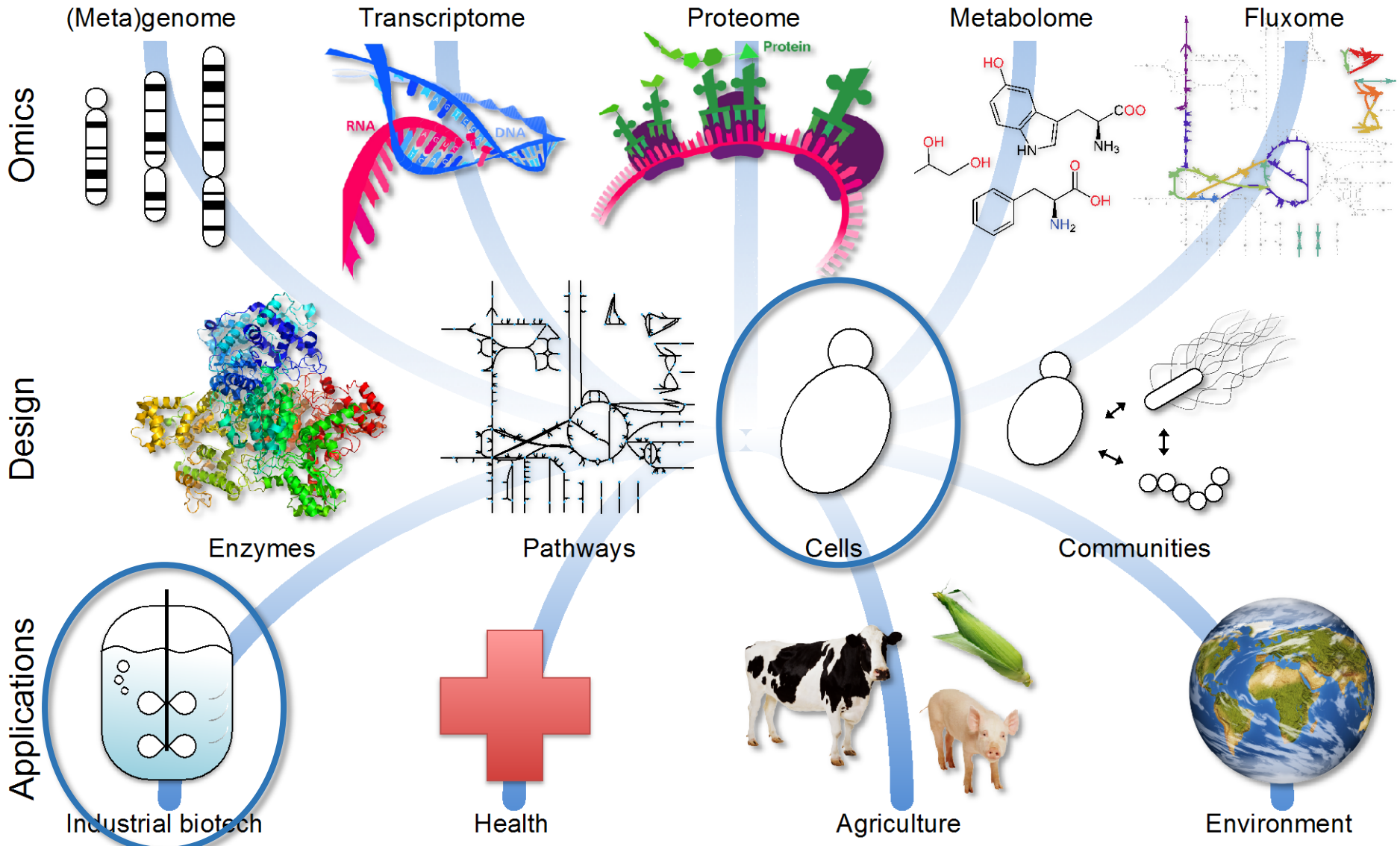
Advanced models and omics data integration

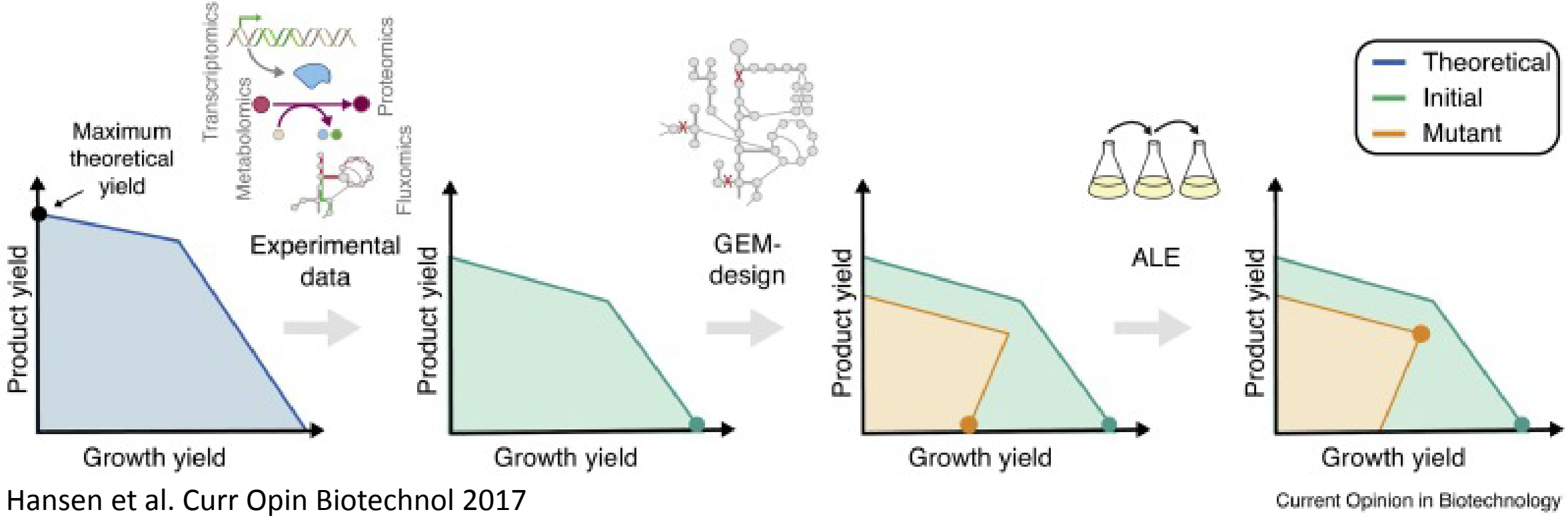
Markus Herrgård
The Novo Nordisk Foundation Center for Biosustainability
Technical University of Denmark

Benjamin Sanchez
Chalmers University of Technology
Gothenburg, Sweden

Omics data types of interest



Ideal workflow for cell factory design

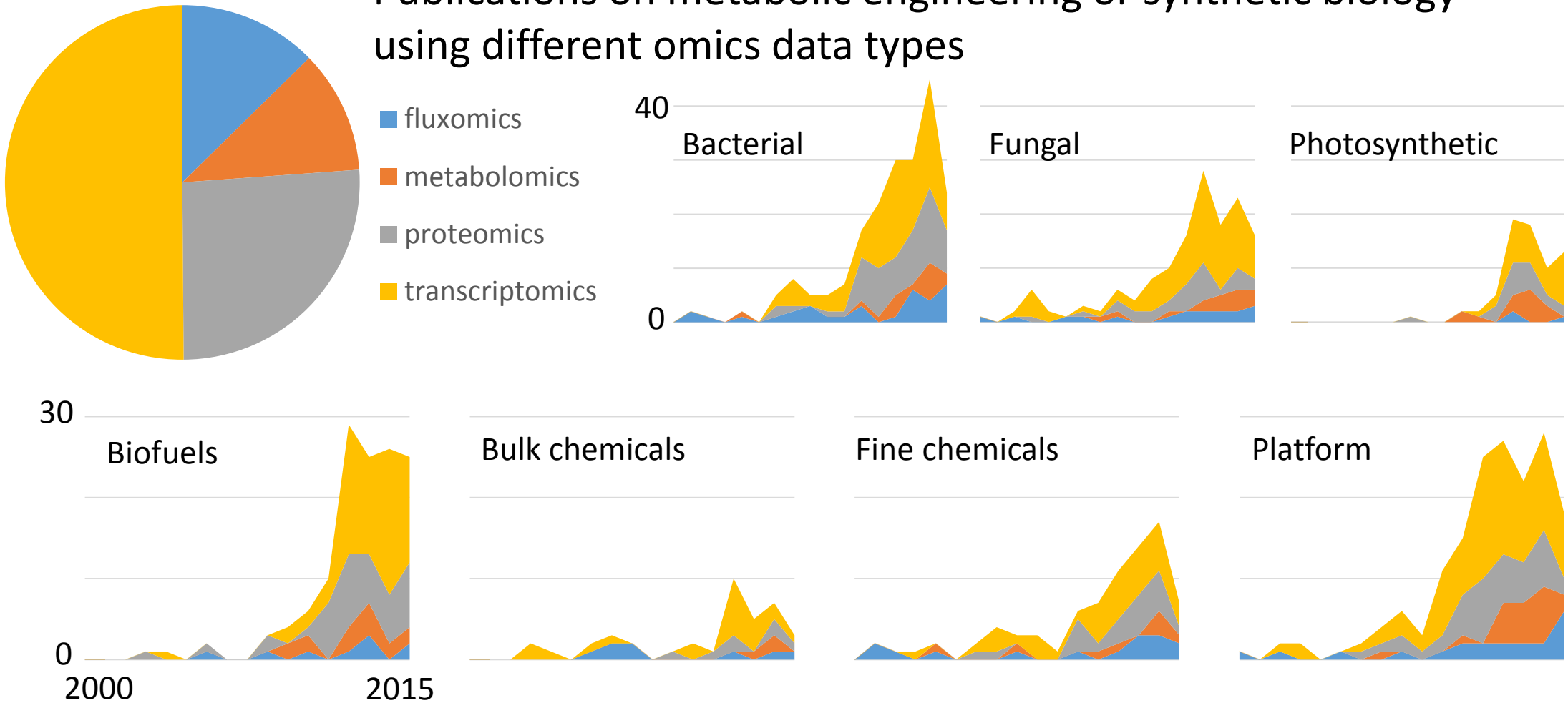


Hansen et al. Curr Opin Biotechnol 2017

Data can be used to reduce uncertainty about the system and to enable better rational design

What data is actually available?

Publications on metabolic engineering or synthetic biology using different omics data types

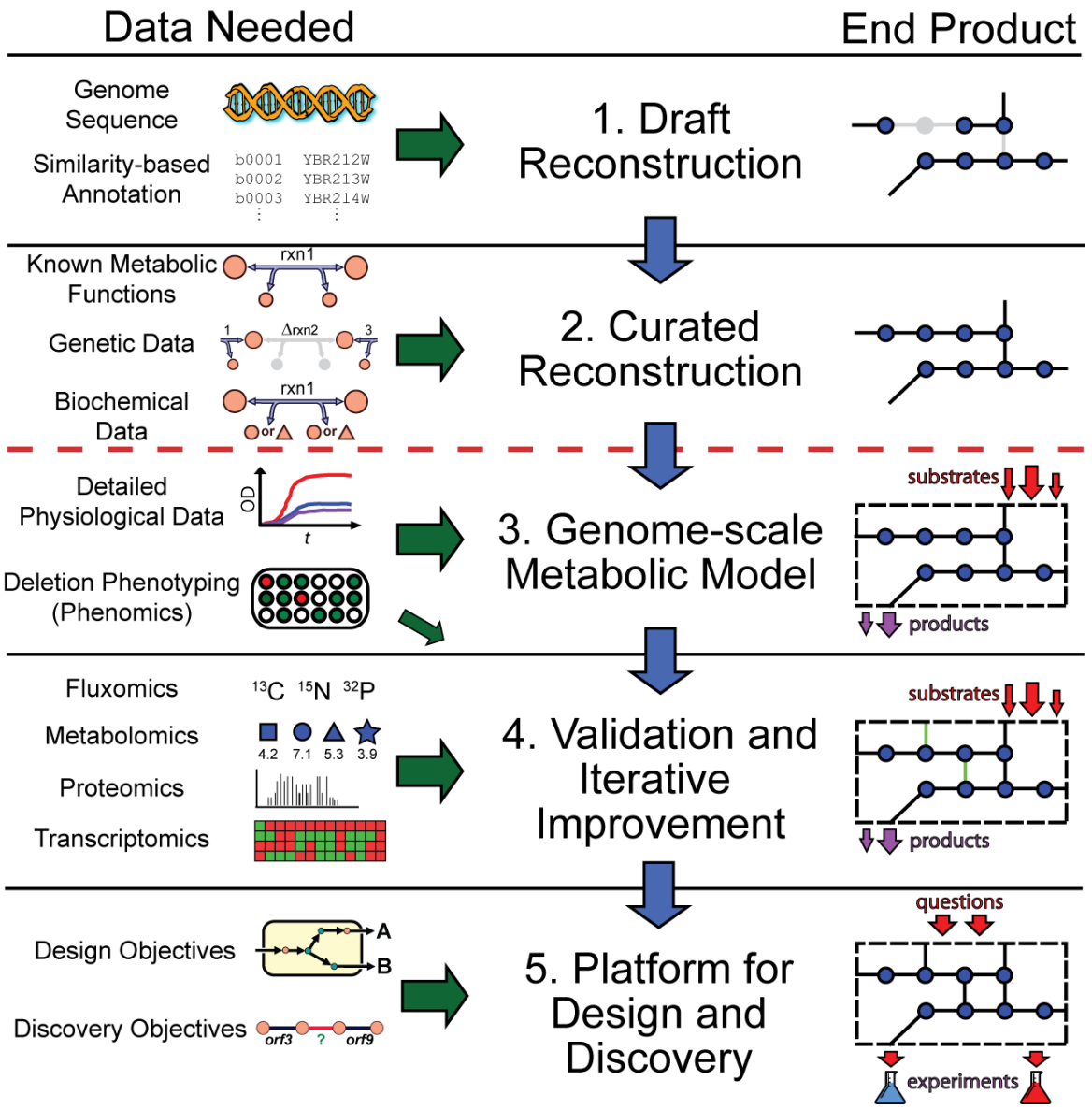


Hansen et al. Curr Opin Biotechnol 2017

Genomics data

Genomic data gives the basic parts list for any type of model

Data is easily available, but it does not tell us anything about the functional state of the cell

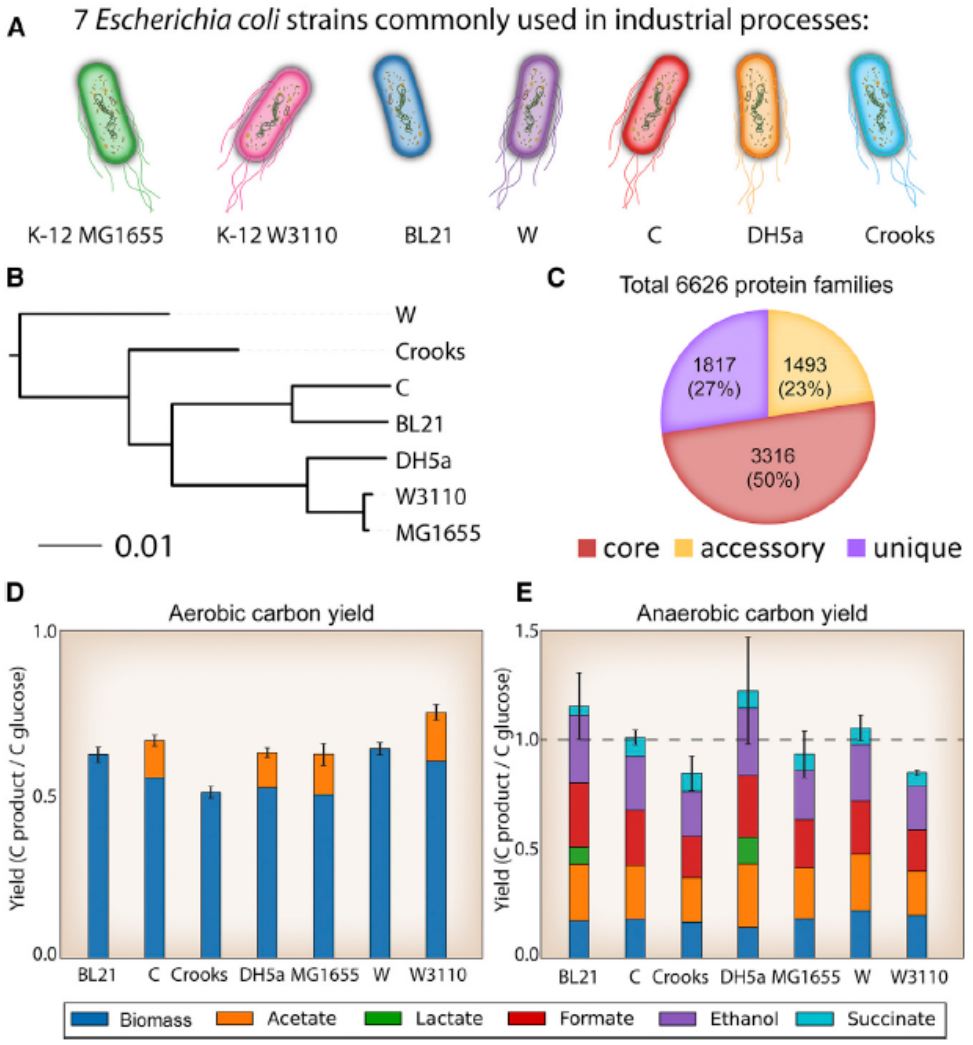


Nature Reviews Microbiology, Feb 2009

Comparative genomics data: Strain-specific models

Genomic data is used to customize models by removing or adding metabolic reactions

Physiological data is needed to parameterize customized models

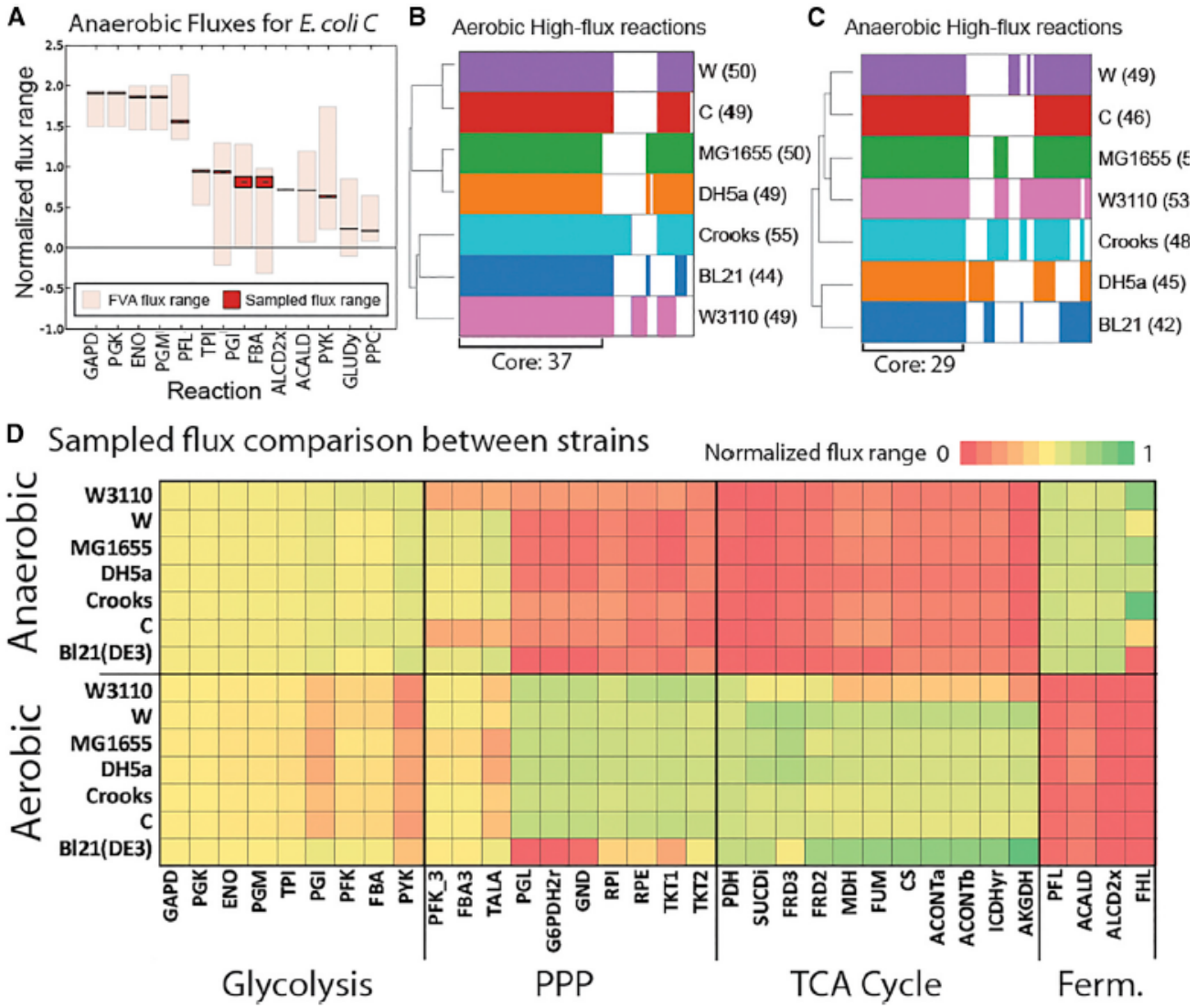


Monk et al. Cell Systems 2016

Predictions from strain-specific models

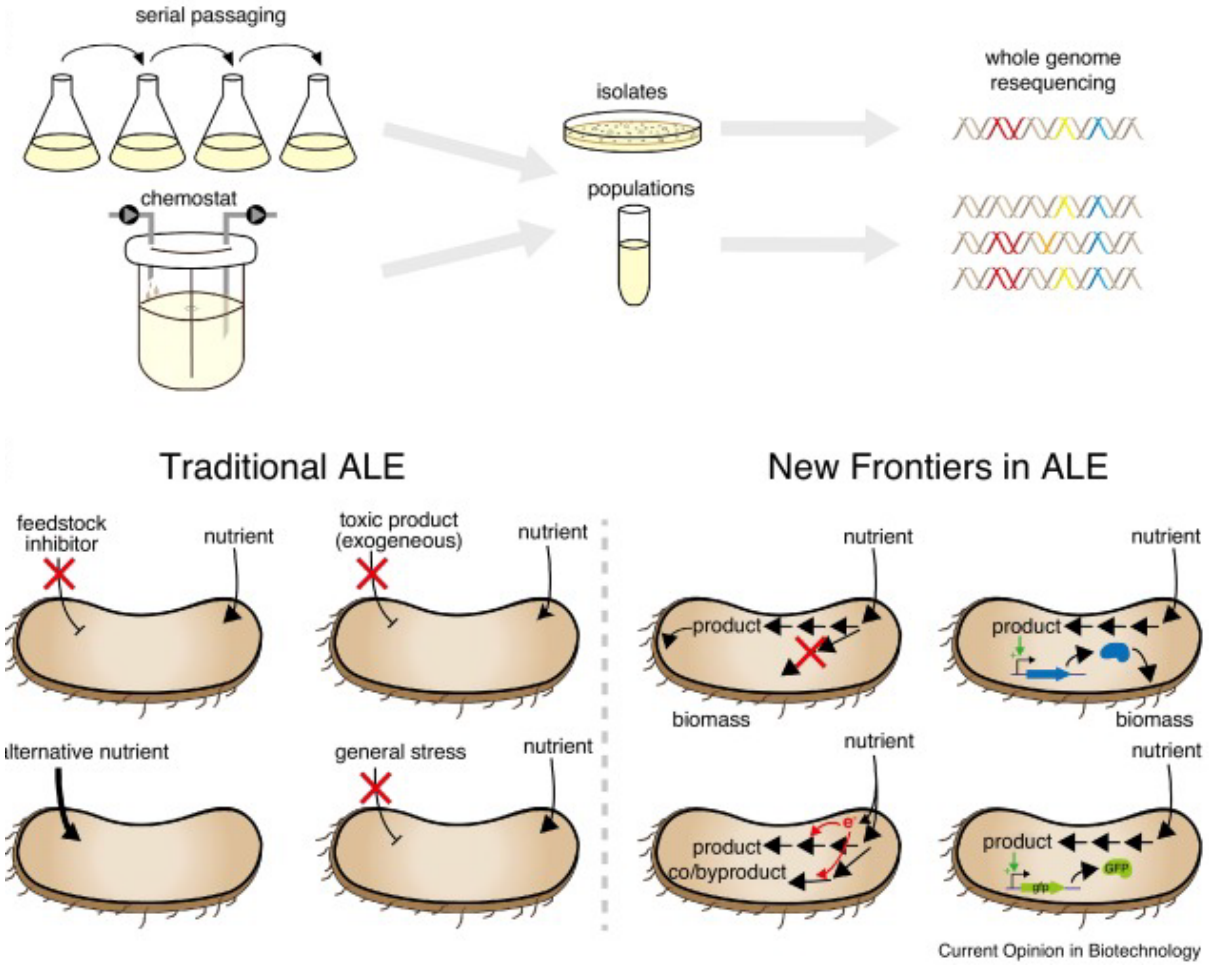
Can aid in deciding what strain to use for production of a specific compound

Caveat: Models do not include information on e.g. Strain-specific protein expression capacity or stress resistance



Monk et al. Cell Systems 2016

What about resequencing data?

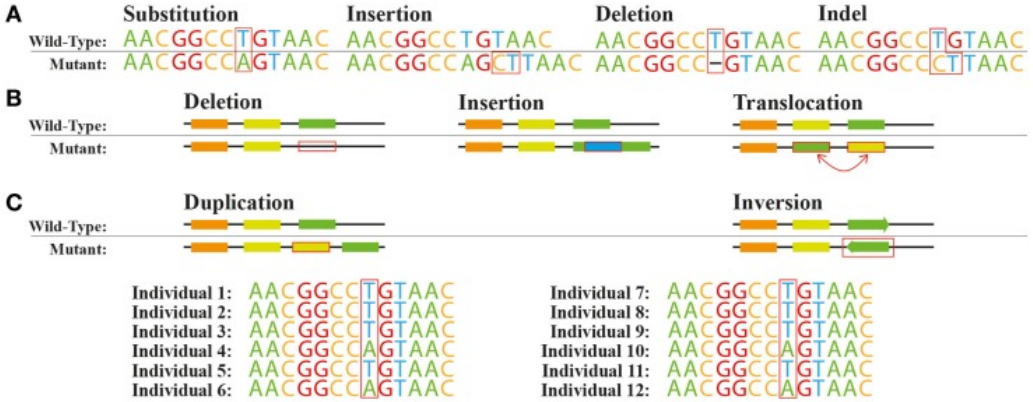


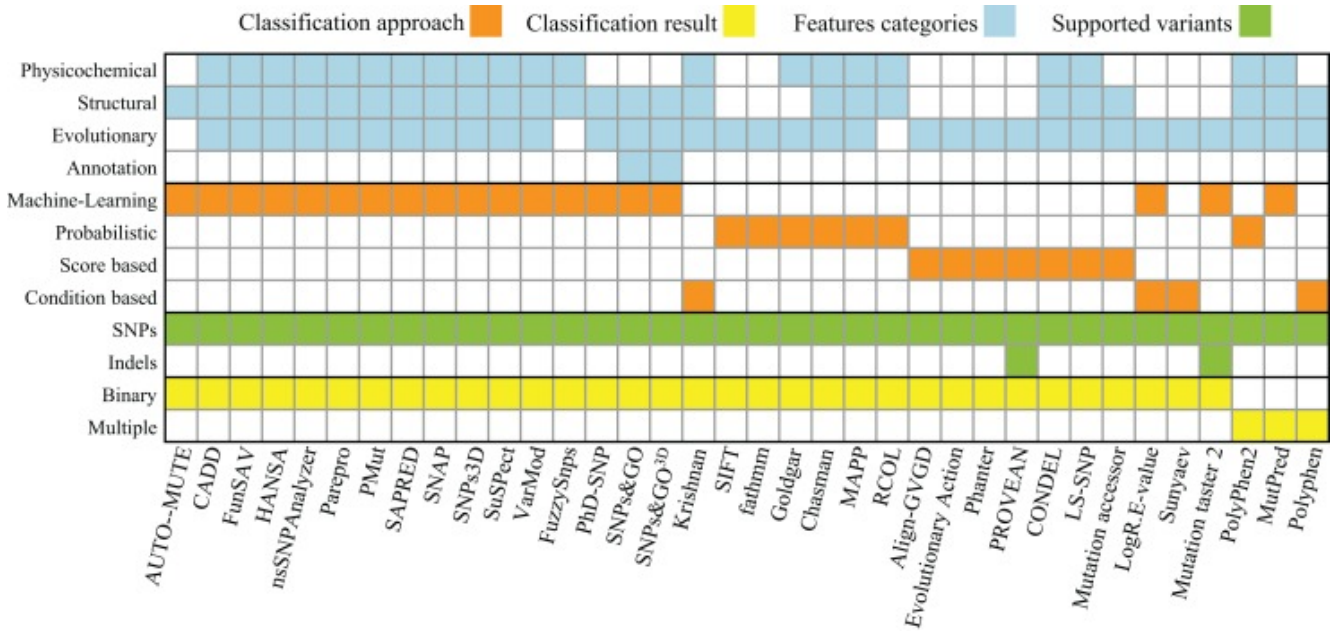
Hansen et al. Curr Opin Biotechnol 2017

Evolutionary engineering and classical strain improvement are standard tools in cell factory development

Next-generation sequencing has made generation of this type of data routine and cheap

Using resequencing data





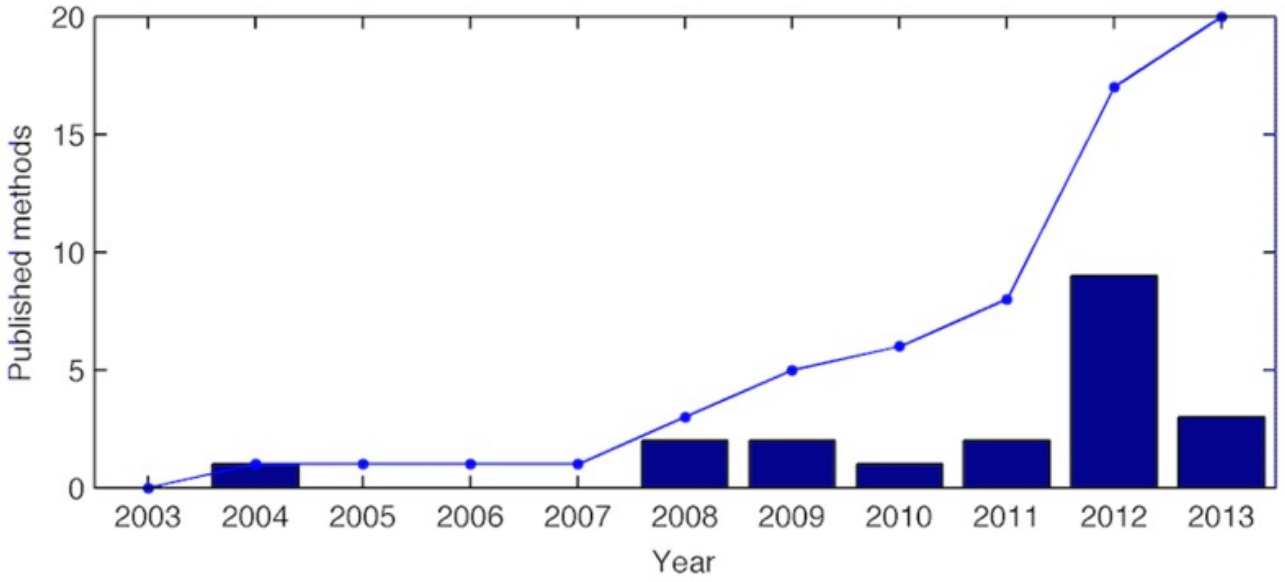
The major challege is predicting effects of mutations on enzyme activity beyond obvious loss-of-function mutations

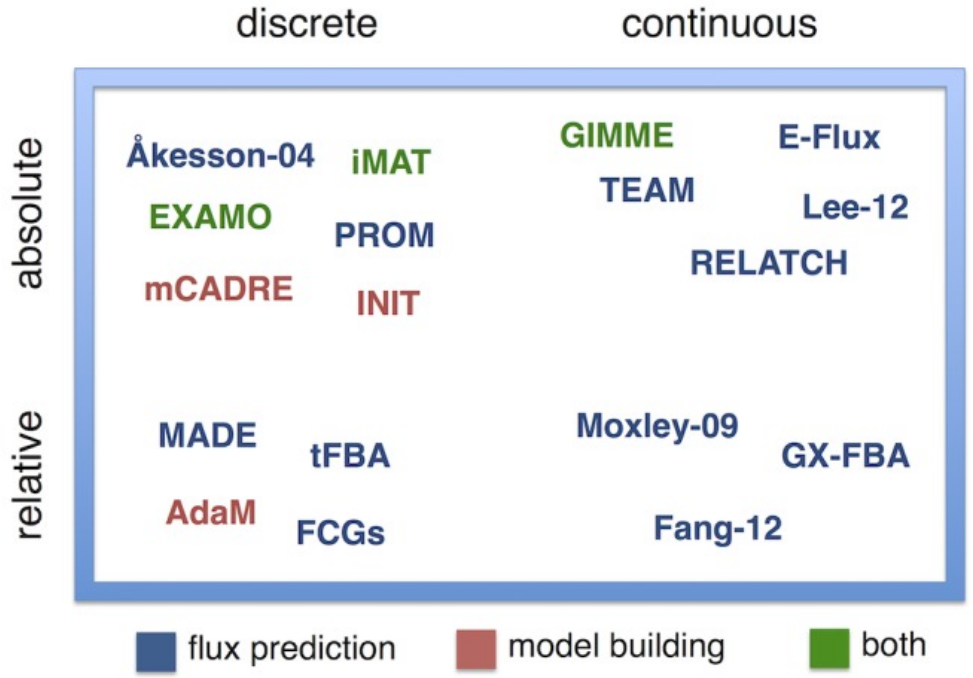
Cardoso et al. Front Bioeng Biotechnol 2016

For predictive modeling need to be able to predict the quantitative effect of mutations on k_{cat} and/or protein level

Transcriptomics data

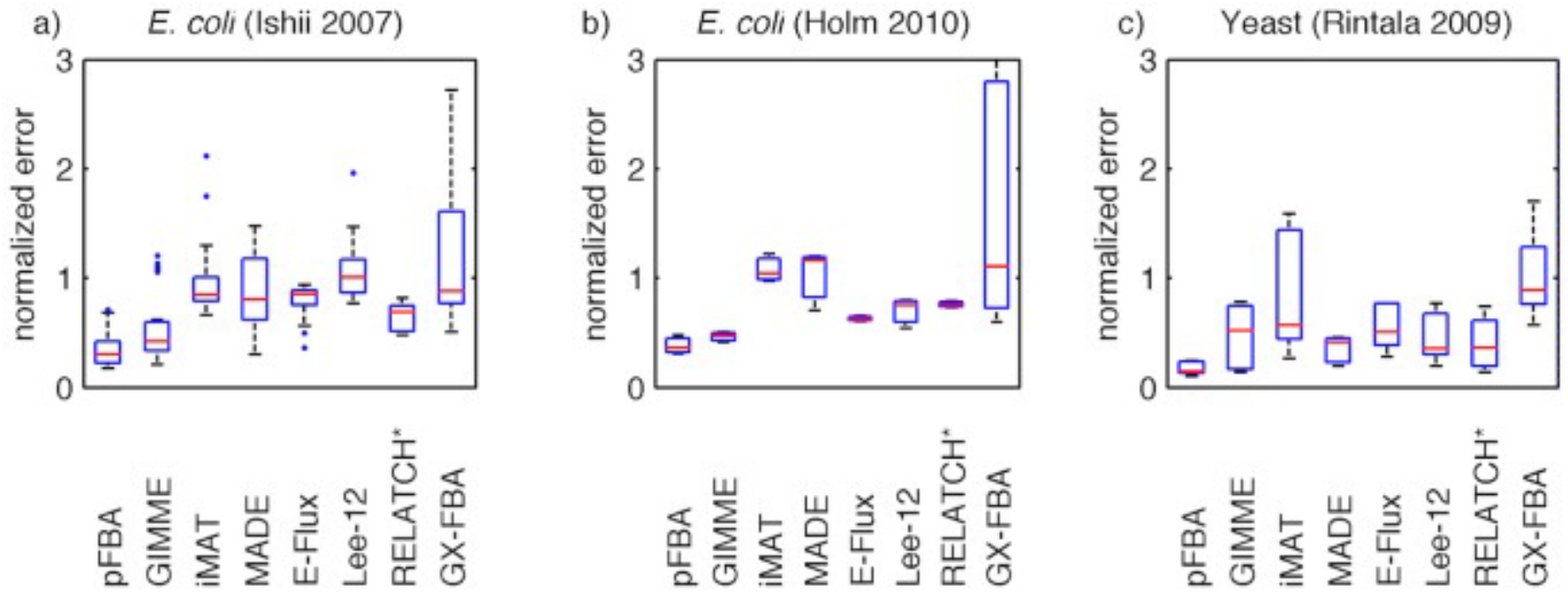
- Transcriptomics data is the most readily available omics data type for cell factory development projects
- Tens of different methods for integrating transcriptomics data with genome-scale metabolic models have been published with the aim to improve flux predictions using this data





Machado & Herrgård PLoS Comp Biol 2014

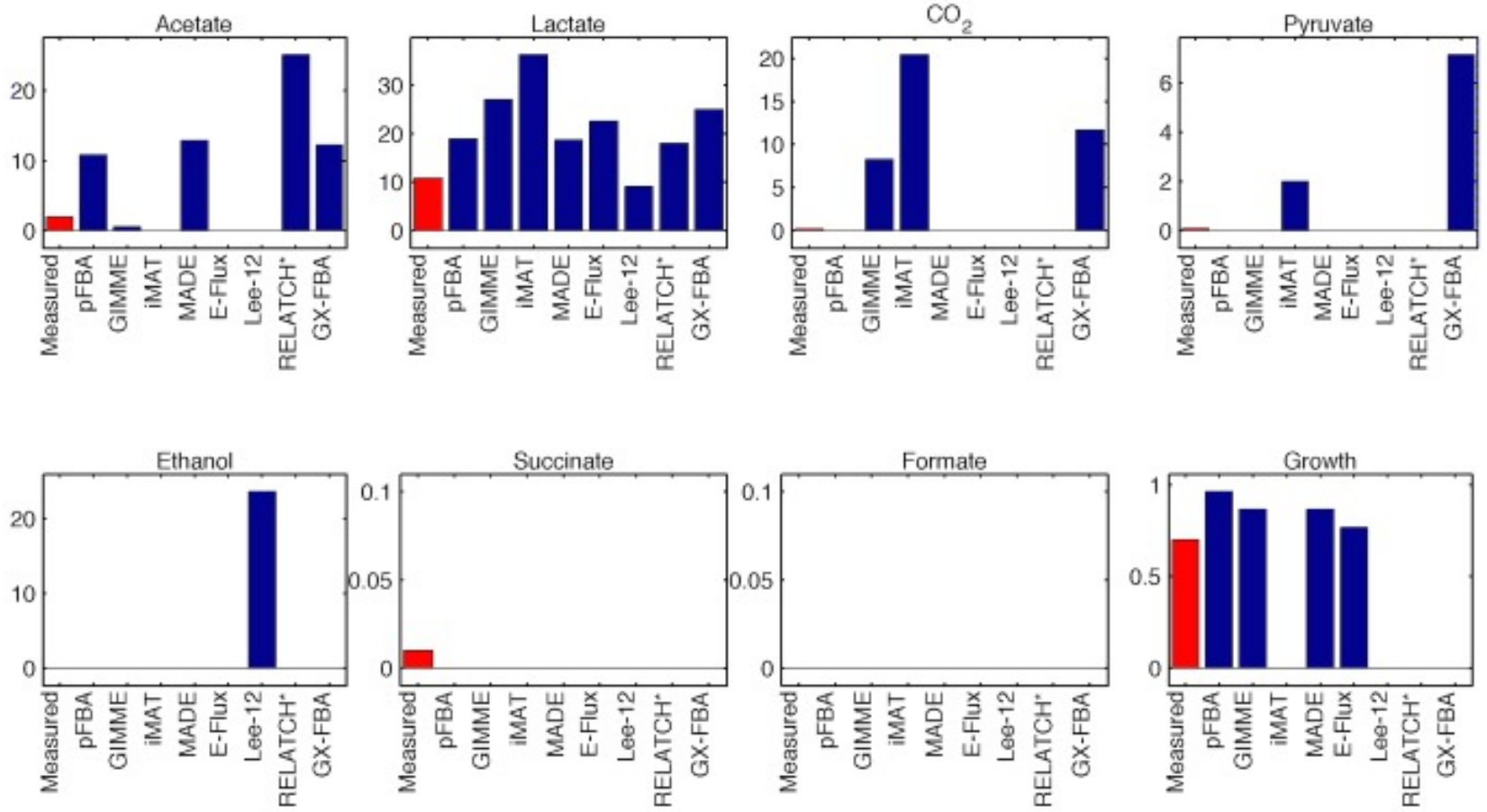
Does transcriptomics data help predicting fluxes?



Not really

Machado & Herrgård PLoS Comp Biol 2014

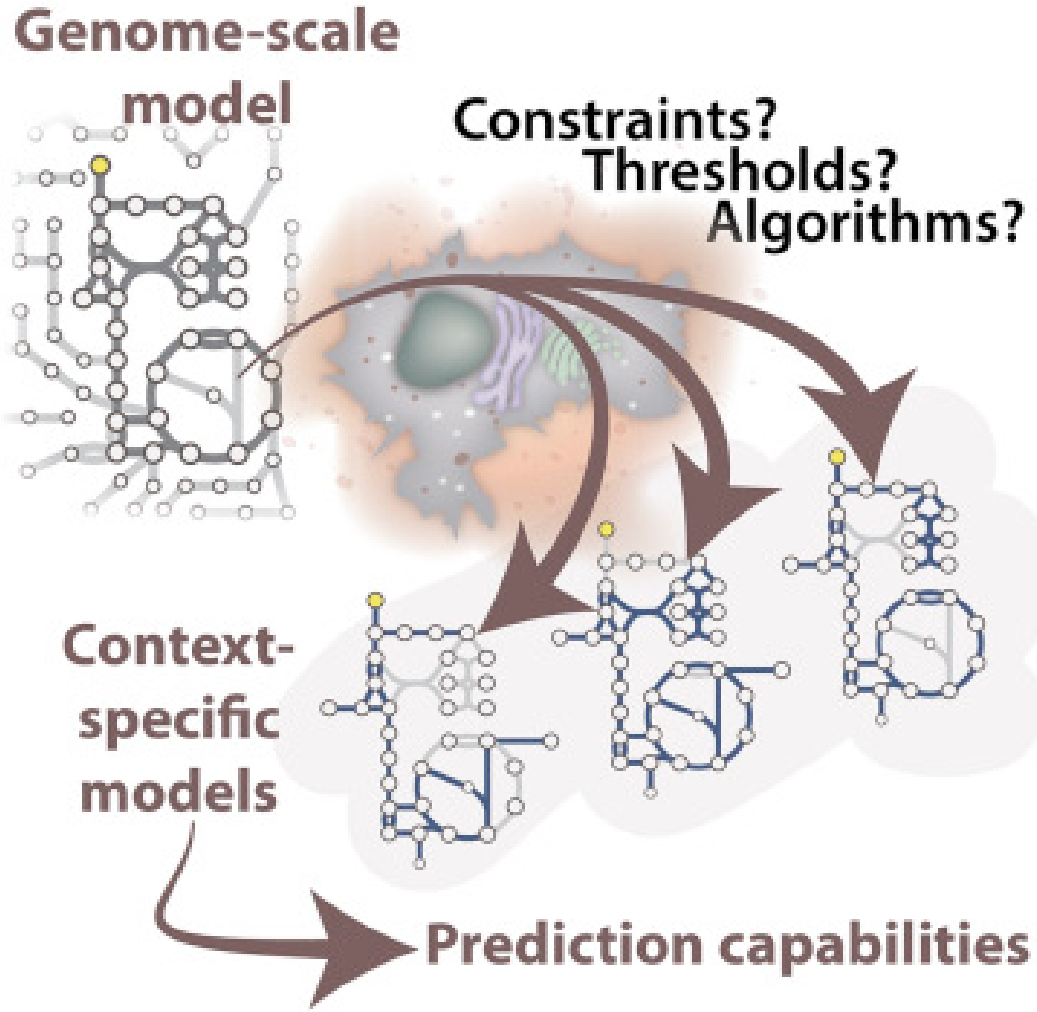
Really?



Yes really, one can not predict fluxes with a traditional genome-scale metabolic model and transcriptome (or proteome) data only

Machado & Herrgård PLoS Comp Biol 2014

Surely there is some way to use transcriptomics data



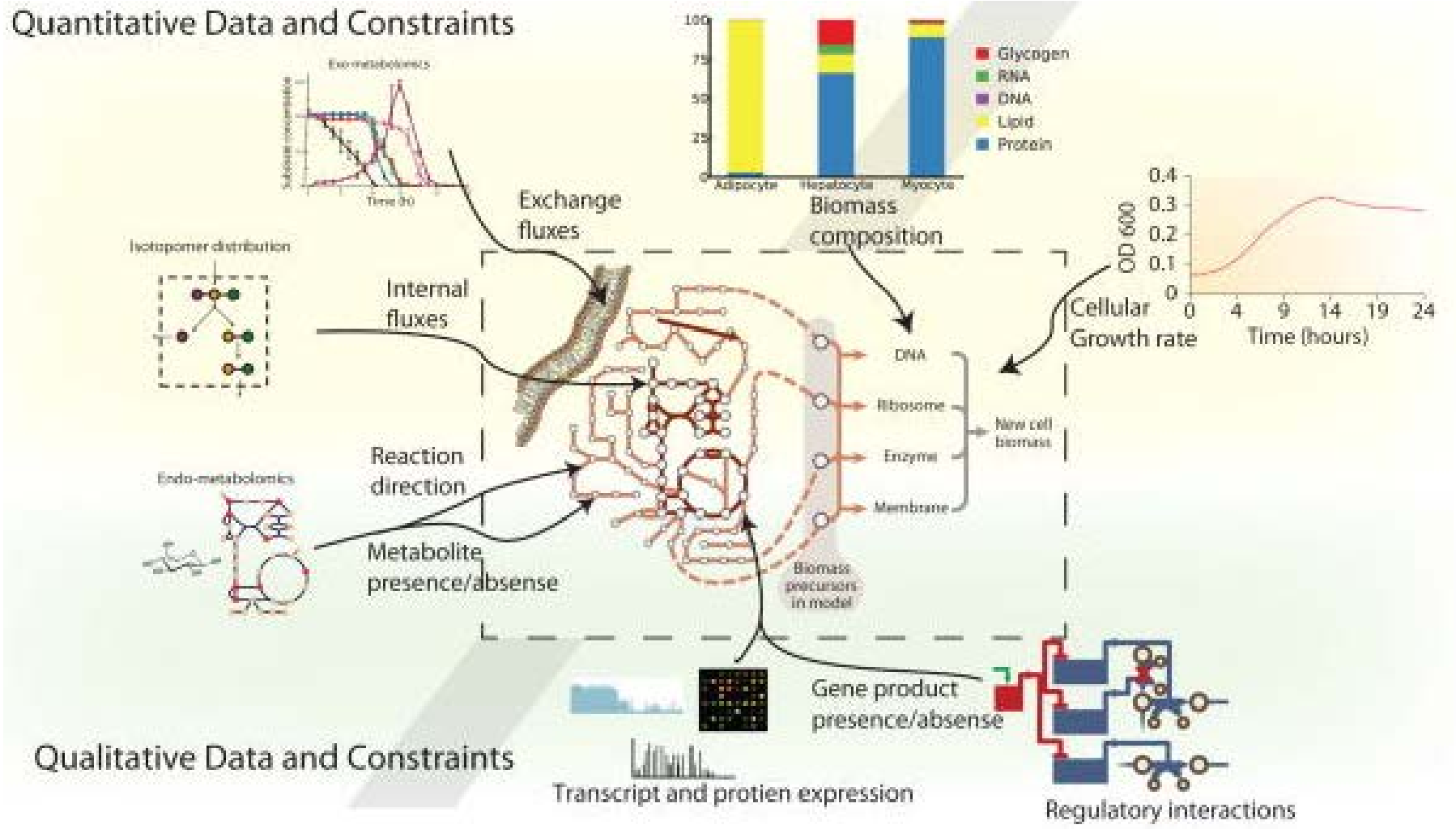
Opdam *et al.* Cell Systems 2017

Building tissue- or cell line-specific models using transcriptomics data seems to work

Caveats:

- There are many methods for this that all give different models
- Data handling (normalization, thresholds etc) also has a strong effect

Quantitative vs qualitative constraints from data

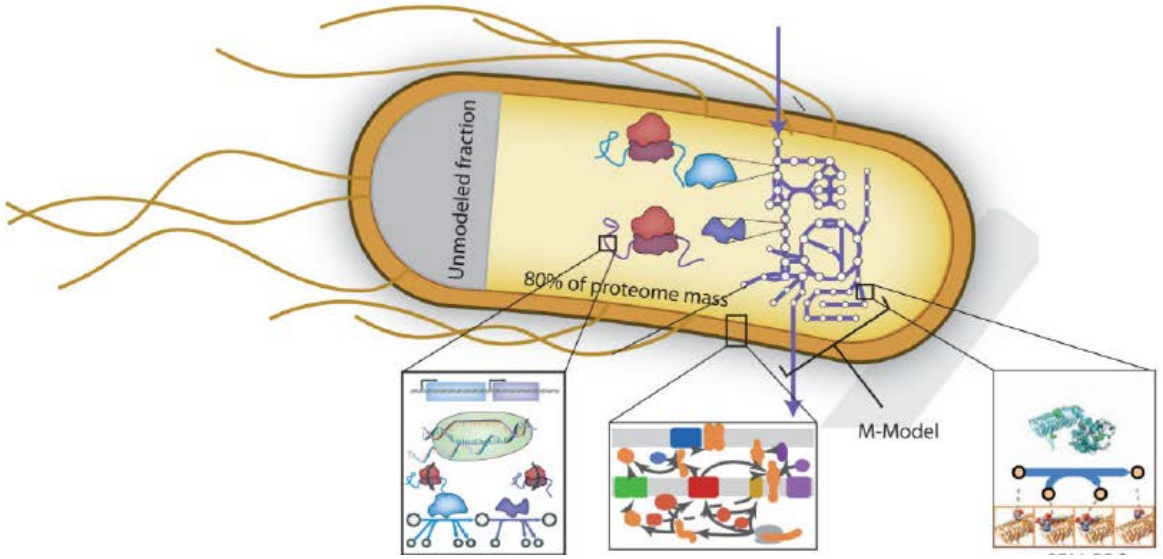


For traditional GEMs genomics, transcriptomics, proteomics and metabolomics act as qualitative constraints

Fluxomics and physiological data (uptake/secretion/growth rates) are the only data types that give quantitative constraints for GEMs

O'Brien et al. Cell 2016

Expanding model scope



In order to use transcriptomics, proteomics or metabolomics quantitatively, need to expand model scope to **represent transcripts, proteins and metabolites explicitly**

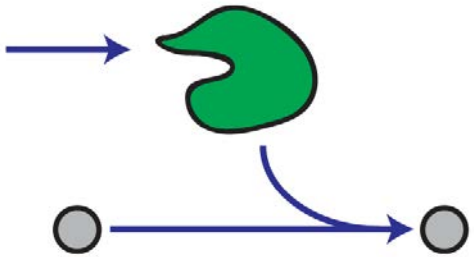
O'Brien et al. Cell 2016

Proteomics data

Approaches for integrating intracellular or secreted proteomics data

1

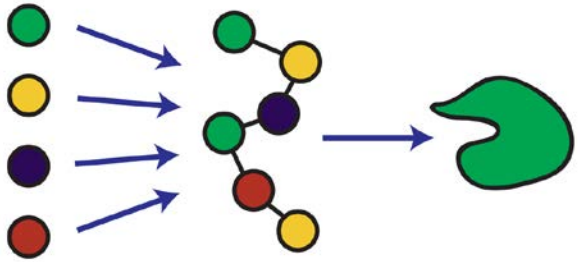
Include enzyme abundances as constraints in metabolic models



GECKO model

2

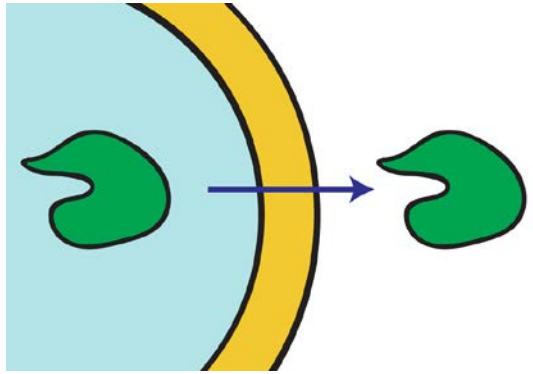
Model protein biosynthesis as part of metabolism



ME model

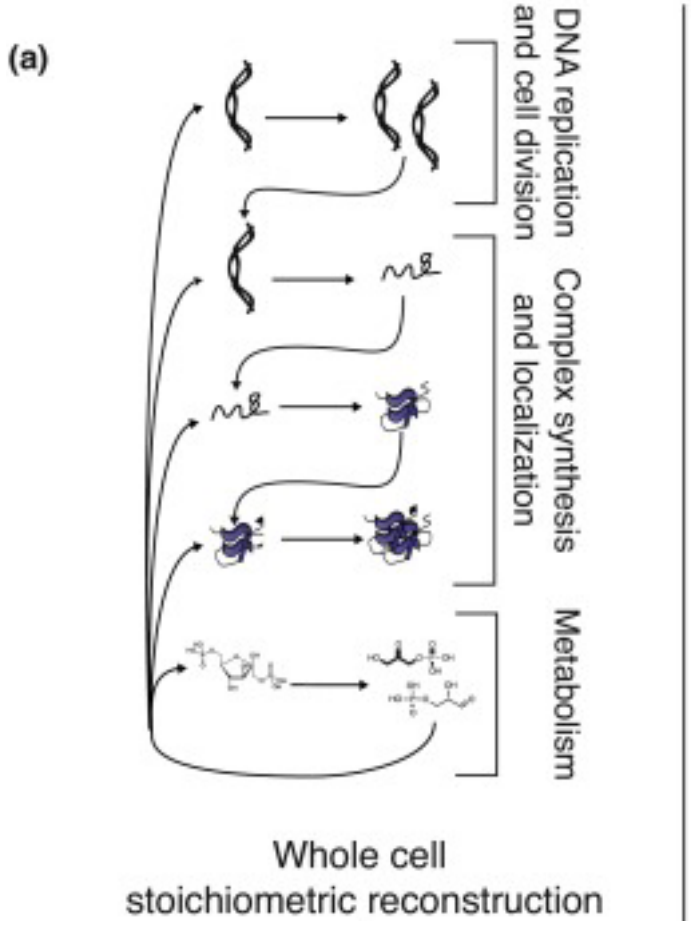
3

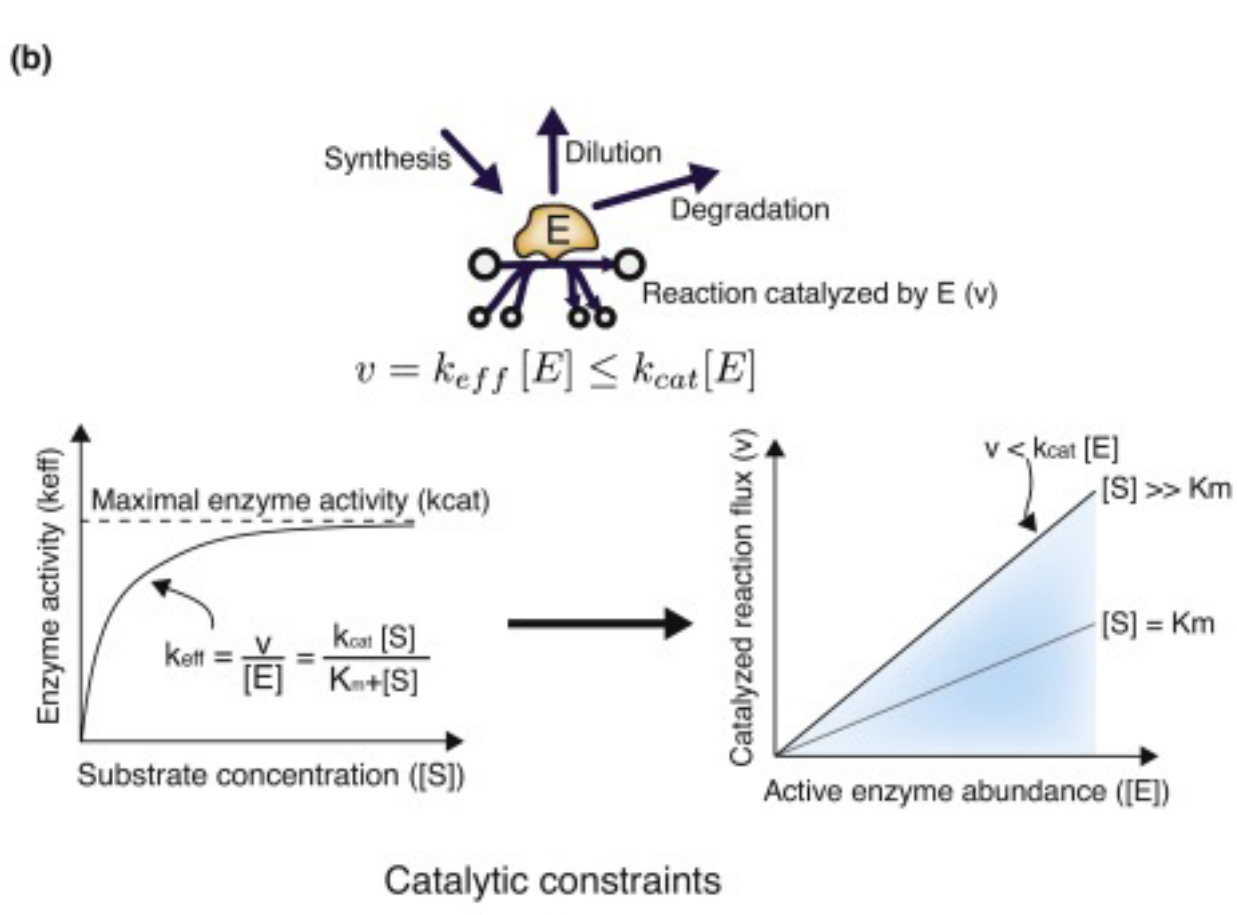
Model protein secretion combined with metabolism



Extended ME model

ME (Metabolism + Expression) models





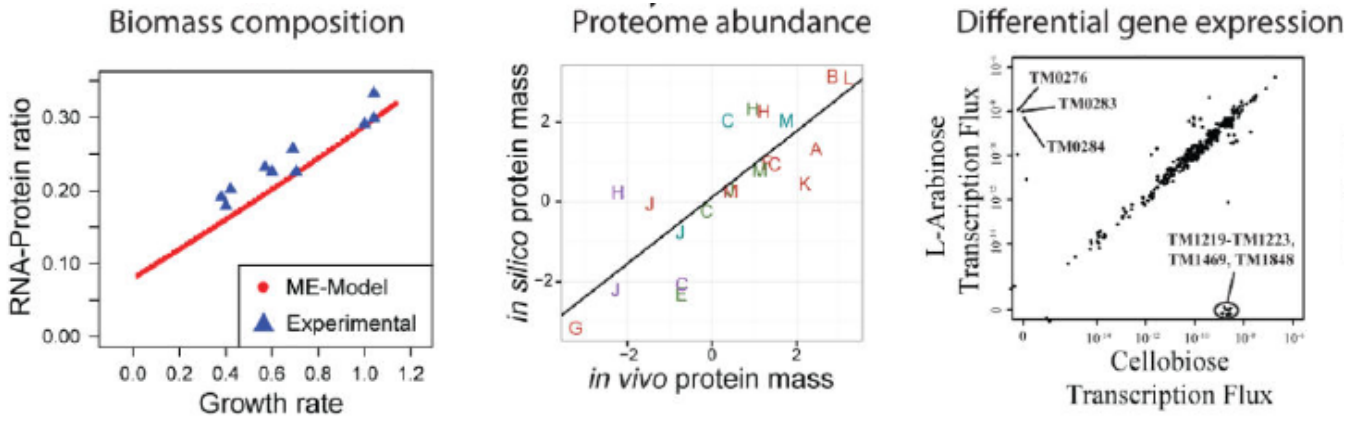
Represent biosynthesis of all transcripts and proteins explicitly in the S matrix

Need to specify coupling parameters (effective kcat) that describe relationship between flux and protein expression

O'Brien et al. Curr Opin Biotechnol 2014

ME model pros and cons

Pros



- No need to specify biomass composition (self consistent)
- Can integrate proteomics (and transcriptomics) data directly
- Can account for cost of expressing heterologous pathways -> improved design predictions

O'Brien et al. Cell 2016

Cons

- Need effective k_{cat}'s
- Much larger and more complex than M-models especially for eukaryotes
- Simulations are computationally expensive
- Strain design calculations are very expensive computationally
- Models currently only available for a few species (*E. coli* and *T. maritima*)

Building a ME model for additional organisms

Model for *Lactococcus lactis* that combines metabolism and expression (ME model)

Components in E model:

- RNA polymerase (5)
- Sigma factor (1)
- Transcription elongation and termination factor (5)
- Ribonuclease (1)
- Degradosome and oligoribonuclease (8)
- Translation initiation factor (3)
- Translation elongation factor (3)
- Release factor (6)
- Ribosomal protein subunit and binding factor (59)
- Aminoacyl-tRNA synthetase (25)
- RNA modification (37)

E model: 153 protein genes

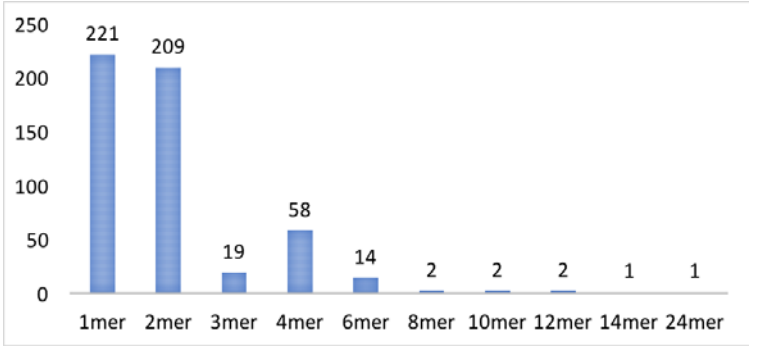
M model: 601 protein genes

ME model: 716 protein genes + 81 RNA genes

Source of information:

1. Gene orthology analysis from ME models from *E. coli* and *T. maritima*
2. Subsystem analysis (RNA and protein metabolism)

Protein stoichiometry:



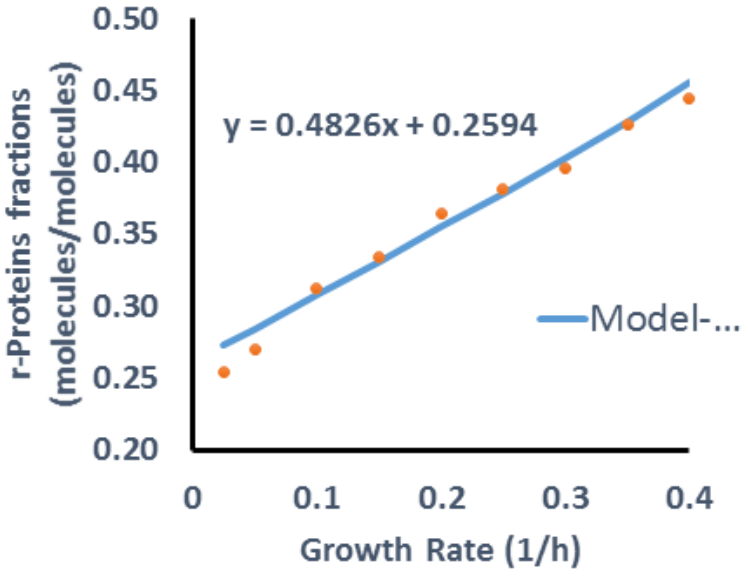
Template reactions:

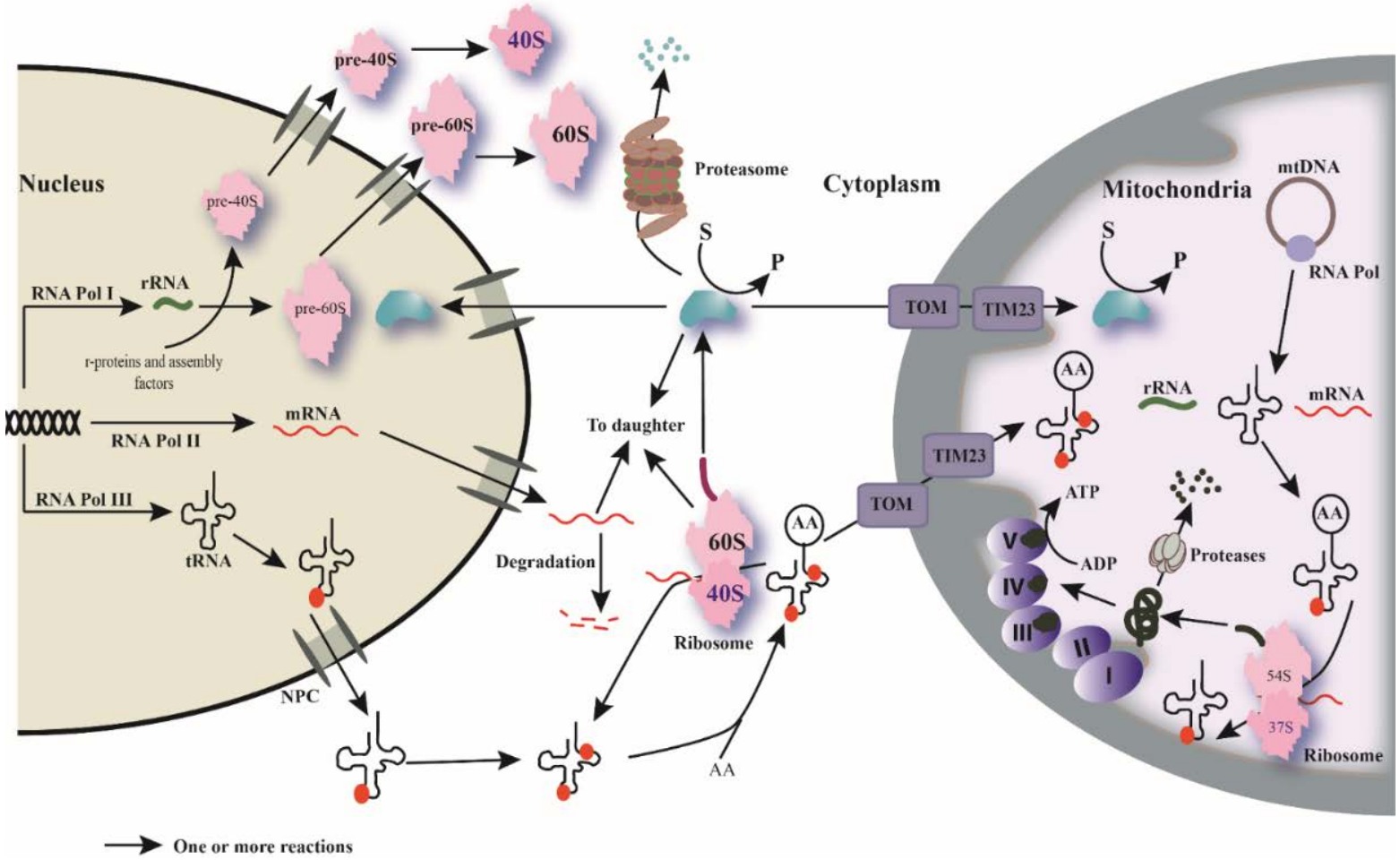
Subprocess	RXN name	RXN formula
Transcription	transcription_initiation_TUCDT-xxxx	1 RNAP_sf[c] + a atp[c] + b ctp[c] + c gtp[c] + d utp[c] => 1 RNAP_TUCDT-xxxx[c] + 1 sf[c] + 15 ppi[c]
Transcription	transcription_binding_rho_dependent_TUCDT-xxxx	1 Transcription_rho_dependent[c] + 1 RNAP_TUCDT-xxxx[c] => 1 Transcription_rho_dependent_RNAP_TUCDT-xxxx[c]
Transcription	transcription_binding_rho_independent_TUCDT-xxxx	1 Transcription_rho_independent[c] + 1 RNAP_TUCDT-xxxx[c] => 1 Transcription_rho_independent_RNAP_TUCDT-xxxx[c]
Transcription	transcription_elongation_rho_dependent_TUCDT-xxxx	1 Transcription_rho_dependent_RNAP_TUCDT-xxxx[c] + (a+3) atp[c] + b ctp[c] + c gtp[c] + d utp[c] + 3 h2o[c] => 1 TUCDT-xx
Transcription	transcription_elongation_rho_independent_TUCDT-xxxx	1 Transcription_rho_independent_RNAP_TUCDT-xxxx[c] + a atp[c] + b ctp[c] + c gtp[c] + d utp[c] => 1 TUCDT-xxxx[c] + 1 ef_1
Transcription	RNAP_sf	1 RNAP[c] + 1 sf[c] => 1 RNAP_sf[c]
Transcription	Transcription_rho_dependent	1 rho_*mer[c] + 1 NusG_*mer[c] + 1 NusA_*mer[c] + 1 GreA_*mer[c] + 1 transcription-repair_coupling_factor_*mer[c] + 1
Transcription	Transcription_rho_independent	1 NusG_*mer[c] + 1 NusA_*mer[c] + 1 GreA_*mer[c] + 1 transcription-repair_coupling_factor_*mer[c] + 1 NusB_*mer[c] =>
Transcription	RNAP	1 llmg_xxxx[c] + ... + 1 llmg_xxxx[c] => 1 RNAP[c]
Cleavage of stable RNA	cleavage_of_TUCDT-xxxx	1 TUCDT-xxxx[c] + a Ribonuclease_A_*mer_primed[c] + ... + a Ribonuclease_Z_*mer_primed[c] + b h2o[c] => 1 rnaxx_unmo
Cleavage of stable RNA	Ribonuclease_A_*mer_primed[c]	1 Ribonuclease_A_*mer[c] => 1 Ribonuclease_A_*mer_primed[c]
rRNA modification	rRNA_modification_rnaxx_1_binding	1 rnaxx_unmodified[c] + 1 llmg_xxxx_*mer[c] + 1 chemical_a[c] => 1 rnaxx_1_bound[c]

ME models for eukaryotes

ME model for *S. cerevisiae*

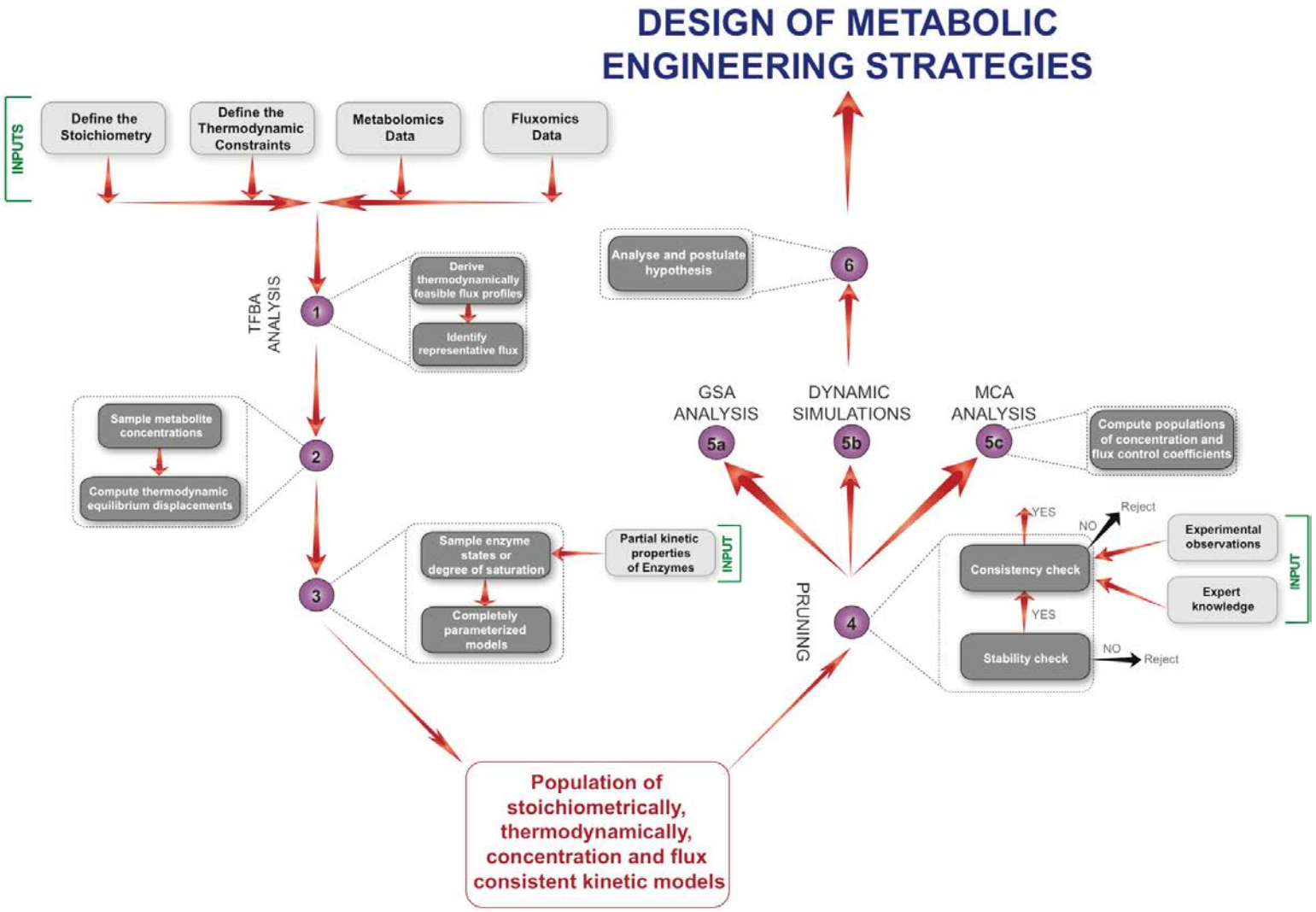
Model can describe growth and proteome levels correctly, e.g. ribosome level





Metabolomics data

- Metabolomics data can be used to identify thermodynamically feasible reaction directions with M models (Thermodynamic FBA, TFBA)
- Full integration of metabolomics data requires using kinetic models
- Requires reducing genome-scale metabolic models to medium scale
- Medium-scale kinetic models are currently available only for *E. coli*



Wang et al., Biophys J, 87, 3750 (2004)
Miskovic & Hatzimanikatis, B&B, 108, 2, 413(2011)

What will be available through DD-DeCaF project?

- M, GECKO, ME and large-scale kinetic models with common identifiers
- Ability to upload data and integrate with the models
- Range of organisms supported from within the project:
  - ***Escherichia coli***
  - ***Saccharomyces cerevisiae***
  - *Kluveromyces marxianus*
  - *Yarrowia lipolytica*
  - *Pseudomonas putida*
  - *Lactococcus lactis*
  - *Bacillus subtilis*
- Models for additional organisms can be added to the platform assuming that model quality checks are passed