Supplementary information:
Thiele *et al*., "A community-driven global reconstruction of human metabolism".

Supplementary Note 1: Comparison of information content in five metabolic resources

Information for reaction and gene comparisons was obtained for the following five human metabolism resources: Reactome[2], HumanCyc[3], KEGG[4], the compartmentalized Edinburgh Human Metabolic Network (compEHMN)[5], and Recon 1[6] (see Supplementary Table 8). For this analysis, 406 exchange reactions were removed from Recon 1 along with 38 reaction duplicates that differed only in their tissue annotation[7]. From KEGG, the pathways of the category 'Metabolism' were selected. KEGG is not a human specific database and therefore, only those reactions that are linked to one or more human genes were selected. From HumanCyc, the two signalling pathways (the 'BMP Signalling Pathway' and the 'MAP kinase cascade') were excluded. From Reactome, ten human pathways focused on (normal) metabolic processes were selected, excluding, for example, signalling and disease-related pathways. All reactions assigned to the selected metabolic pathways were retrieved. Black box events from Reactome, representing reactions for which the molecular details are unspecified or unknown, were ignored if either the input or output was missing. Out-of-date gene and metabolite identifiers were, where possible, transferred to their current identifier or otherwise not taken into account in the comparison. Syntactically incorrect metabolite identifiers were corrected manually.

Database	Format	Version	Source
Recon 1	SBML	1	http://bigg.ucsd.edu/
compEHMN	Excel	2	http://www.ehmn.bioinformatics.ed.ac.uk/
HumanCyc	Text	15.0	http://biocyc.org/download.shtml
KEGG	KGML	58	ftp://ftp.genome.jp/pub/kegg/
Reactome	MySQL database	36	http://reactome.org/download/index.html

Supplementary Table 8: Versions of the databases used in metabolic resource comparisons. Pathway Tools[8] was used to export the content of HumanCyc into textual flat files. All data were downloaded in May 2011.

Entrez Gene identifiers were used to match genes between databases, since it is the only gene identifier provided by all five databases. In HumanCyc, the Entrez Gene identifier is missing for 605 genes. For 282 of these genes, the Entrez Gene identifier could be retrieved via either the Ensembl Gene identifier (181 genes) or the UniProt identifier (101 genes). The remaining 323 genes were excluded from the comparison, as they could not be linked to an Entrez Gene id. For 82% of these, none of the three identifiers were available. The multiple transcript variants in Recon 1 were flattened into a single Entrez Gene identifier. For those databases containing complexes (Recon 1, HumanCyc, and Reactome), the components (genes) of each complex were considered separately in the comparison. Only genes linked to a reaction and/or EC number were included in the comparison.

For the reaction comparison, where possible, the KEGG Compound identifier was used to match metabolites. If the KEGG Compound identifier was missing, metabolites had to match on any of the other metabolite identifiers (KEGG Glycan, ChEBI[9], PubChem Compound[10] or CAS (http://www.cas.org)) or on name, in which case also the chemical formula was required to match. Reactions were considered to

Nature Biotechnology: doi:10.1038/nbt.2488

be the same if all substrates and products (excluding water, electrons and protons) matched, irrespective of compartmentalization.

The results of the comparison (see Supplementary Tables 9 and 10) indicated that five resources agreed only on a relatively small set of enzyme-encoding genes (510) and metabolic reactions (199). These differences indicated that extensive evaluation and manual curation would be required to build a consensus network for human metabolism with uniform content and representation. Additional differences exist in the annotation of the contents among those reconstructions, such as insufficient use of identifiers and the genetic basis for the reactions included[7].

	Total	Unique	1 other database	2 other databases	3 other databases	4 other databases
Recon 1	1490	45 (3%)	90 (6%)	221 (15%)	624 (42%)	510 (34%)
EHMN	2492	128 (5%)	868 (35%)	355 (14%)	631 (25%)	510 (20%)
HumanCyc	3209	759 (24%)	954 (30%)	337 (11%)	649 (20%)	510 (16%)
KEGG	1535	63 (4%)	138 (9%)	272 (18%)	552 (36%)	510 (33%)
Reactome	1180	144 (12%)	116 (10%)	174 (15%)	236 (20%)	510 (43%)

Supplementary Table 9: Comparison of enzyme-encoding genes present in five resources describing human metabolism.

	Total	Unique	1 other databases	2 other databases	3 other databases	4 other databases
Recon 1	2549	1250 (49%)	485 (19%)	346 (14%)	269 (11%)	199 (8%)
EHMN	3695	1832 (50%)	913 (25%)	461 (12%)	290 (8%)	199 (5%)
HumanCyc	1761	905 (51%)	203 (12%)	197 (11%)	257 (15%)	199 (11%)
KEGG	1622	348 (21%)	479 (30%)	357 (22%)	239 (15%)	199 (12%)
Reactome	1131	539 (48%)	100 (9%)	172 (15%)	121 (11%)	199 (18%)

Supplementary Table 10: Comparison of metabolites present in five resources describing human metabolism.

Supplementary Note 2: Biomass composition

A biomass reaction was added to Recon 2 to account for the macromolecular synthesis requirement for proteins, DNA, RNA, lipids, and carbohydrates. The biomass composition was determined as described by Thiele and Palsson[11]. The information was assembled from literature and tailored towards leukaemia cell lines where specific data were available.

The following macromolecular composition was assumed: protein 70.6%, DNA 1.4%, RNA 5.8%, carbohydrates 7.1%, lipids 9.7%, others 5.4%[12, 13]. The main carbohydrate was assumed to be glucose-6-phosphate. 'Others' includes vitamins, small molecules, ions, cofactors, etc. As no further information was found to the compounds included in 'others', the biomass reaction does not capture their fractional contribution. The cellular weight was assumed as 500×10^{-12} g per cell. The fractional contribution of the different metabolites to the macromolecules was estimated using the nucleotide and amino acid sequence of the ~20,000 annotated ORFs. The fractional contribution of each amino acid was assumed to be L-alanine 7.6%, L-arginine 5.4%, L-asparagine 4.2%, L-aspartate 5.3%, L-cysteine 0.7%, L-glutamine 4.9%, L-glutamate 5.8%, L-glycine 8.1%, L-histidine 1.9%, L-isoleucine 4.3%, L-leucine 8.2%, L-lysine 8.9%, L-methionine 2.3%, L-phenylalanine 3.9%, L-proline 6.2%, L-serine 5.9%, L-threonine 4.7%, L-tryptophan 0.2%, L-tyrosine 2.4%, and L-

Nature Biotechnology: doi:10.1038/nbt.2488

valine 5.3%. The fraction contribution of each nucleotide triphosphate to RNA was assumed to be ATP 29.4%, CTP 21.4%, GTP 19.8%, and UPT 29.3%. The fractional contribution of each deoxynucleotide triphosphate to DNA was assumed to be dATP 28.9%, dCTP 20.7%, dGTP 21.7%, and dTTP 28.7%. The data for lipid contribution was obtained from Sheikh et al.[14]: cholestorol 7%, cardiolipin 4%, phosphatidylinositol 8%, phosphatidylcholine 53%, phosphatidylethanolamine 19%, phosphatidylglycerol 1%, sphingomyelin 6%, and phosphatidylserine 2%. The following energy requirements for translational processes were assumed: 4 ATP per peptidyl bond and an average protein length of 333 amino acids. This biomass reaction is a coarse grained approximation of metabolic precursors required and was tailored for lymphocytes when data was available.

Supplementary Note 3: Identification of missing functions in human metabolism

The identification of missing functions was carried out for two dead-end metabolites (1-pyrroline-2-carboxylate and 5-amino-2-oxopentanoate) as described previously[15]. Briefly, the associated blocked reactions were computed using FVA analysis[16]. We then used the Smiley algorithm[17] implemented in the COBRA Toolbox[18] to identify non-organism specific reactions from the KEGG[4] database that could restore flux through the blocked reactions by connecting causative dead-end metabolites to the network. If no flux restoring reactions were identified from KEGG, an artificial transport reaction was added for the causative dead-end metabolite. 20 solutions, comprised of as few reactions as possible, were calculated for each blocked reaction. We reviewed the literature in order to verify the biological relevance of the proposed solution and to generate a plausible hypothesis for gap filling.

Our example concerns the identification of candidate reactions for the dead-end metabolites 1-pyrroline-2-carboxylate and 5-amino-2-oxopentanoate. These metabolites are produced by the peroxisomal D-proline and D-ornithine oxidases (EC 1.4.3.3)[19, 20], respectively (Fig. A). These two metabolites exist in equilibrium with each other through a hydration reaction[21, 22]. The suggested resolving reaction involved conversion of 1-pyrroline-2-carboxylate to L-proline by pyrroline-2-carboxylate reductase (EC 1.5.1.1). This reaction has been observed in various mammals including rat and monkey[22] and, although the cellular localization is unknown, there is evidence that the conversion could be non-enzymatic[22, 23]. L-proline could then be converted to 1-pyrroline-5-carboxylate by L-pipecolate oxidase (EC 1.5.3.7), which has been shown to be localized in the peroxisome and to have broad substrate specificity[24]. It is possible that this reaction could also be catalyzed by L-proline oxidase (EC 1.5.1.2/EC 1.5.99.8); however, this enzyme has thus far only been localized in the cytoplasm and inner mitochondrial membrane[25]. Transport of 1-pyrroline-5-carboxylate out of the peroxisome would then allow 1-pyrroline-5-carboxylate to enter the tricarboxylic acid cycle or urea cycle, in which it plays an anaplerotic role as its tautomeric equivalent L-glutamate-5-semialdehyde[25]. While it is important to state that the proposed solution to the two blocked reactions is a hypothesis, the results highlight the benefits of a systems approach towards network refinement, without which the potential existence of this reaction pathway would not be inducible.

Supplementary Note 4: Basic topological properties of Recon 2

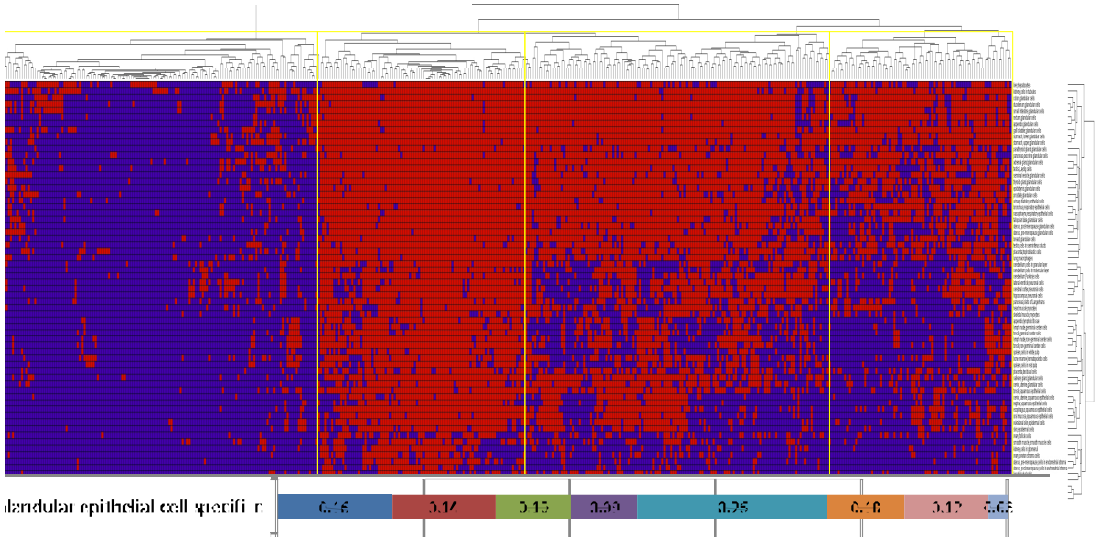
To assess basic topological features of Recon 2, the compound participation for each metabolite was calculated. (Compound participation being the count of how many

Nature Biotechnology: doi:10.1038/nbt.2488

reactions a given metabolite participates.) Similar trends were observed for both Recon 2 and Recon 1 (Fig. B, top). The most connected metabolites in Recon 2 are H+ (cytosol) (1210 reactions), water (cytosol) (868), Na+ (cytosol) (483), Na+ (extracellular) (479), H+ (mitochondrion) (448), and ATP (cytosol) (380). Additionally, reaction length (number of reactants per reaction) was also calculated (Fig. B, bottom). Very few reactions have many participating reactants, with – as can be expected – the artificial biomass reaction containing the greatest number of reactants (41). The average number of reactants per reaction is 4.2±2.6, a result that is comparable with Recon 1.

Nature Biotechnology: doi:10.1038/nbt.2488

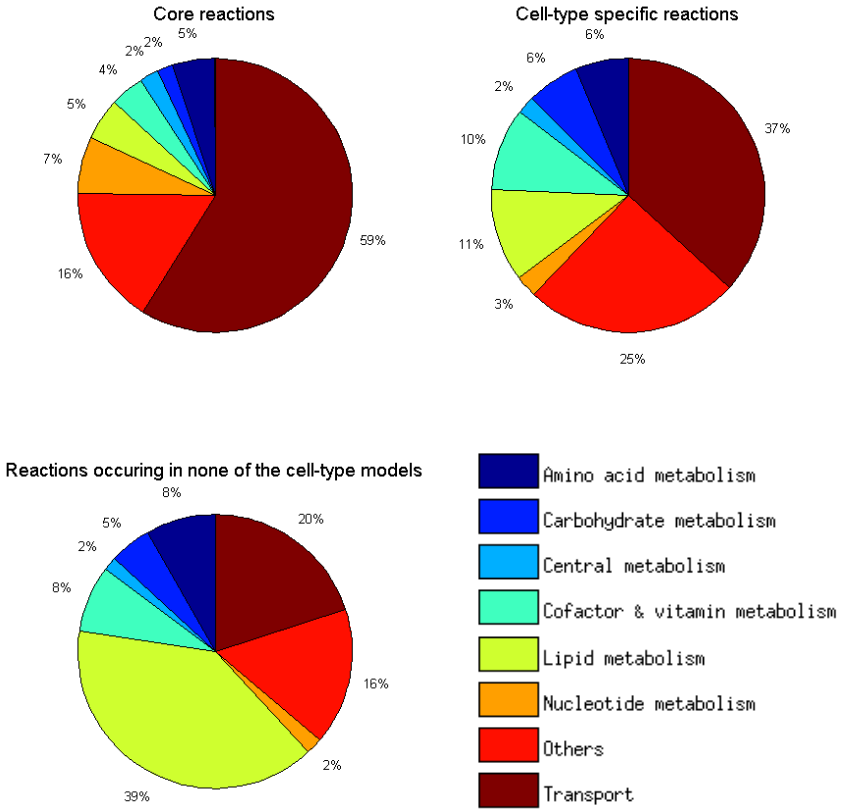
Supplementary Figure 1: Analysis of cell-type distribution of expressed proteins



Supplementary Figure 1: Protein expression data were obtained from the Human Protein Atlas for 65 cell types[26]. Top: Subsystem distribution of the proteins (protein numbers are indicated). Almost a quarter of these gene products belonged to Recon 2 categories of lipid metabolism. Middle: Heat map of hierarchical clustering of tissues and metabolic enzymes, with red and blue indicating presence and absence of enzymes, respectively. On average, there were 214±73 proteins expressed per cell-type, with the highest number of metabolic proteins (345) present in cells in kidney tubules and the lowest number (64) in endometrial stroma cells in post-menopausal uterus. Bottom: Subsystem distribution of enzymes in the four categories of proteins that vary in their expression in different cell types.

Nature Biotechnology: doi:10.1038/nbt.2488

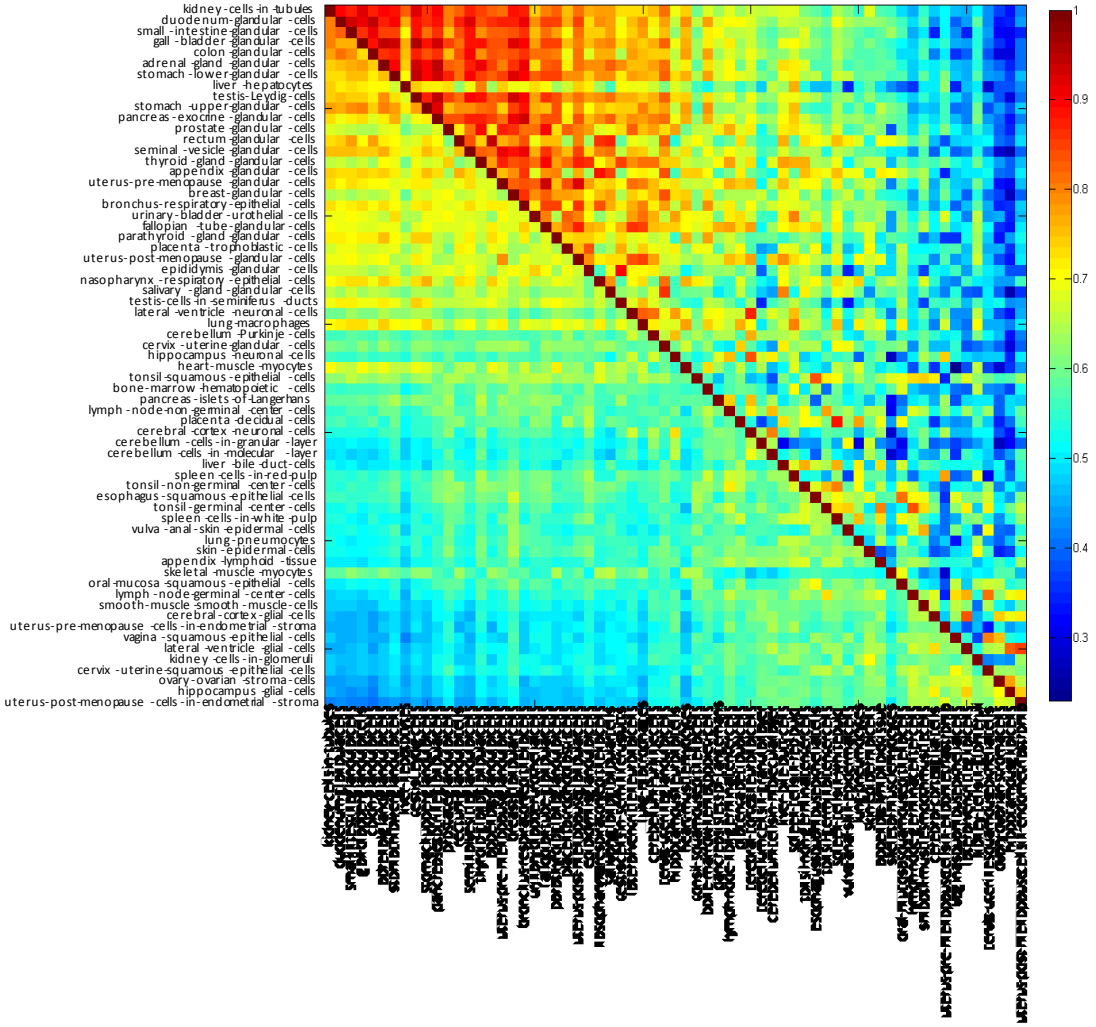
Supplementary Figure 2: Subsystem distribution of the different reaction categories



Supplementary Figure 2: Subsystem distribution of the different reaction categories.

Nature Biotechnology: doi:10.1038/nbt.2488

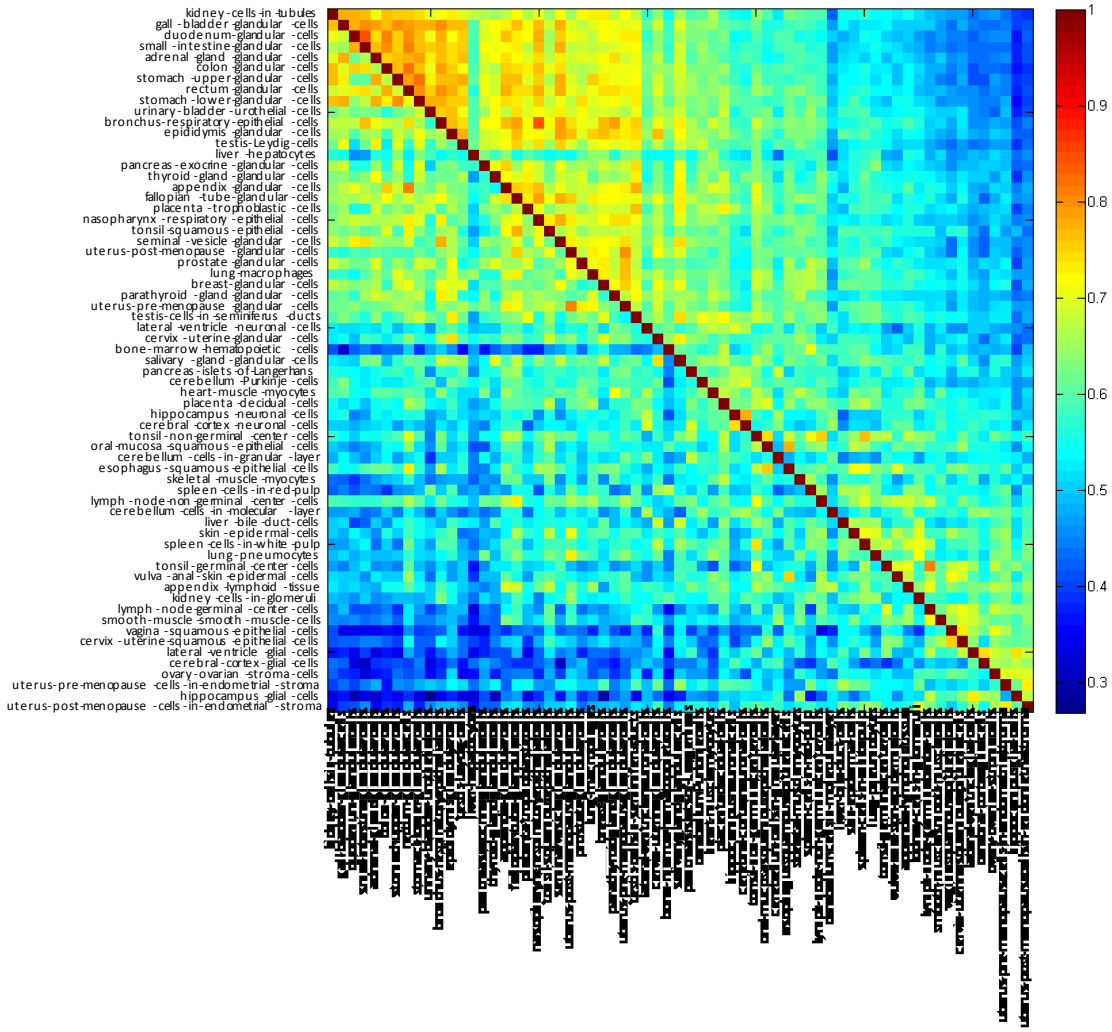
Supplementary Figure 3: Correlation between the 65 draft cell-type specific models



Supplementary Figure 3: Lower triangle: Based on reaction presence/absence. Upper triangle: Based on pathway presence/absence. Pearson correlation coefficients were calculated.

Nature Biotechnology: doi:10.1038/nbt.2488

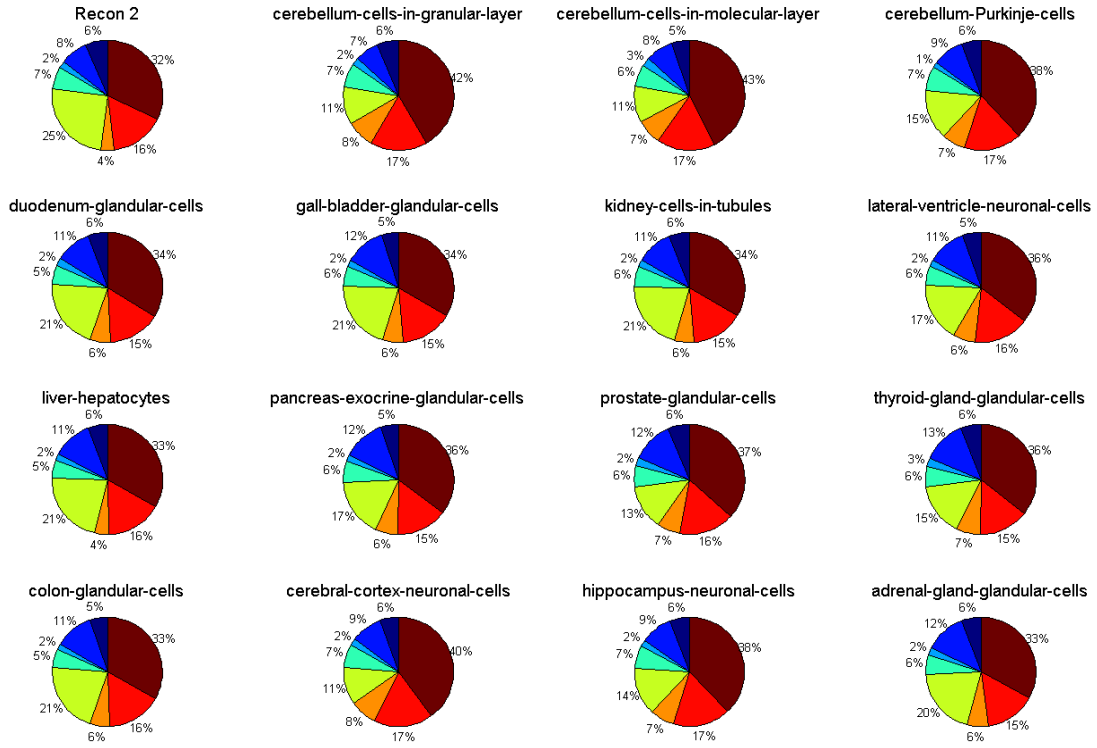
Supplementary Figure 4: Correlation between the 65 draft cell-type specific models



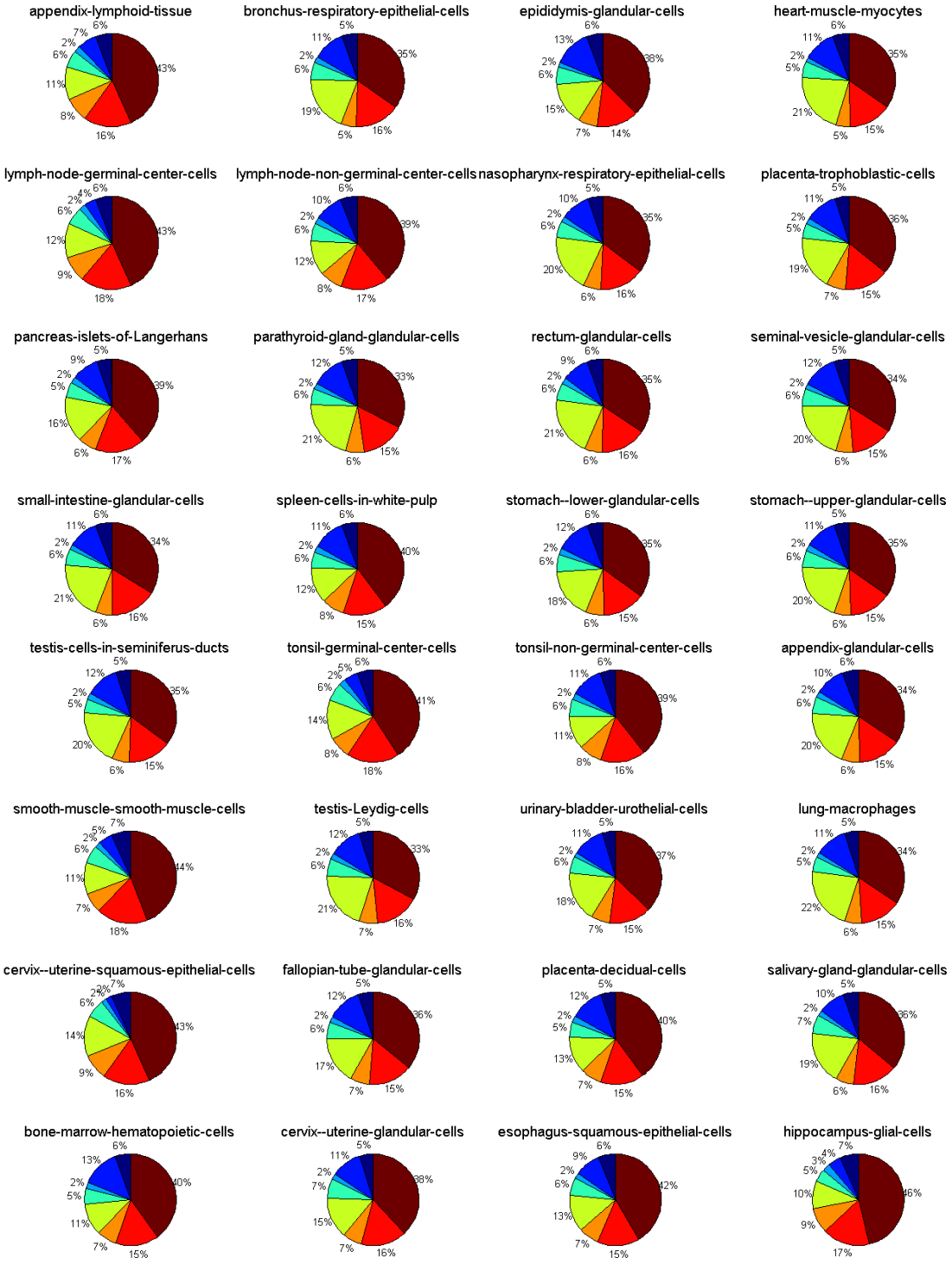
Supplementary Figure 4: Lower triangle: Based on gene presence/absence. Upper triangle: Based on presence/absence of inborn errors of metabolism. Pearson correlation coefficients were calculated.

Nature Biotechnology: doi:10.1038/nbt.2488

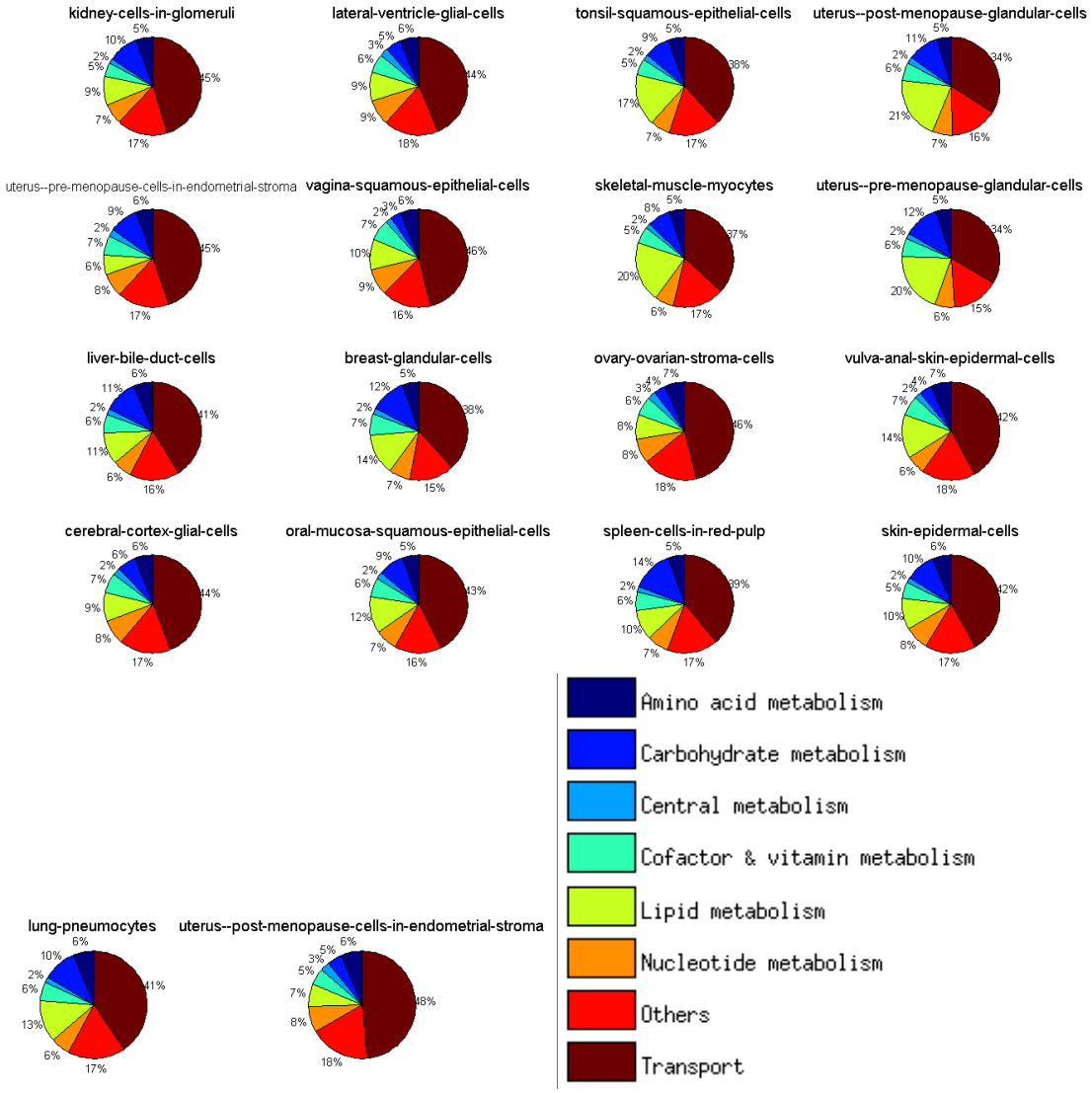
Supplementary Figure 5: Subsystem distribution in the 65 draft cell-type specific metabolic models based on reaction content



Nature Biotechnology: doi:10.1038/nbt.2488



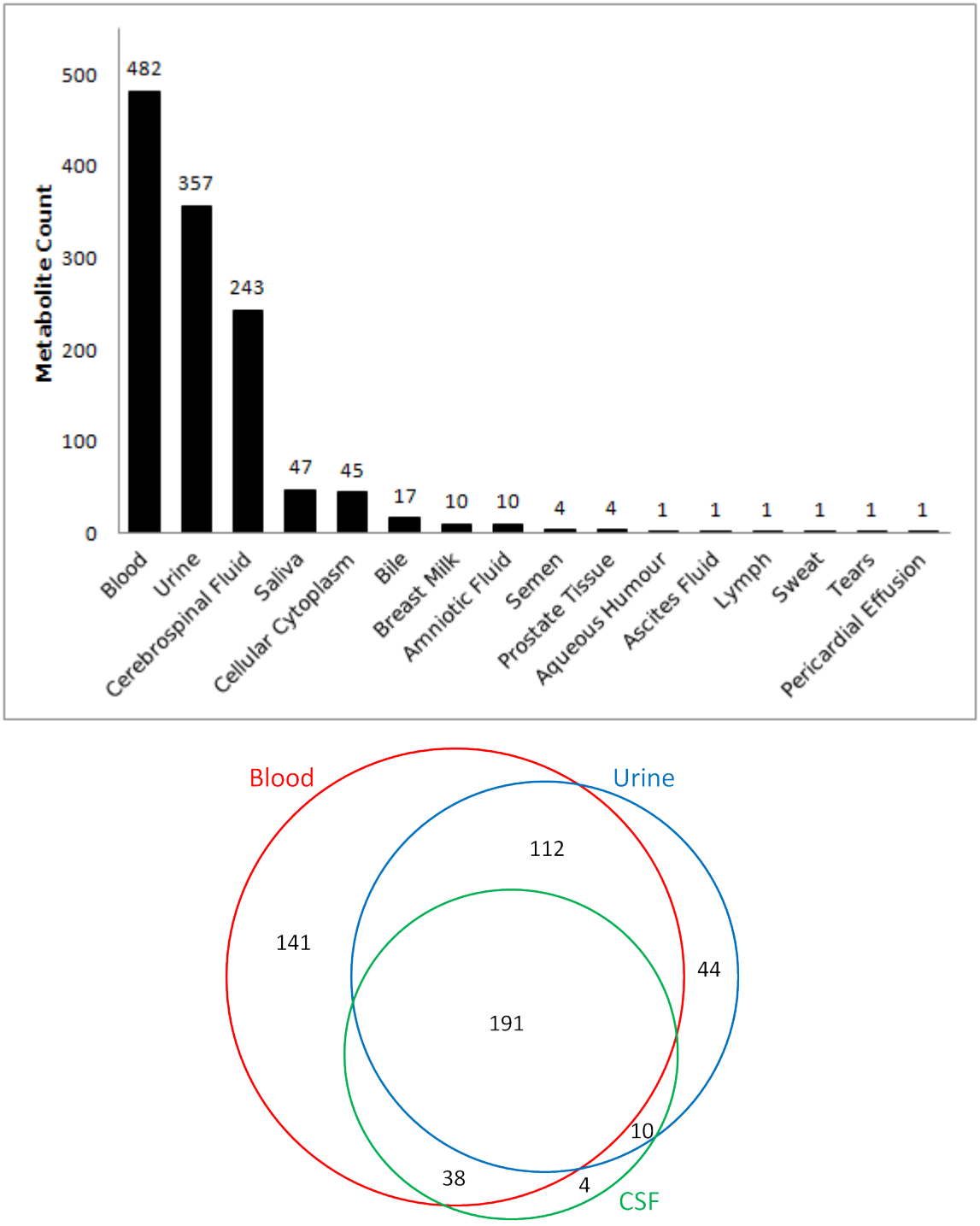
Nature Biotechnology: doi:10.1038/nbt.2488



Supplementary Figure 5: Subsystem distribution in the 65 draft cell-type specific metabolic models based on reaction content.

Nature Biotechnology: doi:10.1038/nbt.2488

Supplementary Figure 6: Analysis of biofluid data



Supplementary Figure 6: Biofluid distribution of the metabolites in Recon 2, as specified in the Human Metabolome Database[27]. The overlap between metabolites found in blood, urine, and cerebrospinal fluid (CSF) is shown in the inset. Most of the metabolites detected in CSF and urine were also present in blood. Only four metabolites were unique to CSF, being (S)-N-methylsalsolinol, methylcobalamin, D-alanine, and S-adenosylmethioninamine.

Nature Biotechnology: doi:10.1038/nbt.2488

Figure A: Semi-automated gap filling

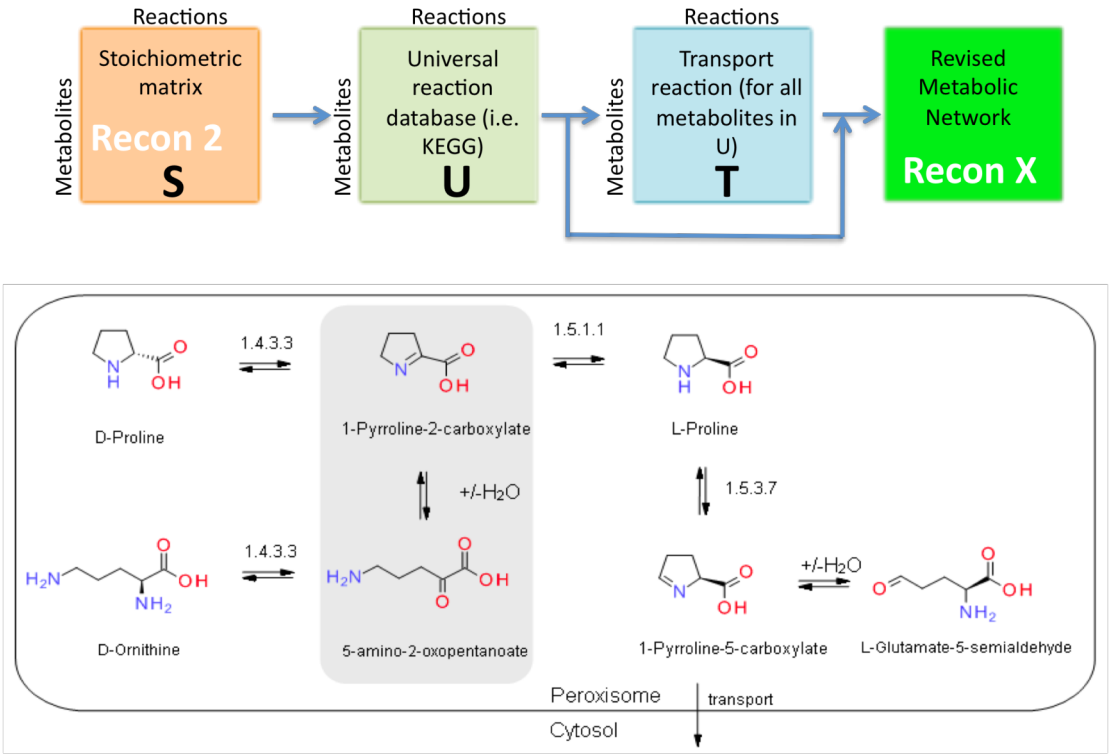


Figure A: Semi-automated gap filling suggests candidate, missing reactions, currently not included in Recon 2. Top: Schematic outline of the computational gap-filling procedure. A minimal number of reactions are borrowed from U or T to be added to S to enable flux through a blocked reaction. Bottom: The oxidation of both D-proline and D-ornithine is catalyzed by peroxisomal D-amino oxidase (EC 1.4.3.3) in Recon 2. Gap filling suggests conversion of 1-pyrroline-2-carboxylate to L-proline through EC 1.5.1.1 activity, thus, connecting the dead end metabolites (shaded grey).

Nature Biotechnology: doi:10.1038/nbt.2488

Figure B: Topological properties of Recon 2 compared with Recon 1

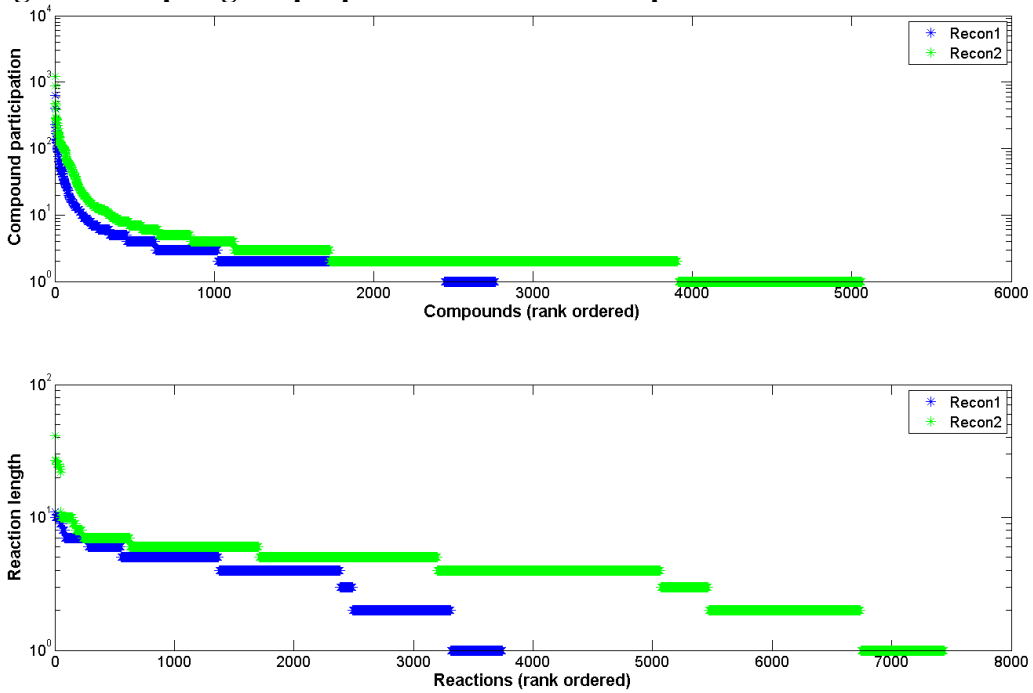


Figure B: Topological properties of Recon 2 compared with Recon 1. Top: Number of reactions each reactant participates in. Bottom: Number of reactants per reaction.

Nature Biotechnology: doi:10.1038/nbt.2488

References

1. Harris, M.A. et al. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* **32**, D258-261 (2004).
2. Matthews, L. et al. Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res* (2008).
3. Romero, P. et al. Computational prediction of human metabolic pathways from the complete human genome. *Genome biology* **6**, R2 (2005).
4. Kanehisa, M. et al. KEGG for linking genomes to life and the environment. *Nucleic Acids Res* **36**, D480-484 (2008).
5. Hao, T., Ma, H.W., Zhao, X.M. & Goryanin, I. Compartmentalization of the Edinburgh Human Metabolic Network. *BMC Bioinformatics* **11**, 393 (2010).
6. 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 of the United States of America* **104**, 1777-1782 (2007).
7. Stobbe, M.D., Houten, S.M., Jansen, G.A., van Kampen, A.H. & Moerland, P.D. Critical assessment of human metabolic pathway databases: a stepping stone for future integration. *BMC systems biology* **5**, 165 (2011).
8. Karp, P.D. et al. Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. *Brief Bioinform* **11**, 40-79 (2010).
9. Degtyarenko, K. et al. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res* **36**, D344-350 (2008).
10. Wheeler, D.L. et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **34**, D173-180 (2006).
11. Thiele, I. & Palsson, B.O. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature protocols* **5**, 93-121 (2010).
12. Bonarius, H.P.J. et al. Metabolic flux analysis of hybridoma cells in different culture media using mass balances. *Biotechnology and Bioengineering* **50**, 299-318 (1996).
13. Savinell, J.M. & Palsson, B.O. Network analysis of intermediary metabolism using linear optimization. I. Development of mathematical formalism. *Journal of theoretical biology* **154**, 421-454 (1992).
14. Sheikh, K., Forster, J. & Nielsen, L.K. Modeling hybridoma cell metabolism using a generic genome-scale metabolic model of Mus musculus. *Biotechnol Prog* **21**, 112-121 (2005).
15. Rolfsson, O., Palsson, B.O. & Thiele, I. The human metabolic reconstruction Recon 1 directs hypotheses of novel human metabolic functions. *BMC systems biology* **5**, 155 (2011).
16. Gudmundsson, S. & Thiele, I. Computationally efficient flux variability analysis. *BMC Bioinformatics* **11**, 489 (2010).
17. Reed, J.L. et al. Systems approach to refining genome annotation. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 17480-17484 (2006).
18. Schellenberger, J. et al. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nature protocols* **6**, 1290-1307 (2011).
19. Molla, G. et al. Characterization of human D-amino acid oxidase. *FEBS Lett* **580**, 2358-2364 (2006).
20. Kawazoe, T., Tsuge, H., Pilone, M.S. & Fukui, K. Crystal structure of human D-amino acid oxidase: context-dependent variability of the backbone conformation of the VAAGL hydrophobic stretch located at the si-face of the flavin ring. *Protein Sci* **15**, 2708-2717 (2006).
21. Meister, A. The alpha-keto analogues of arginine, ornithine, and lysine. *J Biol Chem* **206**, 577-585 (1954).
22. Meister, A., Radhakrishnan, A.N. & Buckley, S.D. Enzymatic synthesis of L-pipecolic acid and L-proline. *J Biol Chem* **229**, 789-800 (1957).
23. Garweg, G., von Rehren, D. & Hintze, U. L-Pipecolate formation in the mammalian brain. Regional distribution of delta1-pyrroline-2-carboxylate reductase activity. *J Neurochem* **35**, 616-621 (1980).
24. Dodt, G. et al. L-Pipecolic acid oxidase, a human enzyme essential for the degradation of L-pipecolic acid, is most similar to the monomeric sarcosine oxidases. *The Biochemical journal* **345 Pt 3**, 487-494 (2000).
25. Phang, J.M., Pandhare, J., Zabirnyk, O. & Liu, Y. PPARgamma and Proline Oxidase in Cancer. *PPAR Res* **2008**, 542694 (2008).
26. Uhlen, M. et al. Towards a knowledge-based Human Protein Atlas. *Nat Biotechnol* **28**, 1248-1250 (2010).
27. Wishart, D.S. et al. HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Res* **37**, D603-610 (2009).

Nature Biotechnology: doi:10.1038/nbt.2488