Metabolomics and transcriptomic data integration in GSMM using CORDA

Berta Bori-Bru[1,2,3]; Núria Aleixandre[1,2,3]; Sergi Pujol-Rigol[1,2,3]; David López-Blanco[3,4,5]; Sergio Madurga[1,2]; Silvia Marin[3,4,5]; Pedro de Atauri[3,4,5]; Francesc Mas[1,2]; Marta Cascante[3,4,5]

[1]Department of Material Science and Physical Chemistry , University of Barcelona, 08028 Barcelona, Spain.
[2]Research Institute of Theoretical and Computational Chemistry of University of Barcelona (IQTCUB), 08028 Barcelona, Spain
[3]Department of Biochemistry and Molecular Biomedicine, Faculty of Biology, University of Barcelona, 08028 Barcelona, Spain.
[4]Institute of Biomedicine of the University of Barcelona (IBUB), University of Barcelona, 08028 Barcelona, Spain.
[5]CIBEREHD, Network Centre for Hepatic and Digestive Diseases, National Spanish Health Institute Carlos III (ISCIII), 28029 Madrid, Spain.

Summary

Cellular metabolism is a complex network of biochemical reactions, and its accurate reconstruction may elucidate essential components in the understanding cellular function, disease mechanisms, and therapeutic targeting.

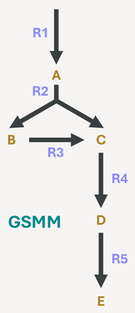
Can we use computational methods to reconstruct condition-specific metabolic models with transcriptomic and metabolomic data?

We developed a Python pipeline that utilizes Genome-Scale Metabolic Models (GSMMs) and the CORDA algorithm to generate condition-specific models and used it to reconstruct the metabolism of several colorectal cancer (CRC) cell lines and analyze the adjustment to metabolomics data, the reproductibility of CRISPR-KO data and the flux distribution through different metabolic pathways.

Results indicate that our pipeline reconstructs models that reproduce metabolomics data accurately, but more fine tunning needs to be made to correctly predict therapeutic targets. The pipeline also elucidates differences in the metabolism of different CRC cell lines.

GSMMs

- GSMMs are represention of metabolic networks containing reactions, metabolites and genes.
- Using steady-state approximation and defining constrains, we can obtain the flux distribution that optimizes an objective function (i.e. biomass production).



GSMM

Reactions

Metabolites

$$\begin{pmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 1 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{pmatrix}$$

Stoichiometric matrix

$$S \times v = 0$$
$$LB_i \leq v_i \leq UB_i$$
$$Max(v_{biomass})$$

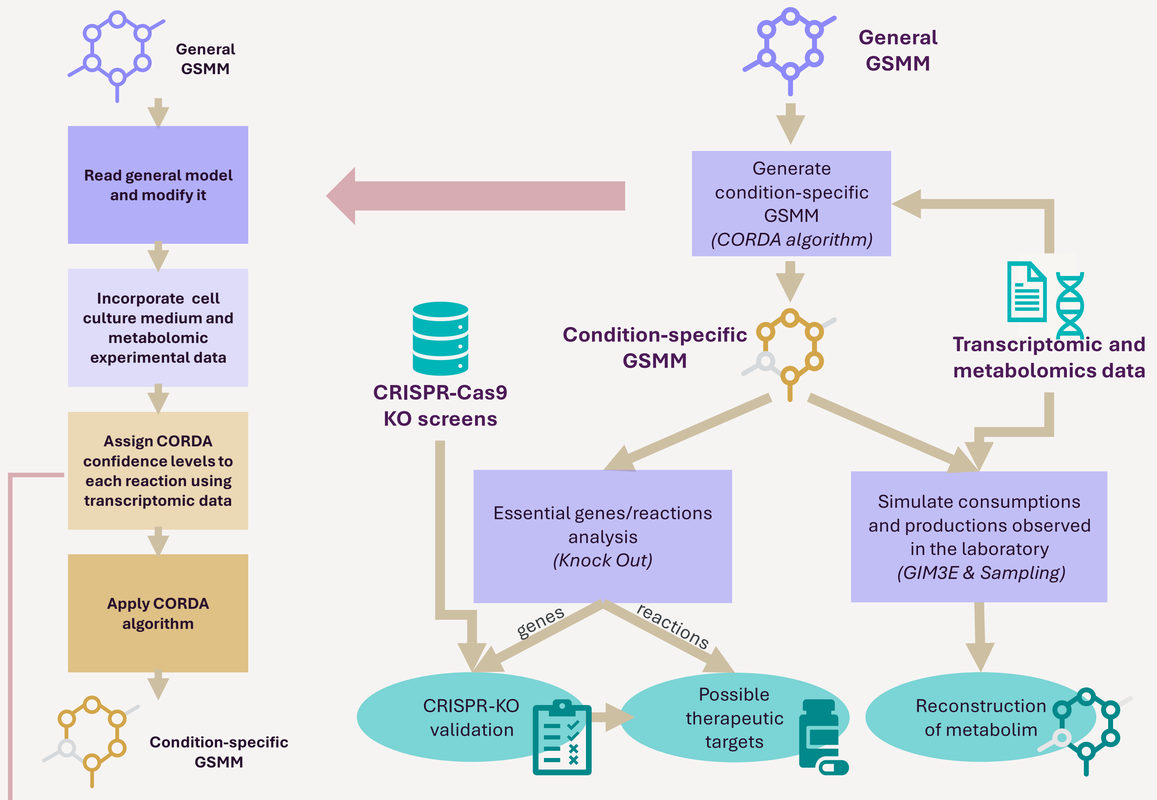
Linear programming problem (Flux Balance Analysis)

CORDA algorithm

- CORDA is an algorithm to reconstruct condition-specific GSMMs based on FBA.
- The algorithm aims to minimize the number of reactions unlikely to be present in a specific condition while maximizing the essential reactions for known metabolic functions.
- Reactions are scored based on available data (e.g., gene expression levels, literature evidence).
- CORDA then tries to reconstruct the minimal viable model that includes the maximum number of high-evidence reactions (higher score) while pruning reactions with low evidence (negative score).

Our pipeline

- We designed a computational pipeline with bash and Python, using the available libraries COBRApy and CORDA to reconstruct GSMMs for several CRC cell lines from transcriptomic and metabolomic data. We compared the reconstructed models with experimental consumption and production fluxes from the metabolomics data and CRISPR-Cas9 Knock Out screenning data from external databases.



- To assign confidence levels or scores to the model's reactions, we first analyze the distribution of gene expression levels and assign confidence levels to the genes. Then, we map these levels to the reactions using Gene-Protein-Reaction (GPR) associations.

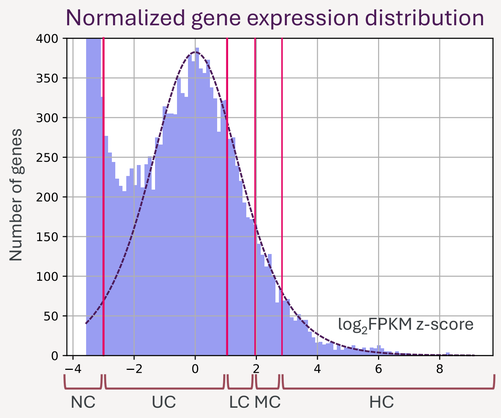
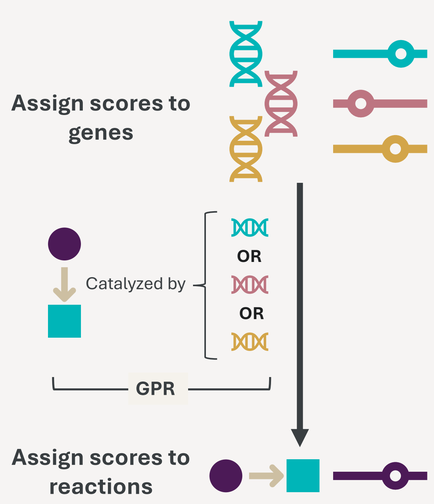


FIG 1. NC = Negative Confidence; **UC** = Unknown Conf. **LC** = Low Conf. **MC** = Mediu Conf. **HC** = High Conf. The gene expression distribution follows a logistic distribution, with most genes having minimal expression.



Results

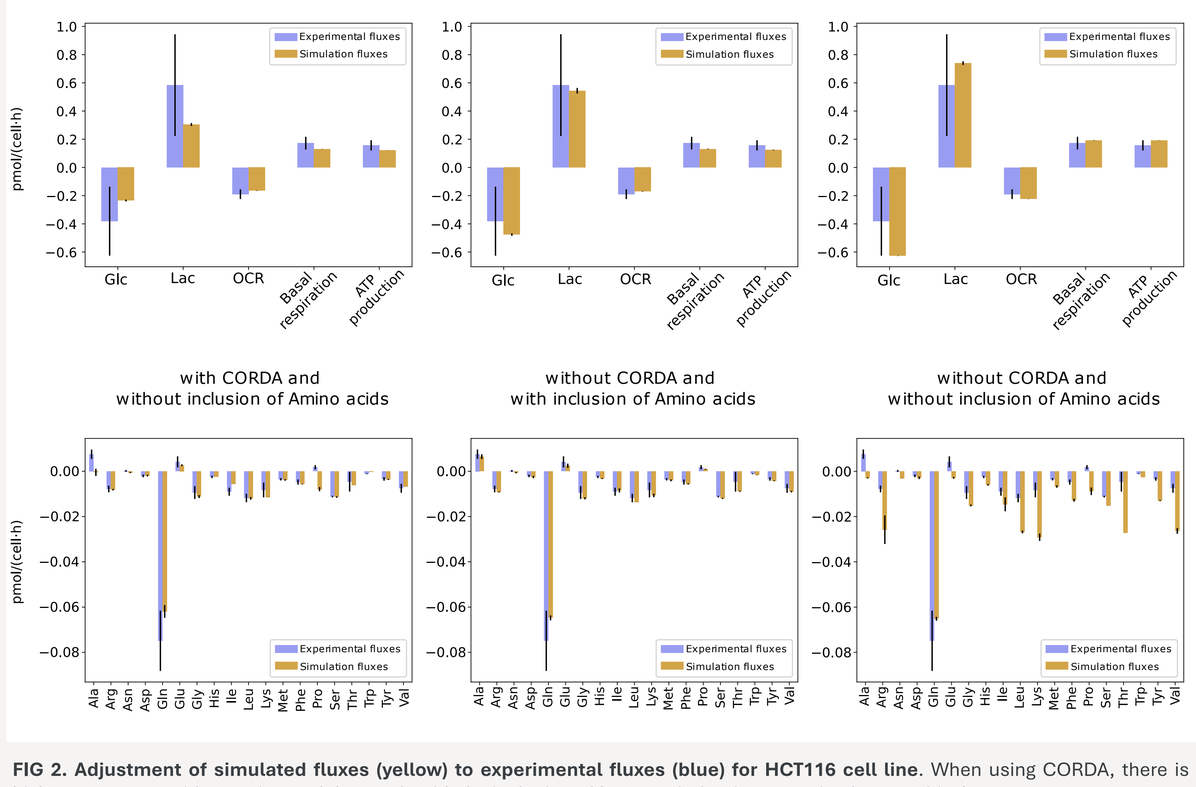
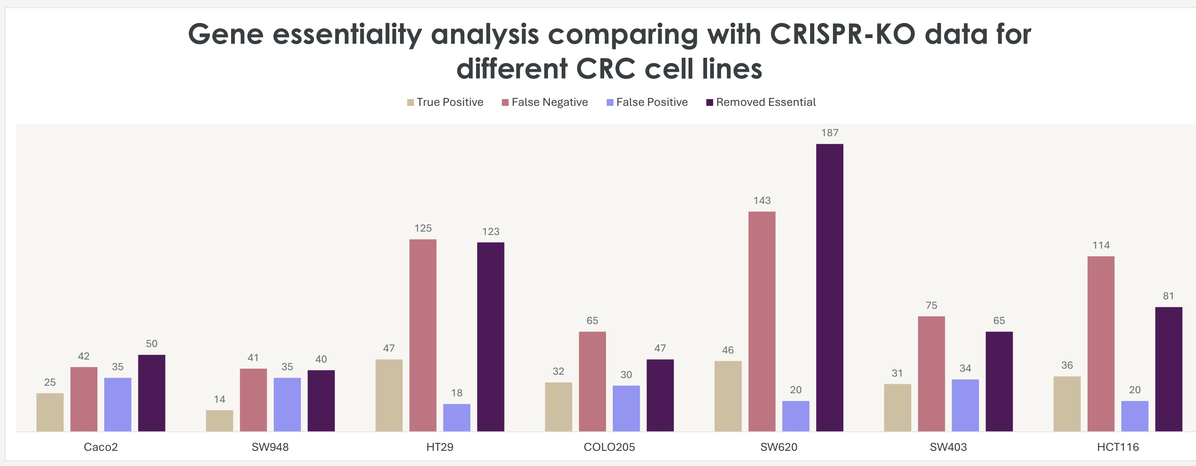


FIG 2. Adjustment of simulated fluxes (yellow) to experimental fluxes (blue) for HCT116 cell line. When using CORDA, there is a higher agreement with experimental data, only with the inclusion of few metabolomic constrains (top graphics).



	Caco2	SW948	HT29	COLO205	SW620	SW403	HCT116
Metabolic Genes	1883	1883	1883	1883	1883	1883	1883
CRISPR Hit Genes	117	95	295	144	376	171	231
% HITS guessed correctly	21	15	16	22	12	18	16

FIG 3. Essential gene analysis for different CRC cell lines. Results show that a high number of essential genes are being removed from the reconstructed models, which could indicate the reconstructed models need to be bigger.

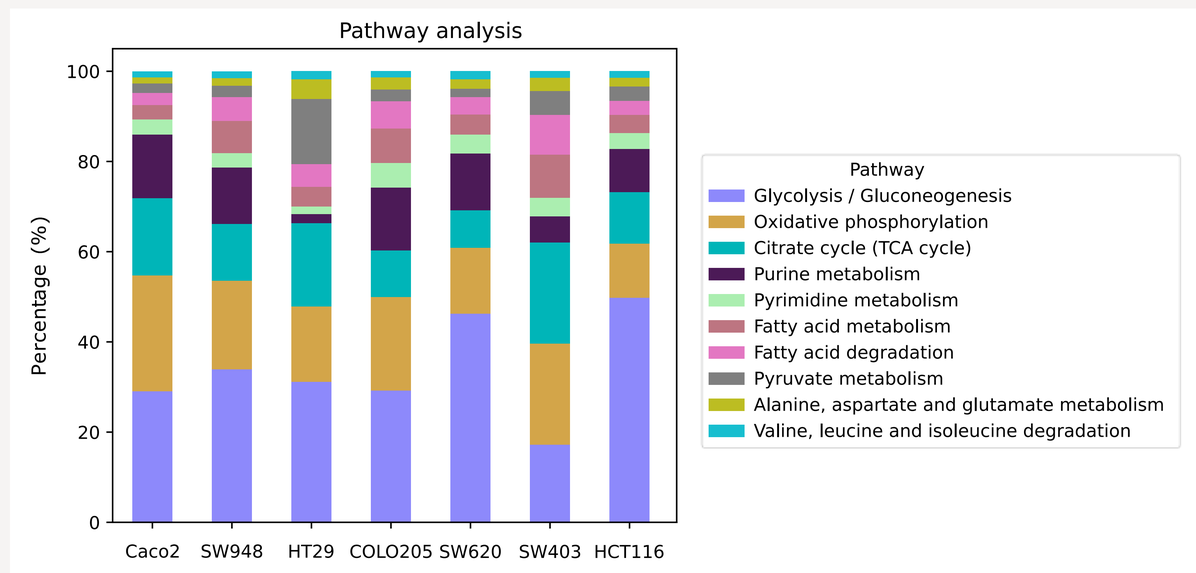


FIG 4. Comparison of flux value distribution through each metabolic pathway for different CRC cell lines. Differences in the usage of the metabolic pathways elucidates differences in the metabolism of the studied cells.

References

J. D. Orth, I. Thiele, B. Ø. Palsson. *Nat. Biotechnol*. **2010**, *28* (3), 245

A. Schultz, A. A. Qutub. *PLoS Comput Biol* **2016**, *12* (3), e1004808.

A. Ebrahim, J. A. Lerman, B. Ø. Palsson, *et al*. *BMC Syst. Biol*. **2013**, *7* (1), 74.

Funding

This work is supported by the Spanish Ministry of Science, Innovation and Universities (PID2020-115051RB-I00) and Centro de Investigación Biomédica en Red en Enfermedades hepáticas y digestivas (CB17/04/00023) from the Instituto de Salud Carlos III; by AGAUR (2021 SGR 00350); ICREA foundation (ICREA-Academia-2021 to MC) of Generalitat de Catalunya and Spanish Structures of Excellence Maria de Maetzu program (CEX2021-001202-M).