ATP yield**	**Theoretical ATP yield** [a]
Glucose - aerobic	32	32	31
Glucose - anaerobic	2	2	2
L-Glutamine - aerobic	ND	22.5	ND
L-Glutamine - anaerobic	ND	1	ND
Fructose - aerobic	ND	32	ND
Fructose - anaerobic	ND	2	ND
Butyrate - aerobic	22	22	21.5
Butyrate - anaerobic	0	0	0
Hexanoate - aerobic	36	34.5	35.25
Hexanoate - anaerobic	0	0	0
Octanoate - aerobic	50	52	49
Octanoate - anaerobic	0	0	0
Decanoate - aerobic	64	67	62.75
Decanoate - anaerobic	0	0	0
Dodecanoate - aerobic	82.5	82	76.5
Dodecanoate - anaerobic	0	0	0
Tetradecanoate - aerobic	92	97	90.25
Tetradecanoate - anaerobic	0	0	0
Hexadecanoate - aerobic	106.75	113	104
Hexadecanoate - anaerobic	0	0	0
Octadecanoate - aerobic	120	129	117.75
Octadecanoate - anaerobic	0	0	0
Arachidate - aerobic	134	143	131.5
Arachidate - anaerobic	0	0	0
Docosanoate - aerobic	147.25	157	145.25
Docosanoate - anaerobic	0	0	0
Lignocerate - aerobic	160.5	168	159
Lignocerate - anaerobic	0	0	0
Hexacosanoate - aerobic	170.75	178.5	172.75
Hexacosanoate - anaerobic	0	0	0

Supplementary Table 2. Quality statistics of crystallographic structures in the GEM-PRO model. Mean sequence identity notes exact amino acid matches between sequence and structure, while mean completeness disregards exact matches.

Property	*H. sapiens*
Mean Sequence Identity	77.9 ± 26.5%
Mean Completeness	78.2 ± 26.2%
Mean Resolution	2.1 ± 0.5 Å

Supplementary Table 3. Quality statistics of the GEM-PRO model, compared to *E. coli*. [a]number of total genes with PDB structures (includes minimally modified) after QC/QA; [b]number of total genes with homology models, note that there may be overlap between PDB and homology model coverage; [c]mean quality score of PDB structures in the GEM-PRO model within a range of (0, 1]; [d]mean quality score of the homology models.

Model	PDB coverage[a]	Homology model coverage[b]	PDB quality score[c]	Homology model quality (C-score)[d]
E. coli	354/1366	1366/1366	0.89	0.42
H. sapiens	1238/3704	2343/3704	0.74	0.68

Supplementary Table 4. Physics-based energetic assessments of PDB structures and homology models in the human GEM-PRO, compared to *E. coli*. PSQS provides an total energetic score, with lower scores indicating better quality. PROCHECK provides geometric measures, and a G-factor below -1 is considered unusual.

Method	Quality measure	*E. coli*	*H. sapiens*
PSQS	Average total score	-0.164 ± 0.12	-0.12 ± 0.12
PROCHECK	Average % of residues in favored positions	87.1% ± 20%	85.7% ± 11%
	Average overall G-factor	-0.10 ± 0.27	-0.96 ± 2.7
Zhang	C-score	0.42 ± 0.06	0.68 ± 0.16

Supplementary Table 5. Top ten ranked mutation hotspots from missense mutations taken from UniProt. The column descriptions are as follows: 10 Å vicinity refers to the number of mutations occurring within a 10 Å sphere of the fatcat residue number; 5Å vicinity refers to the number of mutations occurring within a 5 Å sphere of the fatcat residue number; the representative protein domain ID; the structurally equivalent position in the domain template that maps to missense mutations; the number of sequential amino acids (+/-5 amino acids) in the vicinity of the fatcat residue; the total number of missense mutations that map to the fatcat

domain; the number of occurrences of the same position being mutated in different genes; the number of genes mapping to a shared/common protein domain.

10 Å vicinity	5 Å vicinity	Fatcat domain	Fatcat residue	+/- 5 amino acids (in sequence)	Number of SNPs	Number occurrences	Unique uniprot IDs
6	5	PDP:4RFZAb	545	5	33	1	22
7	5	PDP:4RFZAb	542	5	33	1	22
17	5	d1e2sp_	309	4	76	1	1
15	5	d1e2sp_	307	4	76	1	1
14	5	d1e2sp_	179	2	76	1	1
18	5	d1e2sp_	308	4	76	1	1
6	5	PDP:4RFZAb	543	6	33	2	22
15	5	d1e2sp_	306	4	76	1	1
9	5	d1e2sp_	154	5	76	1	1
8	4	d1e2sp_	244	2	76	1	1

Supplementary Table 6. SNP data from multiple gene variation databases (dbSNP[1], UniProt[2], PharmGKB[3]) analyzed using Recon3D. Missense mutations were mapped to their 3D structural coordinates in shared protein domains and assessed based on the degree of co-occurrence within the same 5 Angstrom sphere. Significant p value indicates that deleterious mutations are much more likely to co-occur within the same sphere than tolerated mutations.

Label of SNP effect (sift)	Number of SNPs	SNPs with structural data	p-value
deleterious	9368	604	p=0.03
tolerated	1982	270	p = 0.1
deleterious - low confidence	177	-	-
tolerated - low confidence	68	-	-

Supplementary Table 7: Summary of large-scale characterization of metabolic response to drugs and their intended use. Included are the database used in collecting relevant information, the number of compounds in the Connectivity Map that were able to be cross-referenced with the corresponding database, the number of mapped links, or properties, and the statistical significance of the associations with predicted drug response.

Property	Cross-referenced Database	Drugs Analyzed	Properties Analyzed	Statistical Significance
Drug Indication	SIDER & PubMed	334 drugs	47 drug indications	p = 4.9x10e-43 (hypergeometric test)

Supplementary Table 8: Summary of schizophrenia drug (antipsychotic) signatures and known metabolic changes in schizophrenia. Included are the Entrez gene name, the manually assigned pathways, the direction of gene expression response induced by drug

treatment found in the drug signature, and the known changes in the metabolic pathway most closely related to the gene found in schizophrenic patients (studies were selected to include those on drug-naïve patients when possible).

Gene	Pathway	Drug Perturbation	Occurrence in Schizophrenia
NDUFS1	electron transport chain	Down	Increased NDUFS1 suggested to be marker of early onset schizophrenia[153]
GMDS	irreversibly degrades mannose precursor to glycosylation	Down	Glycosylation abnormal in schizophrenia[154]. High-mannose glycan side-chains deficient[155]
CYP4A11	fatty acid catabolism	Down	ω-9 fatty acids lower in schizophrenia[156]. CYP4A catabolizes stearoyl-CoA, an ω-9 precursor
SCD	unsaturated fatty acid synthesis	Up	ω-9 fatty acids lower in schizophrenia[156]. Oleic acid, an ω-9 fatty acid, is a primary product of SCD from stearoyl-CoA
ACAT2	cholesterol biosynthesis	Up	Mutations in cholesterol synthesis regulatory genes associated with increased schizophrenia incidence[157]
MSMO1	cholesterol biosynthesis	Up	Mutations in cholesterol synthesis regulatory genes associated with increased schizophrenia incidence[157]
INPP5E	inositol phosphate metabolism	Up	Disturbances in regulation of phosphoinositide signaling system proposed as marker of schizophrenia[158,159]
PIK3C2G	inositol phosphate metabolism	Down	Disturbances in regulation of phosphoinositide signaling system proposed as marker of schizophrenia[158,159]
ENPP2	extracellular lysophosphatidic acid (LPA) synthesis	Up	LPA receptor-deficient mice proposed as model of schizophrenia[160]

Supplementary Table 9: Summary of drugs with significant antipsychotic signatures that are structurally similar (given a tanimoto coefficient based on their respective SMILES descriptors). The sum of the gene expression changes for the signature genes linked to the gene indication signature.

Drug 1	Drug 2	Antipsychotic signature (drug 1)	Antipsychotic signature (drug 2)	Tanimoto coefficient	Drug 1 indication	Drug 2 indication
ouabain	lanatoside_C	-15.347	-13.276	0.923	cardiac glycoside	cardiac glycoside
camptothecin	irinotecan	-15.351	-13.276	0.838	cytotoxic quinoline alkaloid	treat colon cancer and small cell lung cancer
isotretinoin	tretinoin	-15.427	-13.276	1	treat colon cancer and small cell lung cancer	treat acne
doxorubicin	daunorubicin	-15.796	-13.276	1	chemotherapy medication	chemotherapy medication
kinetin	6-benzylaminopurine	-16.523	-13.276	0.711	skin aging	skin aging
netilmicin	sisomicin	-17.005	-13.276	0.909	antibiotic	antibiotic

pergolide	terguride	-18.252	-13.276	0.715	Parkinson's disease, hyperprolactinemia, and restless leg syndrome	hyperprolactinemia
tetryzoline	xylometazoline	-29.404	-13.276	0.721	eye redness caused by certain allergies	nasal congestion, allergic rhinitis, and sinusitis

Supplementary References

1. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29,** 308–311 (2001).
2. Famiglietti, M. L. *et al.* Genetic variations and diseases in UniProtKB/Swiss-Prot: the ins and outs of expert manual curation. *Hum. Mutat.* **35,** 927–935 (2014).
3. Whirl-Carrillo, M. *et al.* Pharmacogenomics knowledge for personalized medicine. *Clin. Pharmacol. Ther.* **92,** 414–417 (2012).
4. Murray, R. K. *et al.* A Lange Medical Book: Harper's Illustrated Biochemistry. (2009).
5. Haemmerle, G. *et al.* Hormone-sensitive lipase deficiency in mice causes diglyceride accumulation in adipose tissue, muscle, and testis. *J. Biol. Chem.* **277,** 4806–4815 (2002).
6. Schmitz, G. & Ruebsaamen, K. Metabolism and atherogenic disease association of lysophosphatidylcholine. *Atherosclerosis* **208,** 10–18 (2010).
7. Abra, R. M. & Quinn, P. J. A novel pathway for phosphatidylcholine catabolism in rat brain homogenates. *Biochim. Biophys. Acta* **380,** 436–441 (1975).
8. Clarke, N. & Dawson, R. M. Enzymic formation of Glycerol 1:2-cyclic phosphate. *Biochem. J* **153,** 745–747 (1976).
9. Ferrer, J. *et al.* Mitochondrial glycerol-3-phosphate dehydrogenase. Cloning of an alternatively spliced human islet-cell cDNA, tissue distribution, physical mapping, and identification of a polymorphic genetic marker. *Diabetes* **45,** 262–266 (1996).
10. Ueda, N., Tsuboi, K., Uyama, T. & Ohnishi, T. Biosynthesis and degradation of the endocannabinoid 2-arachidonoylglycerol. *Biofactors* **37,** 1–7 (2011).
11. Wang, J. & Ueda, N. Biology of endocannabinoid synthesis system. *Prostaglandins Other Lipid Mediat.* **89,** 112–119 (2009).
12. Costa, C. C. *et al.* 3-, 6- and 7-hydroxyoctanoic acids are metabolites of medium-chain triglycerides and excreted in urine as glucuronides. *J. Mass Spectrom.* **31,** 633–638 (1996).
13. Chen, R. *et al.* Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* **148,** 1293–1307 (2012).
14. Goodman, B. E. Insights into digestion and absorption of major nutrients in humans. *Adv. Physiol. Educ.* **34,** 44–53 (2010).
15. Claus, S. P. *et al.* Systemic multicompartmental effects of the gut microbiome on mouse metabolic phenotypes. *Mol. Syst. Biol.* **4,** 219 (2008).
16. Delaney, J. *et al.* Phenylacetylglycine, a putative biomarker of phospholipidosis: its origins and relevance to phospholipid accumulation using

amiodarone treated rats as a model. *Biomarkers* **9,** 271–290 (2004).

17. Meiser, J., Weindl, D. & Hiller, K. Complexity of dopamine metabolism. *Cell Commun. Signal.* **11,** 34 (2013).
18. Tender, E. U. Evaluation of population newborn screening practices for rare disorders in Member States of the European Union. (2012).
19. Abdenur, J. E. *et al.* Aromatic l-aminoacid decarboxylase deficiency: unusual neonatal presentation and additional findings in organic acid analysis. *Mol. Genet. Metab.* **87,** 48–53 (2006).
20. Lindblad, B., Lindstedt, S. & Steen, G. On the enzymic defects in hereditary tyrosinemia. *Proc. Natl. Acad. Sci. U. S. A.* **74,** 4641–4645 (1977).
21. Rolfsson, O., Palsson, B. Ø. & Thiele, I. The human metabolic reconstruction Recon 1 directs hypotheses of novel human metabolic functions. *BMC Syst. Biol.* **5,** 155 (2011).
22. Chalmers, R. *Organic Acids in Man: Analytical Chemistry, Biochemistry and Diagnosis of the Organic Acidurias*. (Springer Netherlands, 2012).
23. Rinaldo, P., O'Shea, J. J., Welch, R. D. & Tanaka, K. Stable isotope dilution analysis of n-hexanoylglycine, 3-phenylpropionylglycine and suberylglycine in human urine using chemical ionization gas chromatography/mass spectrometry selected ion monitoring. *Biomed. Environ. Mass Spectrom.* **18,** 471–477 (1989).
24. Wishart, D. S. *et al.* HMDB: the Human Metabolome Database. *Nucleic Acids Res.* **35,** D521–6 (2007).
25. Shin, S.-Y. *et al.* An atlas of genetic influences on human blood metabolites. *Nat. Genet.* **46,** 543–550 (2014).
26. Illig, T. *et al.* A genome-wide perspective of genetic variation in human metabolism. *Nat. Genet.* **42,** 137–141 (2010).
27. Holle, R., Happich, M., Löwel, H., Wichmann, H. E. & MONICA/KORA Study Group. KORA--a research platform for population based health research. *Gesundheitswesen* **67 Suppl 1,** S19–25 (2005).
28. Mittelstrass, K. *et al.* Discovery of sexual dimorphisms in metabolic and genetic biomarkers. *PLoS Genet.* **7,** e1002215 (2011).
29. Burgard, P. *et al.* Report on the practices of newborn screening for rare disorders implemented in Member States of the European Union, Candidate, Potential Candidate and EFTA Countries. *EU Tender "Evaluation of population newborn screening practices for rare disorders in Member States of the European Union* 53–57 (2011).
30. Engelke, U. F. H., Oostendorp, M. & Wevers, R. A. NMR Spectroscopy of Body Fluids as a Metabolomics Approach to Inborn Errors of Metabolism. in *The Handbook of Metabonomics and Metabolomics* 375–412 (Elsevier, 2007).
31. Smilowitz, J. T. *et al.* The human milk metabolome reveals diverse oligosaccharide profiles. *J. Nutr.* **143,** 1709–1718 (2013).
32. Jain, M. *et al.* Metabolite profiling identifies a key role for glycine in rapid cancer cell proliferation. *Science* **336,** 1040–1044 (2012).
33. Gille, C. *et al.* HepatoNet1: a comprehensive metabolic reconstruction of the human hepatocyte for the analysis of liver physiology. *Mol. Syst. Biol.* **6,** (2010).
34. Duarte, N. C. *et al.* Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proceedings of the National Academy of Sciences* **104,** 1777–1782 (2007).
35. Ma, H. *et al.* The Edinburgh human metabolic network reconstruction and its functional analysis. *Mol. Syst. Biol.* **3,** 135 (2007).

36. Thiele, I. *et al.* A community-driven global reconstruction of human metabolism. *Nat. Biotechnol.* **31,** 419–425 (2013).

37. Maglott, D., Ostell, J., Pruitt, K. D. & Tatusova, T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.* **33,** D54–8 (2005).

38. Hastings, J. *et al.* The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res.* **41,** D456–63 (2013).

39. Kim, S. *et al.* PubChem Substance and Compound databases. *Nucleic Acids Res.* **44,** D1202–13 (2016).

40. Sahoo, S., Aurich, M. K., Jonsson, J. J. & Thiele, I. Membrane transporters in a human genome-scale metabolic knowledgebase and their implications for disease. *Front. Physiol.* **5,** 91 (2014).

41. Partha, R. & Raman, K. Revisiting robustness and evolvability: evolution in weighted genotype spaces. *PLoS One* **9,** e112792 (2014).

42. Ishidoh, K. *et al.* Quinolinate phosphoribosyl transferase, a key enzyme in de novo NAD(+) synthesis, suppresses spontaneous cell death by inhibiting overproduction of active-caspase-3. *Biochim. Biophys. Acta* **1803,** 527–533 (2010).

43. Bordbar, A., Jamshidi, N. & Palsson, B. O. iAB-RBC-283: A proteomically derived knowledge-base of erythrocyte metabolism that can be used to simulate its physiological and patho-physiological states. *BMC Syst. Biol.* **5,** 1–12 (2011).

44. Bordbar, A. *et al.* A multi-tissue type genome-scale metabolic network for analysis of whole-body systems physiology. *BMC Syst. Biol.* **5,** 180 (2011).

45. Summers, L. K. *et al.* Uptake of individual fatty acids into adipose tissue in relation to their presence in the diet. *Am. J. Clin. Nutr.* **71,** 1470–1477 (2000).

46. Lanham-New, S. A., Macdonald, I. A. & Roche, H. M. *Nutrition and Metabolism*. (John Wiley & Sons, 2011).

47. Kodicek, E. Storage of Vitamins in Liver. *Proc. Nutr. Soc.* **13,** 125–135 (1954).

48. Koutsari, C., Ali, A. H., Mundi, M. S. & Jensen, M. D. Storage of circulating free fatty acid in adipose tissue of postabsorptive humans: quantitative measures and implications for body fat distribution. *Diabetes* **60,** 2032–2040 (2011).

49. Brady, S., Siegel, G., Wayne Albers, R. & Price, D. *Basic Neurochemistry: Molecular, Cellular and Medical Aspects*. (Academic Press, 2005).

50. Rucker, R. B., Zempleni, J., Suttie, J. W. & McCormick, D. B. *Handbook of Vitamins, Fourth Edition*. (CRC Press, 2007).

51. Gossell-Williams, M. *et al.* Dietary intake of choline and plasma choline concentrations in pregnant women in Jamaica. *West Indian Med. J.* **54,** 355–359 (2005).

52. Kinross, J. M., Darzi, A. W. & Nicholson, J. K. Gut microbiome-host interactions in health and disease. Genome Med 2011; 3: 14. *Copyright© 2012 Massachusetts Medical Society*

53. Heinken, A. & Thiele, I. Anoxic Conditions Promote Species-Specific Mutualism between Gut Microbes In Silico. *Appl. Environ. Microbiol.* **81,** 4049–4061 (2015).

54. Huang, S.-T. *et al.* Serum total p-cresol and indoxyl sulfate correlated with stage of chronic kidney disease in renal transplant recipients. *Transplant. Proc.* **44,** 621–624 (2012).

55. Doessegger, L. *et al.* Increased levels of urinary phenylacetylglycine associated with mitochondrial toxicity in a model of drug-induced phospholipidosis. *Ther Adv Drug Saf* **4,** 101–114 (2013).

56. Winchester, J. F., Hostetter, T. H. & Meyer, T. W. p-Cresol sulfate: further understanding of its cardiovascular disease potential in CKD. *Am. J. Kidney Dis.* **54,** 792–794 (2009).

57. Sahoo, S., Haraldsdóttir, H. S., Fleming, R. M. T. & Thiele, I. Modeling the effects of commonly used drugs on human metabolism. *FEBS J.* **282,** 297–317 (2015).

58. Sahoo, S. & Thiele, I. Predicting the impact of diet and enzymopathies on human small intestinal epithelial cells. *Hum. Mol. Genet.* **22,** 2705–2722 (2013).

59. Mori, T. A. & Woodman, R. J. The independent effects of eicosapentaenoic acid and docosahexaenoic acid on cardiovascular risk factors in humans. *Curr. Opin. Clin. Nutr. Metab. Care* **9,** 95–104 (2006).

60. Millán, J. *et al.* Lipoprotein ratios: Physiological significance and clinical usefulness in cardiovascular prevention. *Vasc. Health Risk Manag.* **5,** 757–765 (2009).

61. Alnouti, Y. Bile Acid sulfation: a pathway of bile acid elimination and detoxification. *Toxicol. Sci.* **108,** 225–246 (2009).

62. Perreault, M. *et al.* Role of glucuronidation for hepatic detoxification and urinary elimination of toxic bile acids during biliary obstruction. *PLoS One* **8,** e80994 (2013).

63. Trottier, J. *et al.* Human UDP-glucuronosyltransferase (UGT)1A3 enzyme conjugates chenodeoxycholic acid in the liver. *Hepatology* **44,** 1158–1170 (2006).

64. Court, M. H. *et al.* Quantitative distribution of mRNAs encoding the 19 human UDP-glucuronosyltransferase enzymes in 26 adult and 3 fetal tissues. *Xenobiotica* **42,** 266–277 (2012).

65. Perreault, M. *et al.* The Human UDP-glucuronosyltransferase UGT2A1 and UGT2A2 enzymes are highly active in bile acid glucuronidation. *Drug Metab. Dispos.* **41,** 1616–1620 (2013).

66. Dawson, P. A. & Karpen, S. J. Intestinal transport and metabolism of bile acids. *J. Lipid Res.* **56,** 1085–1099 (2015).

67. Araya, Z. & Wikvall, K. 6alpha-hydroxylation of taurochenodeoxycholic acid and lithocholic acid by CYP3A4 in human liver microsomes. *Biochim. Biophys. Acta* **1438,** 47–54 (1999).

68. Bodin, K., Lindbom, U. & Diczfalusy, U. Novel pathways of bile acid metabolism involving CYP3A4. *Biochim. Biophys. Acta* **1687,** 84–93 (2005).

69. Gudmundsson, S. & Thiele, I. Computationally efficient flux variability analysis. *BMC Bioinformatics* **11,** 489 (2010).

70. Kunnen, S. & Van Eck, M. Lecithin:cholesterol acyltransferase: old friend or foe in atherosclerosis? *J. Lipid Res.* **53,** 1783–1799 (2012).

71. Andersson, L., Sternby, B. & Nilsson, A. Hydrolysis of phosphatidylethanolamine by human pancreatic phospholipase A2. Effect of bile salts. *Scand. J. Gastroenterol.* **29,** 182–187 (1994).

72. Quek, L.-E. *et al.* Reducing Recon 2 for steady-state flux analysis of HEK cell culture. *J. Biotechnol.* **184,** 172–178 (2014).

73. Heller, S., McNaught, A., Stein, S., Tchekhovskoi, D. & Pletnev, I. InChI - the worldwide chemical structure identifier standard. *J. Cheminform.* **5,** 7 (2013).

74. Anderson, E., Veith, G. D., Weininger, D. & Environmental Research Laboratory (Duluth, M.). *SMILES, a Line Notation and Computerized Interpreter for Chemical Structures*. (U.S. Environmental Protection Agency, Environmental Research Laboratory, 1987).

75. Thiele, I. & Palsson, B. Ø. Reconstruction annotation jamborees: a community approach to systems biology. *Mol. Syst. Biol.* **6,** 361 (2010).

76. Hao, T., Ma, H.-W., Zhao, X.-M. & Goryanin, I. Compartmentalization of the Edinburgh human metabolic network. *BMC Bioinformatics* **11,** 393 (2010).

77. UniProt Consortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* **40,** D71–5 (2012).

78. Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh, A. OMIM. org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* **43,** D789–D798 (2015).

79. Thiele, I. & Palsson, B. Ø. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat. Protoc.* **5,** 93–121 (2010).

80. Heirendt L, Arreckx S, Pfau T, Mendoza SN, Richelle A, Heinken A, Haraldsdottir HS, Keating SM, Vlasov V, Wachowiak J, et al.: Creation and analysis of biochemical constraint-based models: the COBRA Toolbox v3.0. arXiv:1710.04038 [q-bio.QM].

81. Pornputtapong, N., Nookaew, I. & Nielsen, J. Human metabolic atlas: an online resource for human metabolism. *Database* **2015,** bav068 (2015).

82. Malhi, H., Irani, A. N., Gagandeep, S. & Gupta, S. Isolation of human progenitor liver epithelial cells with extensive replication capacity and differentiation into mature hepatocytes. *J. Cell Sci.* **115,** 2679–2688 (2002).

83. Senyo, S. E. *et al.* Mammalian heart renewal by pre-existing cardiomyocytes. *Nature* **493,** 433–436 (2013).

84. Li, Y. & Wingert, R. A. Regenerative medicine for the kidney: stem cell prospects & challenges. *Clin. Transl. Med.* **2,** 11 (2013).

85. van de Poll, M. C. G., Soeters, P. B., Deutz, N. E. P., Fearon, K. C. H. & Dejong, C. H. C. Renal metabolism of amino acids: its role in interorgan amino acid exchange. *Am. J. Clin. Nutr.* **79,** 185–197 (2004).

86. Potter, J. E., James, M. J. & Kandutsch, A. A. Sequential cycles of cholesterol and dolichol synthesis in mouse spleens during phenylhydrazine-induced erythropoiesis. *J. Biol. Chem.* **256,** 2371–2376 (1981).

87. Chen, T. C. *et al.* Factors that influence the cutaneous synthesis and dietary sources of vitamin D. *Arch. Biochem. Biophys.* **460,** 213–217 (2007).

88. Agren, R. *et al.* Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling. *Mol. Syst. Biol.* **10,** 721 (2014).

89. Swainston, N. *et al.* Recon 2.2: from reconstruction to model of human metabolism. *Metabolomics* **12,** 109 (2016).

90. Haraldsdóttir, H. S., Thiele, I. & Fleming, R. M. T. Quantitative assignment of reaction directionality in a multicompartmental human metabolic reconstruction. *Biophys. J.* **102,** 1703–1711 (2012).

91. Nilsson, A., Mardinoglu, A. & Nielsen, J. Predicting growth of the healthy infant using a genome scale metabolic model. *npj Systems Biology and Applications* **3,** 3 (2017).

92. Mardinoglu, A. *et al.* Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease. *Nat. Commun.* **5,** 3083 (2014).

93. Brunk, E. *et al.* Systems Biology of the Structural Proteome. *BMC Syst. Biol.* **just accepted,** (2016).

94. McCloskey, D., Palsson, B. Ø. & Feist, A. M. Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli. *Mol. Syst. Biol.* **9,** 661 (2013).

95. Pruitt, K. D. *et al.* RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.* **42,** D756–63 (2014).

96. Cokelaer, T., Pultz, D., Harder, L. M., Serra-Musach, J. & Saez-Rodriguez, J. BioServices: a common Python package to access biological Web Services programmatically. *Bioinformatics* **29,** 3241–3242 (2013).

97. Kinsella, R. J. *et al.* Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database* **2011,** bar030 (2011).

98. Xu, D. & Zhang, Y. Ab Initio structure prediction for Escherichia coli: towards genome-wide protein structure modeling and fold assignment. *Sci. Rep.* **3,** (2013).

99. Zhou, H., Gao, M., Kumar, N. & Skolnick, J. SUNPRO: Structure and function predictions of proteins from representative organisms. (2012).

100. Roy, A., Kucukural, A. & Zhang, Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.* **5,** 725–738 (2010).

101. Hamelryck, T. & Manderick, B. PDB file parser and structure class implemented in Python. *Bioinformatics* **19,** 2308–2310 (2003).

102. Zhang, Y. I-TASSER: Fully automated protein structure prediction in CASP8. *Proteins: Struct. Funct. Bioinf.* **77,** 100–113 (2009).

103. Jaroszewski, L., Pawlowski, K. & Godzik, A. Multiple Model Approach: Exploring the Limits of Comparative Modeling. *J. Mol. Med.* **4,** 294–309 (1998).

104. Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **26,** 283–291 (1993).

105. Rose, A. S. & Hildebrand, P. W. NGL Viewer: a web application for molecular visualization. *Nucleic Acids Res.* **43,** W576–9 (2015).

106. Ye, Y. & Godzik, A. Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics* **19 Suppl 2,** ii246–55 (2003).

107. Alexandrov, N. & Shindyalov, I. PDP: protein domain parser. *Bioinformatics* **19,** 429–430 (2003).

108. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44,** D457–62 (2016).

109. Sud, M. *et al.* LMSD: LIPID MAPS structure database. *Nucleic Acids Res.* **35,** D527–32 (2007).

110. Forster, M., Pick, A., Raitner, M., Schreiber, F. & Brandenburg, F. J. The system architecture of the BioPath system. *In Silico Biol.* **2,** 415–426 (2002).

111. Williams, A. J., Tkachenko, V., Golotvin, S., Kidd, R. & McCann, G. ChemSpider-building a foundation for the semantic web by hosting a crowd sourced databasing platform for chemistry. *J. Cheminform.* **2,** 1–1 (2010).

112. Haraldsdóttir, H. S. & Fleming, R. M. T. Identification of Conserved Moieties in Metabolic Networks by Graph Theoretical Analysis of Atom Transition Networks. *PLoS Comput. Biol.* **12,** e1004999 (2016).

113. Famili, I. & Palsson, B. O. The convex basis of the left null space of the stoichiometric matrix leads to the definition of metabolically meaningful pools. *Biophys. J.* **85,** 16–26 (2003).

114. Wiechert, W. 13C metabolic flux analysis. *Metab. Eng.* **3,** 195–206 (2001).

115. Rahman, S. A. *et al.* Reaction Decoder Tool (RDT): extracting features from chemical reactions. *Bioinformatics* **32,** 2065–2066 (2016).

116. First, E. L., Gounaris, C. E. & Floudas, C. A. Stereochemically consistent reaction mapping and identification of multiple reaction mechanisms through integer linear optimization. *J. Chem. Inf. Model.* **52,** 84–92 (2012).

117. Kumar, A. & Maranas, C. D. CLCA: maximum common molecular substructure queries within the MetRxn database. *J. Chem. Inf. Model.* **54,** 3417–3438 (2014).

118. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* **Chapter 7,** Unit7.20 (2013).

119. Noronha, A. *et al.* ReconMap: an interactive visualization of human metabolism. *Bioinformatics* **33,** 605–607 (2017).

120. Matsuoka, Y., Funahashi, A., Ghosh, S. & Kitano, H. Modeling and simulation using CellDesigner. *Methods Mol. Biol.* **1164,** 121–145 (2014).

121. Hucka, M. *et al.* Systems Biology Markup Language (SBML) Level 2: Structures and Facilities for Model Definitions. (2008). doi:10.1038/npre.2008.2715.1

122. Hucka, M. *et al.* The Systems Biology Markup Language (SBML): Language Specification for Level 3 Version 1 Core. *Nature Precedings* (2010). doi:10.1038/npre.2010.4959

123. SBML Level 3 Version 1 Core Release 2 | COMBINE. Available at: http://identifiers.org/combine.specifications/sbml.level-3.version-1.core.release-2. (Accessed: 24th May 2017)

124. Gauges, R., Rost, U., Sahle, S., Wengler, K. & Bergmann, F. T. The Systems Biology Markup Language (SBML) Level 3 Package: Layout, Version 1 Core. *J. Integr. Bioinform.* **12,** 267 (2015).

125. Rodriguez, N. *et al.* JSBML 1.0: providing a smorgasbord of options to encode systems biology models. *Bioinformatics* **31,** 3383–3386 (2015).

126. Dräger, A. *et al.* JSBML: a flexible Java library for working with SBML. *Bioinformatics* **27,** 2167–2168 (2011).

127. Le Novère, N. *et al.* Minimum information requested in the annotation of biochemical models (MIRIAM). *Nat. Biotechnol.* **23,** 1509–1515 (2005).

128. Wimalaratne, S. M. *et al.* SPARQL-enabled identifier conversion with Identifiers.org. *Bioinformatics* **31,** 1875–1877 (2015).

129. Courtot, M. *et al.* Controlled vocabularies and semantics in systems biology. *Mol. Syst. Biol.* **7,** 543 (2011).

130. King, Z. A. *et al.* BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Res.* **44,** D515–22 (2016).

131. Reed, J. L., Famili, I., Thiele, I. & Palsson, B. O. Towards multidimensional genome annotation. *Nat. Rev. Genet.* **7,** 130–141 (2006).

132. Römer, M. *et al.* ZBIT Bioinformatics Toolbox: A Web-Platform for Systems Biology and Expression Data Analysis. *PLoS One* **11,** e0149263 (2016).

133. King, Z. A. *et al.* Escher: A Web Application for Building, Sharing, and Embedding Data-Rich Visualizations of Biological Pathways. *PLoS Comput. Biol.* **11,** e1004321 (2015).

134. Moodie, S., Le Novère, N., Demir, E., Mi, H. & Villéger, A. Systems Biology Graphical Notation: Process Description language Level 1 Version 1.3. *J. Integr. Bioinform.* **12,** 263 (2015).

135. Thomas, A., Rahmanian, S., Bordbar, A., Palsson, B. Ø. & Jamshidi, N. Network reconstruction of platelet metabolism identifies metabolic signature for aspirin resistance. *Sci. Rep.* **4,** 3925 (2014).

136. Lahiry, P., Torkamani, A., Schork, N. J. & Hegele, R. A. Kinase mutations in human disease: interpreting genotype-phenotype relationships. *Nat. Rev. Genet.* **11,** 60–74 (2010).

137. Robinson, D. R., Wu, Y. M. & Lin, S. F. The protein tyrosine kinase family of the human genome. *Oncogene* **19,** 5548–5557 (2000).

138. Paul, M. K. & Mukhopadhyay, A. K. Tyrosine kinase - Role and significance in Cancer. *Int. J. Med. Sci.* **1,** 101–115 (2004).

139. Majerus, P. W. & York, J. D. Phosphoinositide phosphatases and disease. *J. Lipid Res.* **50 Suppl,** S249–54 (2009).

140. Andersen, J. N. *et al.* A genomic perspective on protein tyrosine phosphatases: gene structure, pseudogenes, and genetic disease linkage. *FASEB J.* **18,** 8–30 (2004).

141. Pedersen, P. L. Transport ATPases into the year 2008: a brief overview related to types, structures, functions and roles in health and disease. *J. Bioenerg. Biomembr.* **39,** 349–355 (2007).

142. Page, M. J. & Di Cera, E. Serine peptidases: classification, structure and function. *Cell. Mol. Life Sci.* **65,** 1220–1236 (2008).

143. Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489,** 519–525 (2012).

144. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499,** 214–218 (2013).

145. Brennan, C. W. *et al.* The somatic genomic landscape of glioblastoma. *Cell* **155,** 462–477 (2013).

146. Villa, G. R. *et al.* An LXR-Cholesterol Axis Creates a Metabolic Co-Dependency for Brain Cancers. *Cancer Cell* (2016).

147. Gatto, F., Miess, H., Schulze, A. & Nielsen, J. Flux balance analysis predicts essential genes in clear cell renal cell carcinoma metabolism. *Sci. Rep.* **5,** 10738 (2015).

148. Gatto, F. & Nielsen, J. Pan-cancer analysis of the metabolic reaction network. *bioRxiv* 050187 (2016). doi:10.1101/050187

149. Agren, R. *et al.* The RAVEN toolbox and its use for generating a genome-scale metabolic model for Penicillium chrysogenum. *PLoS Comput.*

Biol. **9,** e1002980 (2013).

150. Zielinski, D. C. *et al.* Pharmacogenomic and clinical data link non-pharmacokinetic metabolic dysregulation to drug side effect pathogenesis. *Nat. Commun.* **6,** 7101 (2015).
151. Lamb, J. *et al.* The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313,** 1929–1935 (2006).
152. Kuhn, M., Letunic, I., Jensen, L. J. & Bork, P. The SIDER database of drugs and side effects. *Nucleic Acids Res.* (2015). doi:10.1093/nar/gkv1075
153. Taurines, R. *et al.* Expression analyses of the mitochondrial complex I 75-kDa subunit in early onset schizophrenia and autism spectrum disorder: increased levels as a potential biomarker for early onset schizophrenia. *Eur. Child Adolesc. Psychiatry* **19,** 441–448 (2010).
154. Stanta, J. L. *et al.* Identification of N-glycosylation changes in the CSF and serum in patients with schizophrenia. *J. Proteome Res.* **9,** 4476–4489 (2010).
155. Bauer, D., Haroutunian, V., Meador-Woodruff, J. H. & McCullumsmith, R. E. Abnormal glycosylation of EAAT1 and EAAT2 in prefrontal cortex of elderly patients with schizophrenia. *Schizophr. Res.* **117,** 92–98 (2010).
156. Assies, J. *et al.* Significantly reduced docosahexaenoic and docosapentaenoic acid concentrations in erythrocyte membranes from schizophrenic patients compared with a carefully matched control group. *Biol. Psychiatry* **49,** 510–522 (2001).
157. Le Hellard, S. *et al.* Polymorphisms in SREBF1 and SREBF2, two antipsychotic-activated transcription factors controlling cellular lipogenesis, are associated with schizophrenia in German and Scandinavian samples. *Mol. Psychiatry* **15,** 463–472 (2010).
158. Kaiya, H., Nishida, A., Imai, A., Nakashima, S. & Nozawa, Y. Accumulation of diacylglycerol in platelet phosphoinositide turnover in schizophrenia: a biological marker of good prognosis? *Biol. Psychiatry* **26,** 669–676 (1989).
159. Strunecká, A. & Rípová, D. What can the investigation of phosphoinositide signaling system in platelets of schizophrenic patients tell us? *Prostaglandins Leukot. Essent. Fatty Acids* **61,** 1–5 (1999).
160. Harrison, S. M. *et al.* LPA1 receptor-deficient mice have phenotypic changes observed in psychiatric disease. *Mol. Cell. Neurosci.* **24,** 1170–1179 (2003).