Model SEED Tutorial Part 1: Technical Primer

Christopher Henry, Scott Devoid, Matt DeJongh, Aaron Best, Ross Overbeek, and Rick Stevens

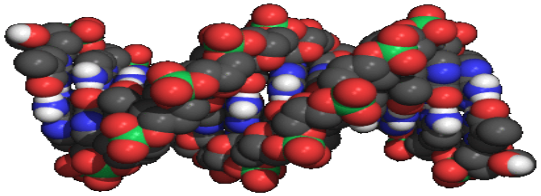
Presented by: Christopher Henry

August 31-September 2

Fellowship for Interpretation of Genomes

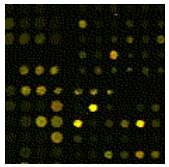
HOPE COLLEGE

Predicting Phenotypes from Genotypes **—the prediction of system level behavior from collections of functional components**



DNA (storage)

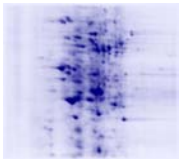
Transcription



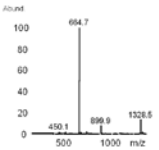
Translation

Proteins

Gene Expression

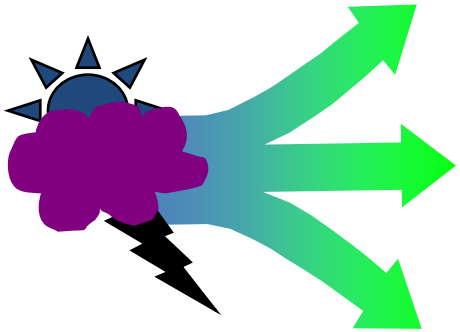


Proteomics



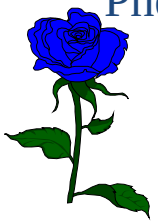
Biochemical Circuitry

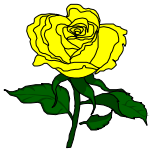
Metabolomics

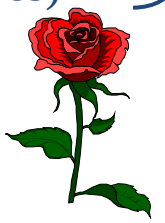


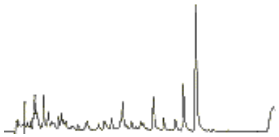
Environment

Phenotypes (Organism Properties)









Adapted From Bruno Sobral VBI

Genome + Environment = Phenotype

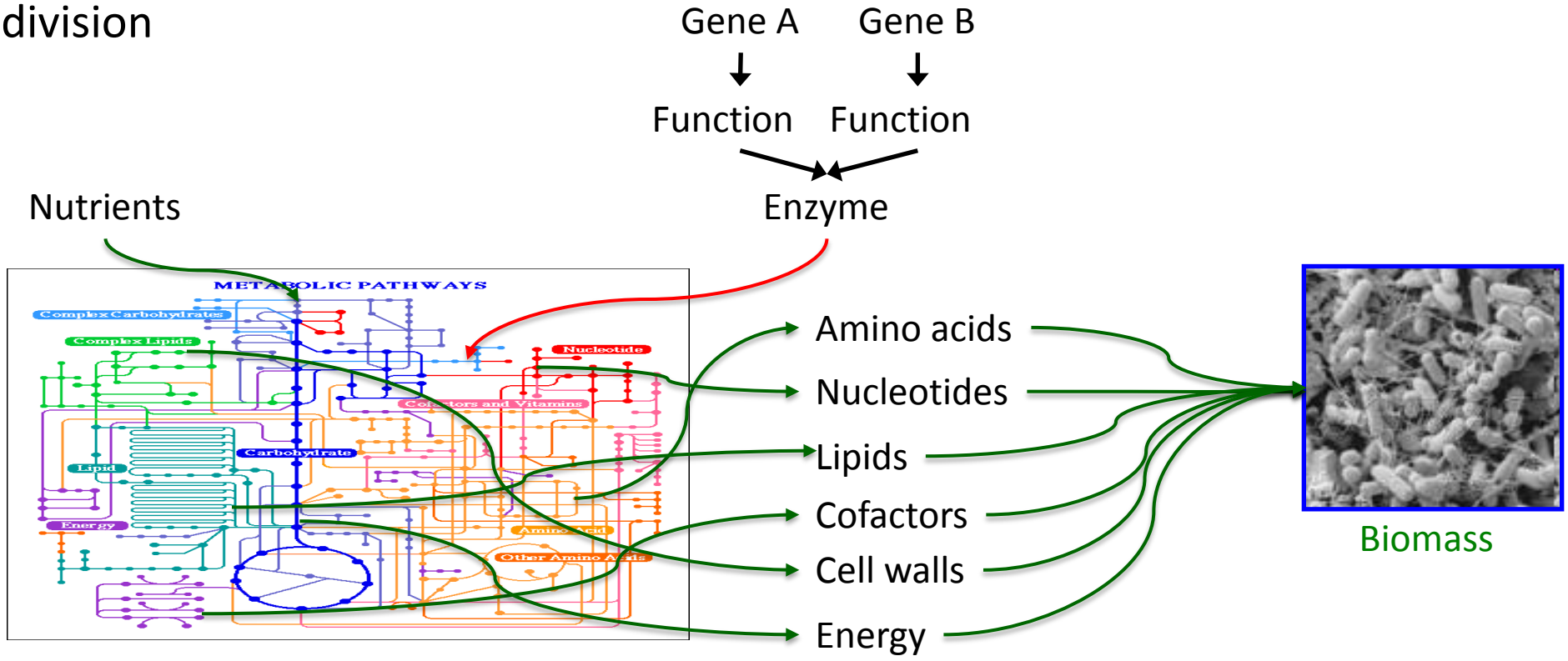
Metabolic Modeling is One Key to Predicting Phenotype from Genotype

What is a metabolic model?

1.) A list of all reactions involved in the metabolic pathways

2.) A list of rules associating reaction activity to gene activity

3.) A biomass reaction listing essential building blocks needed for growth and division



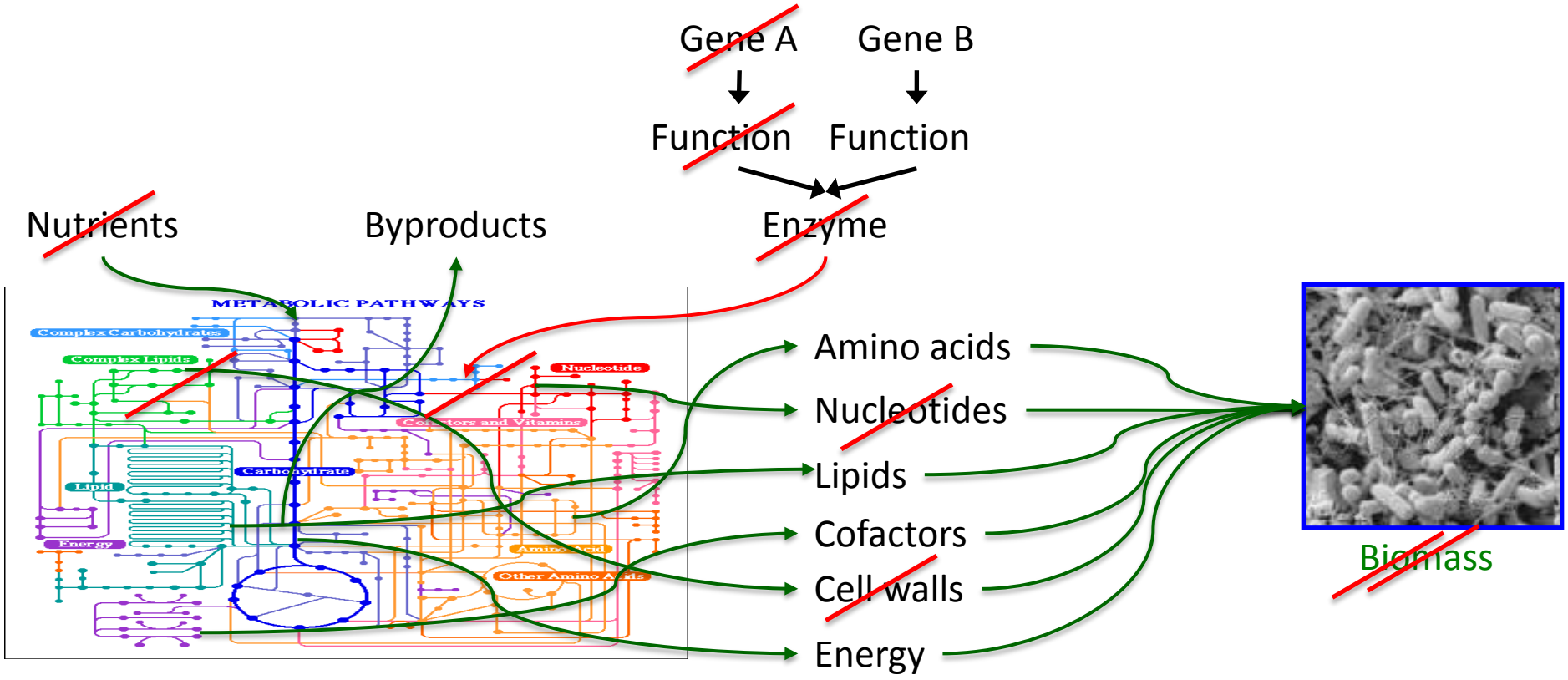
Metabolic Modeling is One Key to Predicting Phenotype from Genotype

What can a metabolic model do?

1.) Predict culture conditions and possible responses to environment changes.

2.) Predict metabolic capabilities from genotype.

3.) Predict impact of genetic perturbations



Metabolic Modeling is One Key to Predicting Phenotype from Genotype

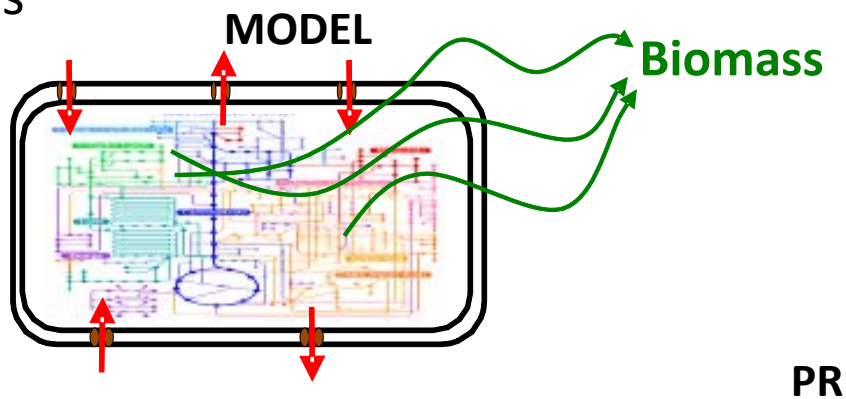
What can a metabolic model do?

1.) Predict culture conditions and possible responses to environment changes.

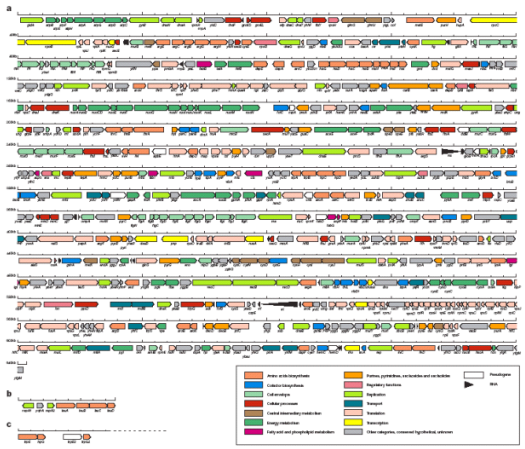
2.) Predict metabolic capabilities from genotype.

3.) Predict impact of genetic perturbations

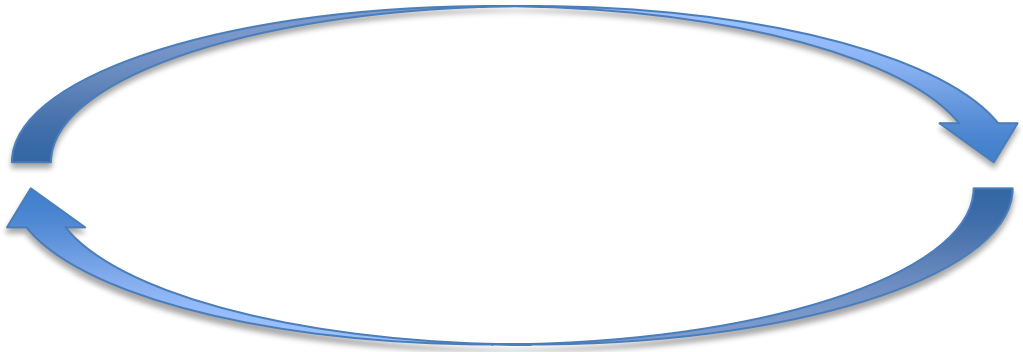
4.) Linking annotations to observed organism behavior enabling validation and correction of annotations



ANNOTATION



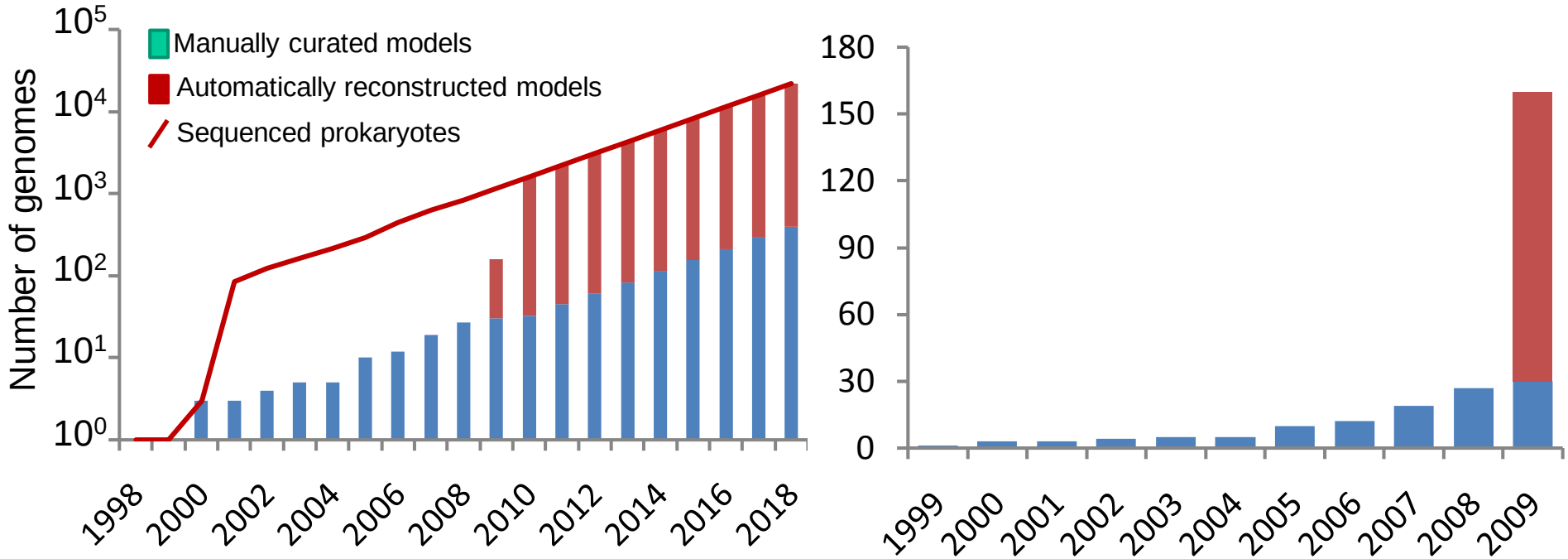
PREDICTION



RECONCILIATION

PHENOTYPE

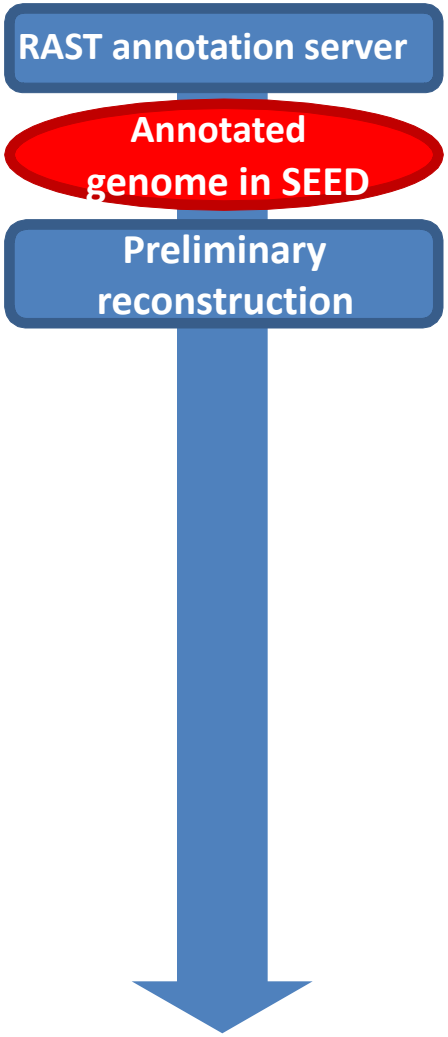
Model reconstruction lags behind genome sequencing



- ≈**1000** completely sequenced prokaryotes vs ≈**30** published genome-scale models
- Models are often constructed one-at-a-time by individuals working independently
- Model building typically begins by identifying bidirectional best hits with *E. coli*
- Current process results in replication of work, propagation of errors, and extensive manual curation
- Bottom line: it previously required approximately one year to produce a complete model

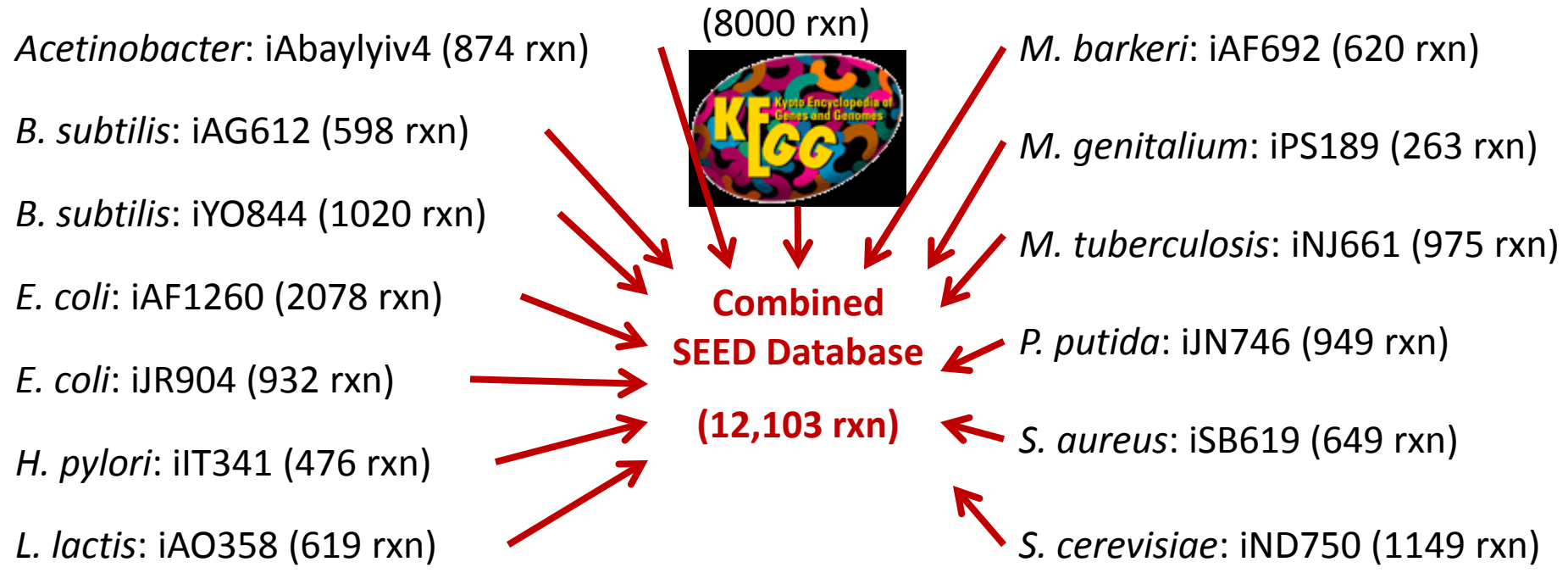
www.theseed.org/models/

Model SEED: Converting Annotated Genomes into Genome-scale Metabolic Models



Biochemistry Database in the SEED

•A biochemistry database was constructed combining content from the **KEGG** and **13** published genome-scale models into a non-redundant set of compounds and reactions



•Reactions were then mapped to the functional roles in the SEED based on EC number, substrate names, and enzyme names:

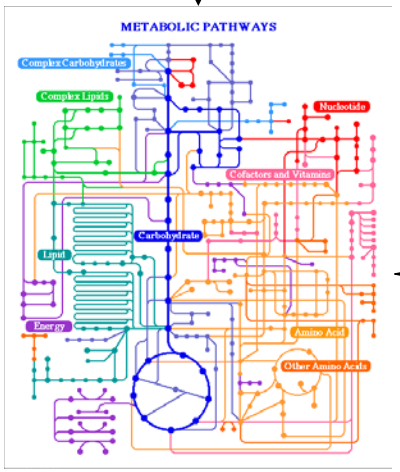
REACTION	COMPLEX	FUNCTIONAL ROLE	GENE
$NAD^+ + NADPH \leftrightarrow NADH + NADP^+$	Gene complex	NAD(P) transhydrogenase subunit beta (EC 1.6.1.2)	peg.100
		NAD(P) transhydrogenase alpha subunit (EC 1.6.1.2)	peg.101

www.theseed.org/models/

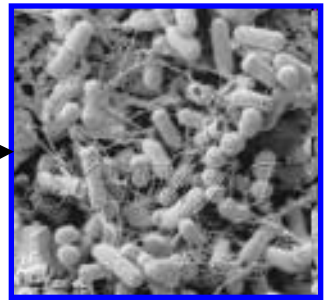
Biomass Objective Function

- To test growth of the model, we build a biomass objective function template

Nutrients



ATP+H2O→ADP+Pi	Universal	Energy
dATP, dGTP, dCTP, dTTP	Universal	DNA
ATP, GTP, CTP, UTP	Universal	RNA
Amino acids	Universal	Protein
Cofactors and ions	Depends on genome	Misc
Various acylglycerols	Depends on genome	Lipids
Peptioglycan	Any genome with cell wall	Cell wall
Teichoic acid	Gram positive	Cell wall
Core lipid A	Gram negative	Cell wall

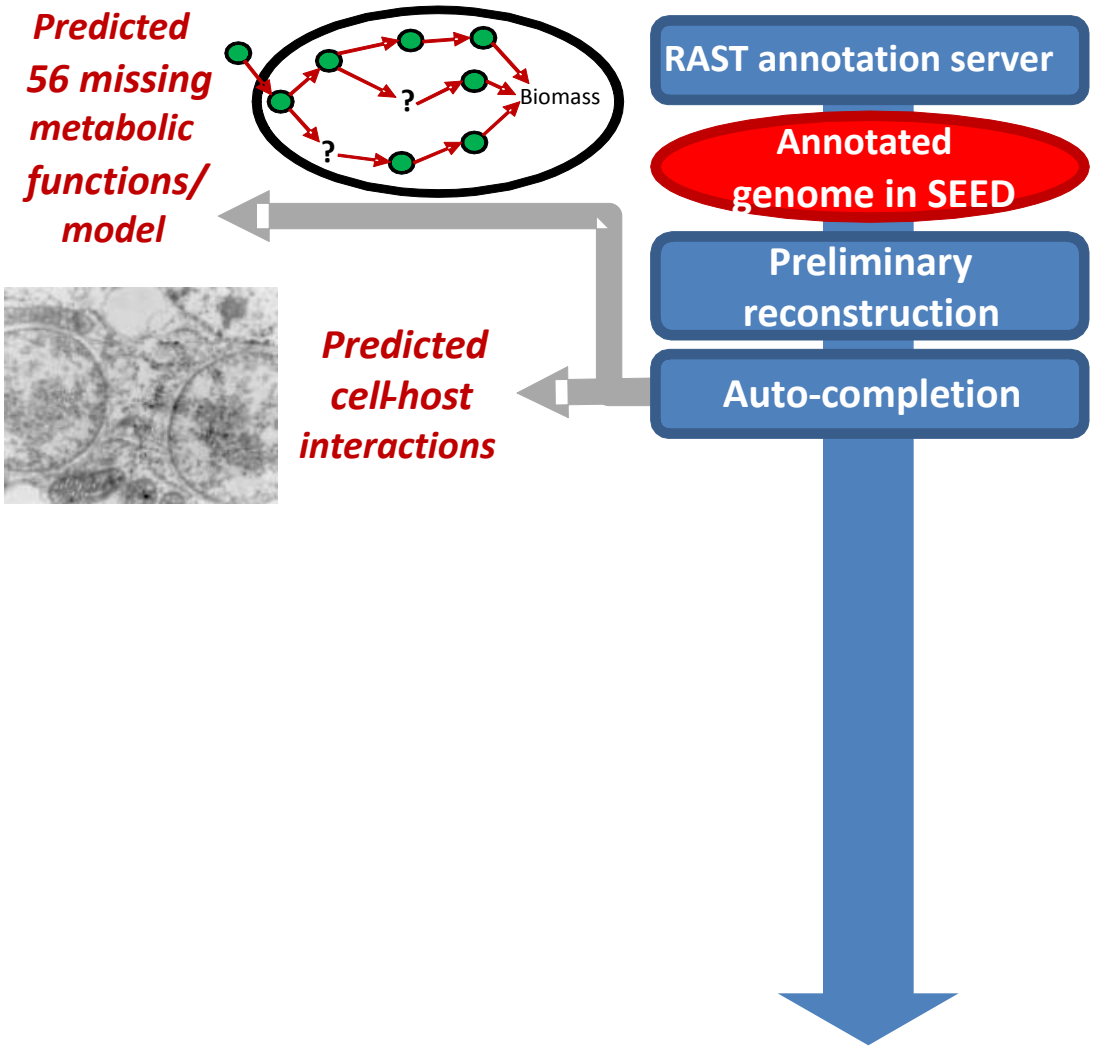


Biomass

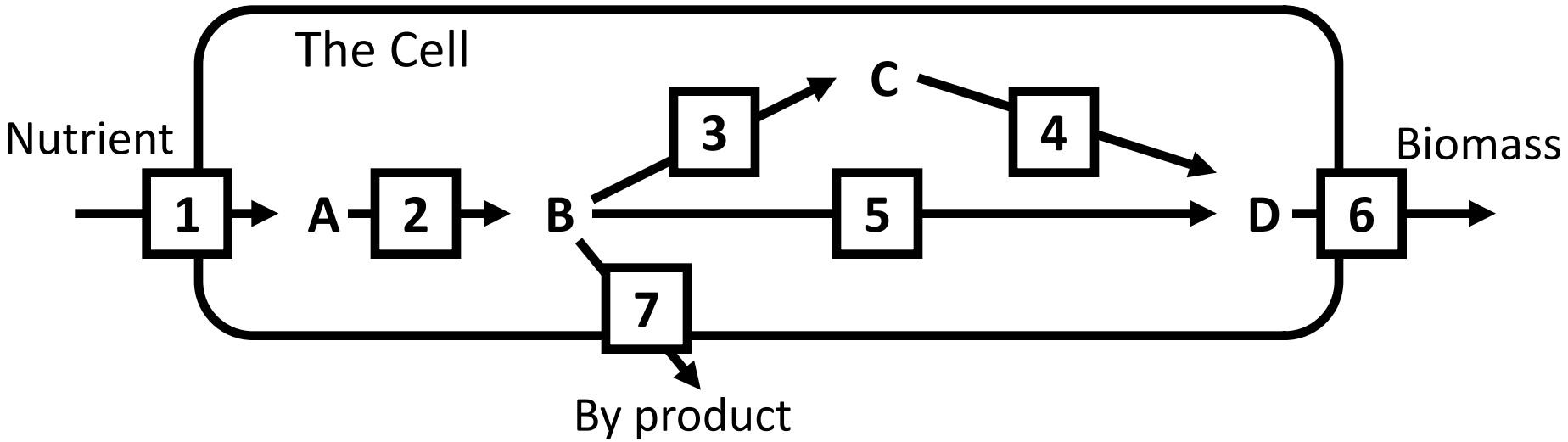
- Each biomass component may be rejected from the biomass reaction of a model based on the following criteria:
  - Subsystem representation
  - Functional role presence
  - Taxonomy
  - Cell wall types

www.theseed.org/models/

Model SEED: Converting Annotated Genomes into Genome-scale Metabolic Models



Flux Balance Analysis



Assuming Steady State:

No internal metabolite is allowed to accumulate

Thus, reaction rates are constrained by mass balances

For example:

At Steady State:

$v_1 = v_2$

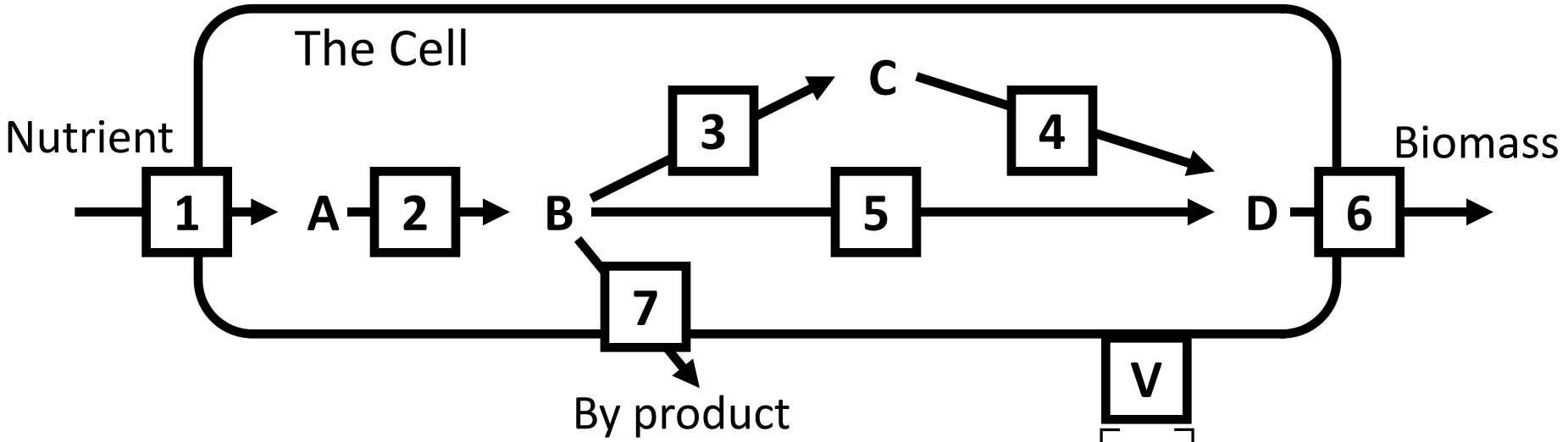
$v_2 = v_3 + v_5 + v_7$

$v_3 = v_4$

$v_4 + v_5 = v_6$

www.theseed.org/models/

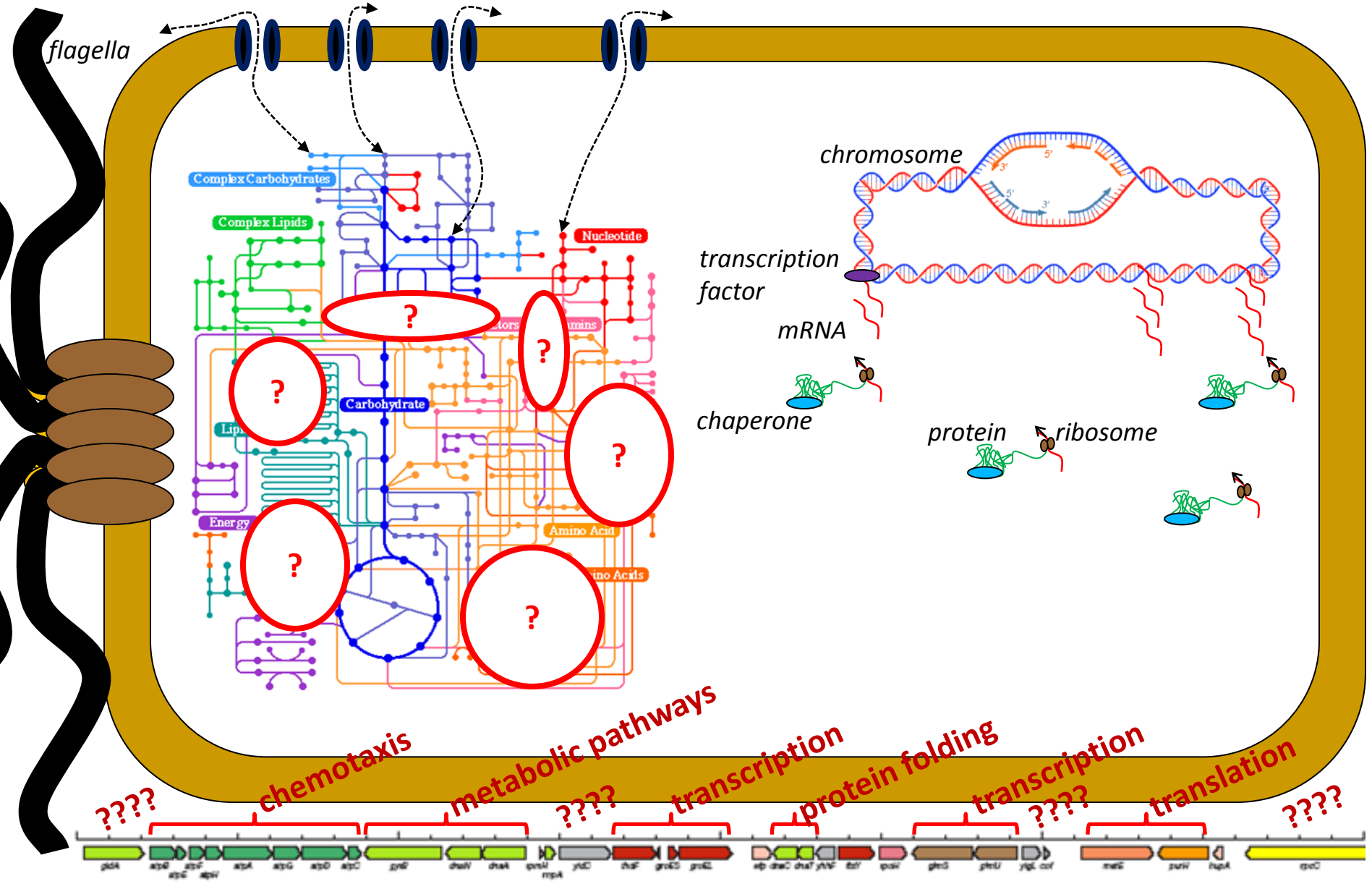
Flux Balance Analysis



$$
\begin{array}{c} \\ A \\ B \\ C \\ D \end{array}
\begin{array}{c}
\begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\
\begin{bmatrix}
1 & -1 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & -1 & 0 & -1 & 0 & -1 \\
0 & 0 & 1 & -1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 & -1 & 0
\end{bmatrix}
\end{array}
\bullet
\begin{array}{c} V \\ \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \\ v_7 \end{bmatrix} \end{array}
=
\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}
$$

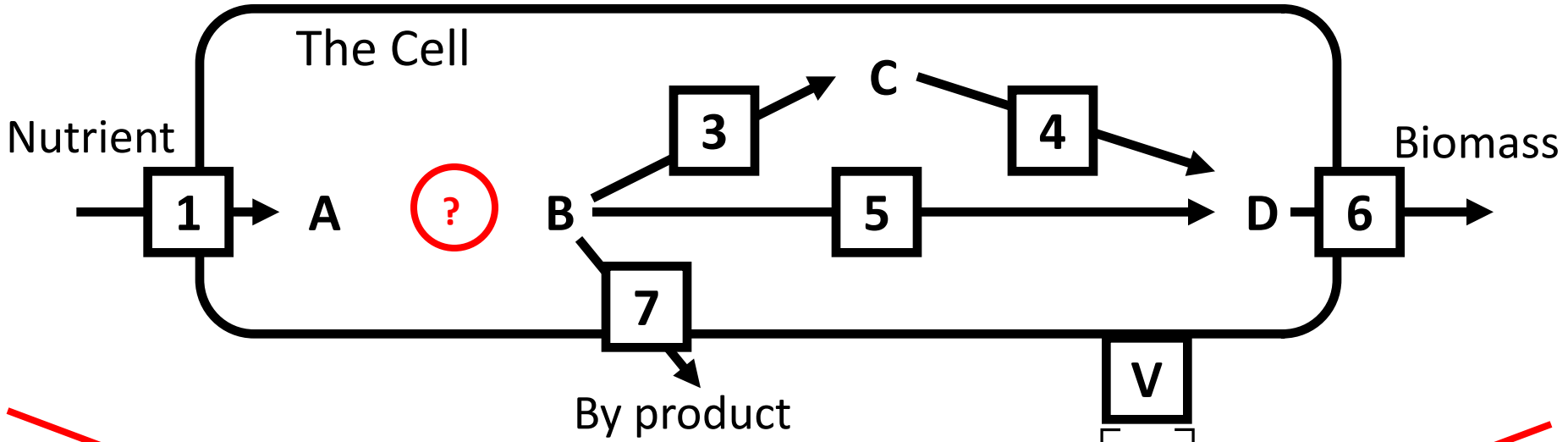
www.theseed.org/models/

Genome Annotations Contain Knowledge Gaps



www.theseed.org/models/

Flux Balance Analysis



$$
\begin{array}{c} \\ A \\ B \\ C \\ D \end{array}
\begin{array}{c}
\begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\
\begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & -1 & 0 & -1 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & -1 & 0 \end{bmatrix}
\end{array}
\bullet
\begin{array}{c} V \\ \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \\ v_7 \end{bmatrix} \end{array}
=
\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}
$$

www.theseed.org/models/

Model Auto-completion Optimization

Objective:

Penalizing reversibility adjustments

$$\text{Minimize} \sum_{i=1}^{r} \left[z_{i,forward,\text{not in model}} + z_{i,reverse,\text{reversible not in model}} + 2z_{i,reverse,\text{irreversible not in model}} + z_{i,reverse,\text{irreversible in model}} \right]$$

Penalizing addition of reactions to the model

Subject to:

Mass balance constraints:

$$\begin{array}{c} \\ \text{Compounds in model} \\ \text{Compounds not in model} \end{array} \begin{array}{c} \begin{array}{cc} \text{Reactions in model} & \text{Reactions not in model} \end{array} \\ \begin{bmatrix} \boldsymbol{N_{core}} & \boldsymbol{N_{db}} \\ \boldsymbol{0} & \boldsymbol{N_{db}} \end{bmatrix} \end{array} \bullet \begin{bmatrix} \boldsymbol{v_{core}} \\ \boldsymbol{v_{db}} \end{bmatrix} = 0$$

Use variable constraints:

$$0 \le v_{i,forward} \le v_{Max} z_{i,forward}$$

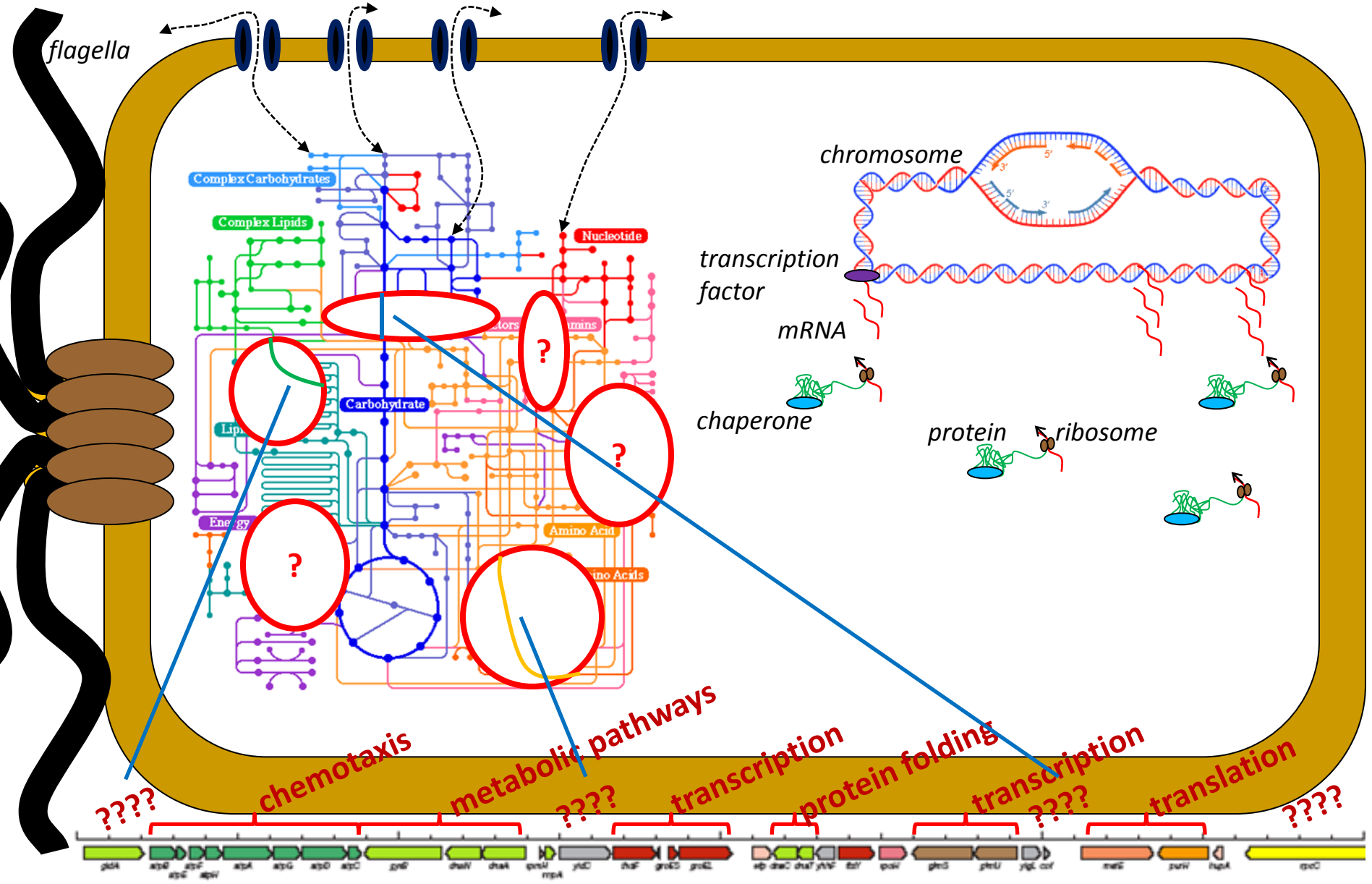
$$0 \le v_{i,reverse} \le v_{Max} z_{i,reverse}$$

Forcing positive growth:

$$v_{biomass} > 0$$

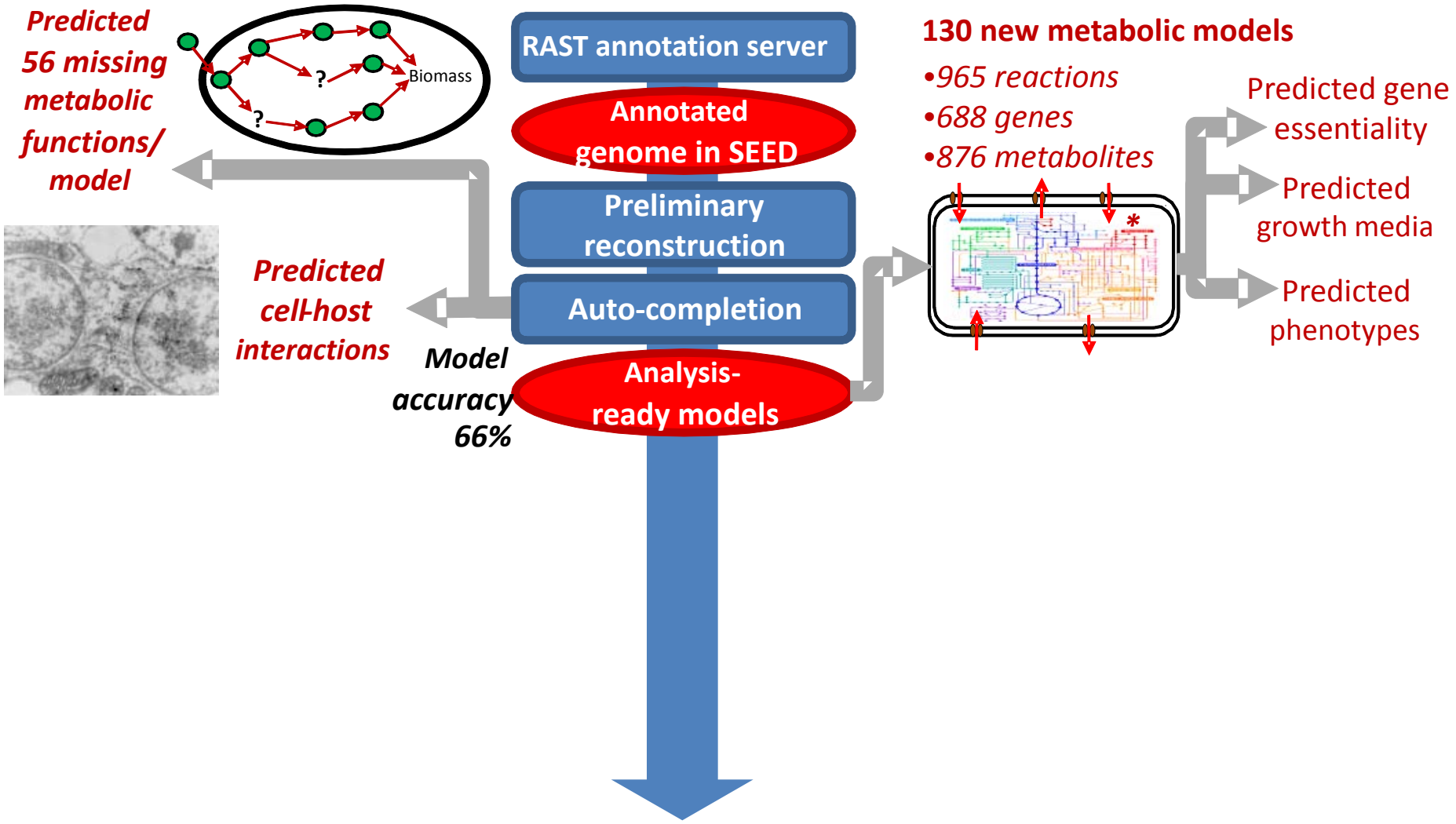
www.theseed.org/models/

Genome Annotation: the Subsystems Approach



www.theseed.org/models/

Model SEED: Converting Annotated Genomes into Genome-scale Metabolic Models

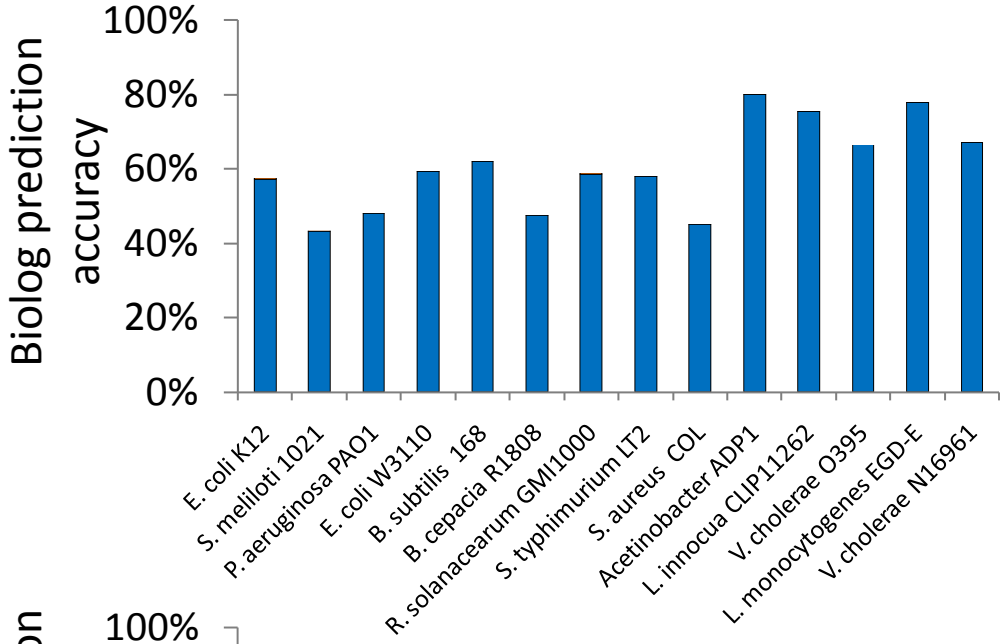


Accuracy Before Optimization

Biolog phenotype data

•SEED models were used to predict the output of 14 biolog phenotyping arrays

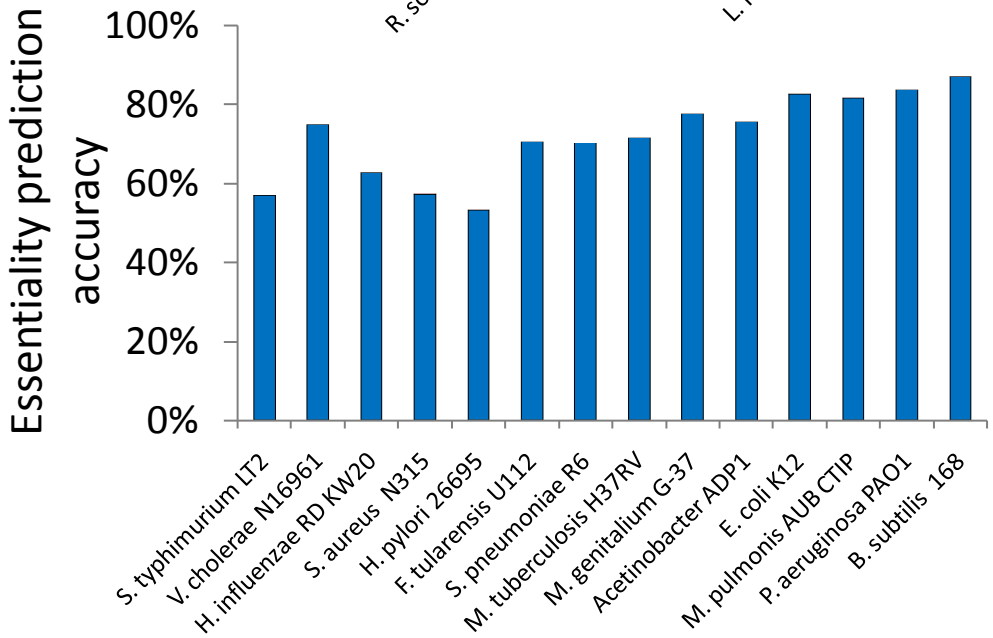
•Average accuracy: 60%



Essentiality data

•SEED models were used to predict essential genes for 14 experimental gene essentiality datasets

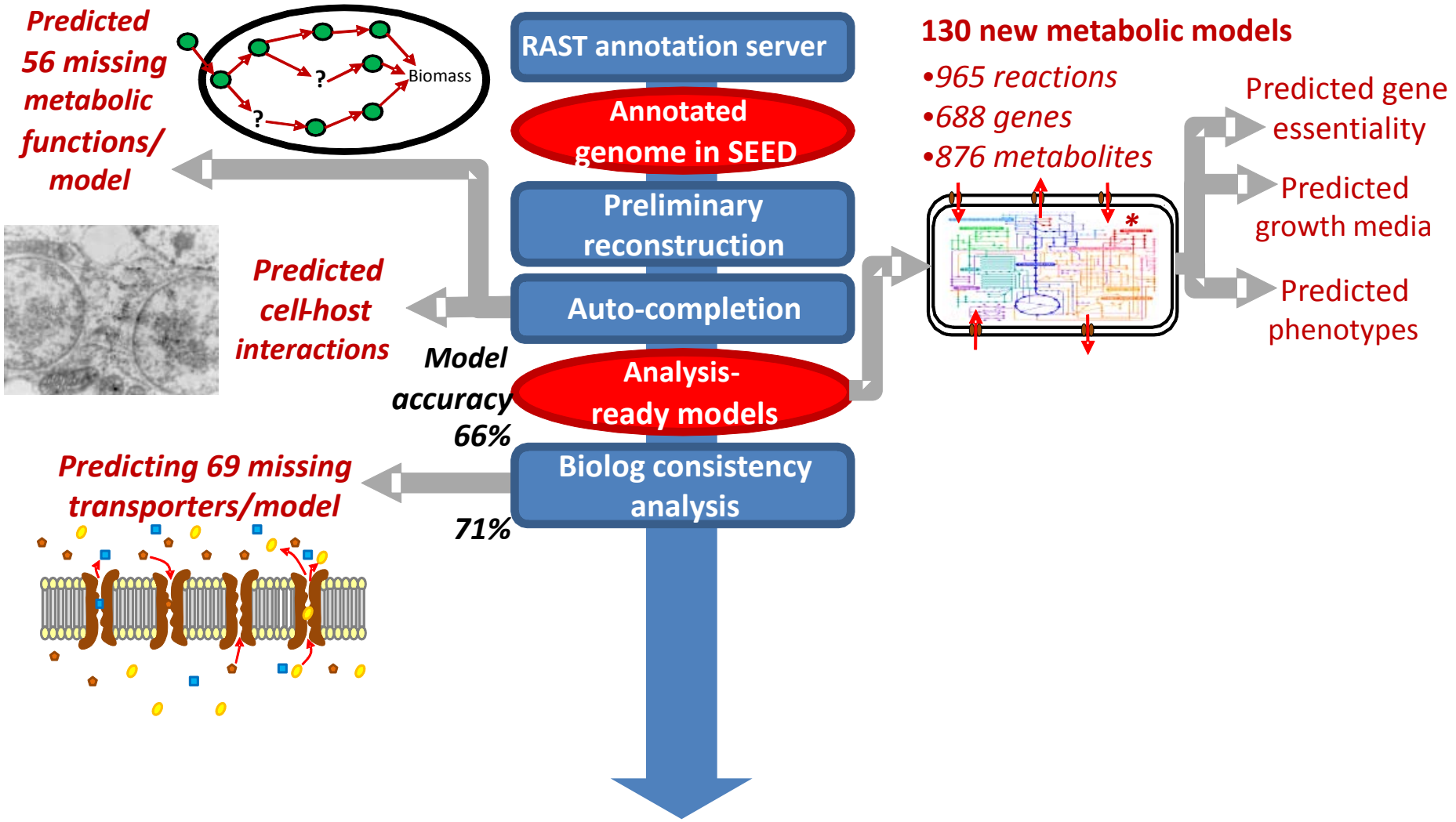
•Average accuracy: 72%



Overall accuracy: 66%

www.theseed.org/models/

Model SEED: Converting Annotated Genomes into Genome-scale Metabolic Models



Biolog Consistency Analysis

Biolog phenotype data

•Add transporters for Biolog nutrients if missing from models

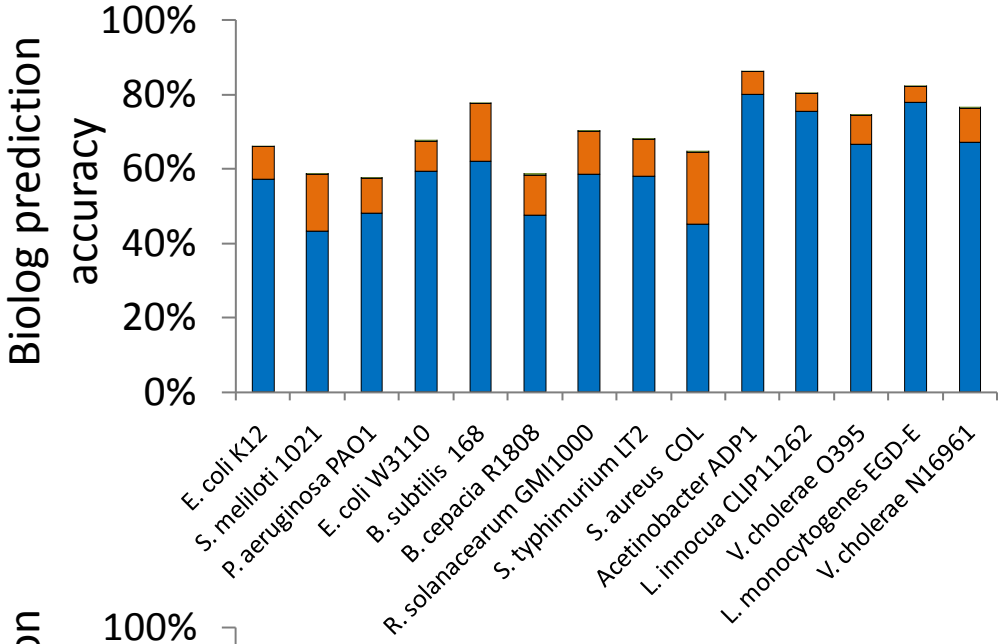
•69 transporters added to each model on average

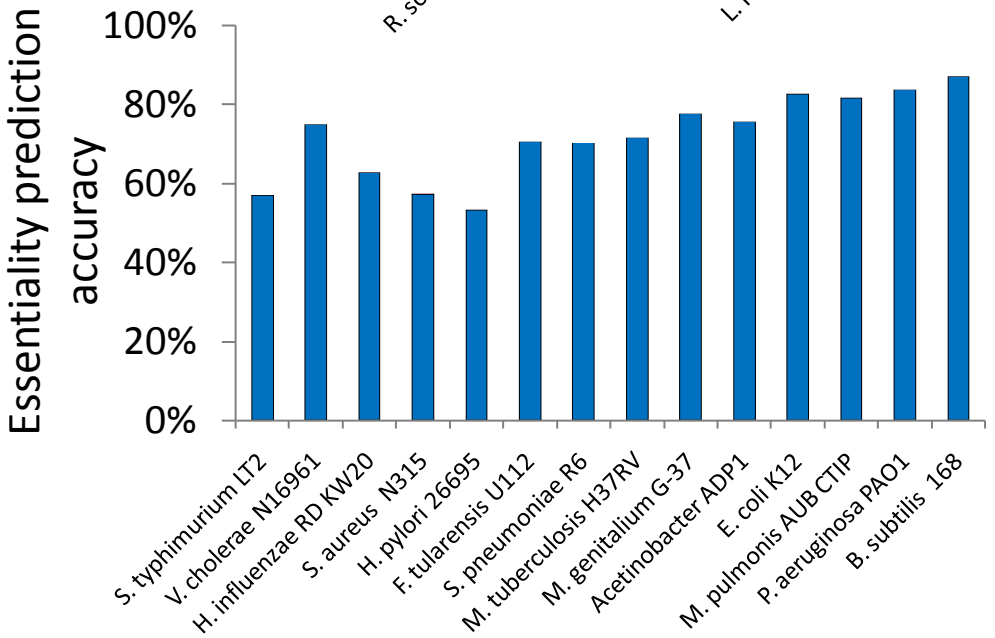
•Average accuracy: 70%

Essentiality data

•Accuracy unchanged: 72%

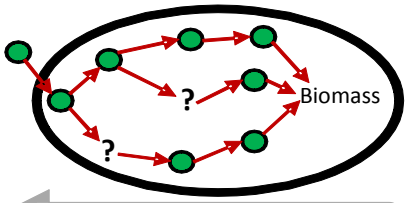
Overall accuracy: 71%



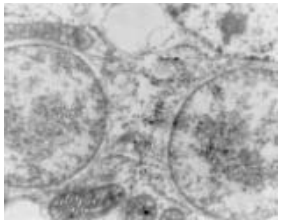


www.theseed.org/models/

Model SEED: Converting Annotated Genomes into Genome-scale Metabolic Models

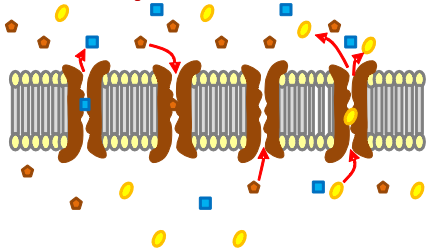


Predicted 56 missing metabolic functions/ model



Predicted cell-host interactions

Predicting 69 missing transporters/model



RAST annotation server

Annotated genome in SEED

Preliminary reconstruction

Auto-completion

Analysis-ready models

Model accuracy

66%

Biolog consistency analysis

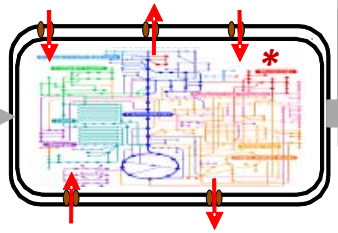
71%

Gene essentiality consistency analysis

74%

130 new metabolic models

- *965 reactions*
- *688 genes*
- *876 metabolites*

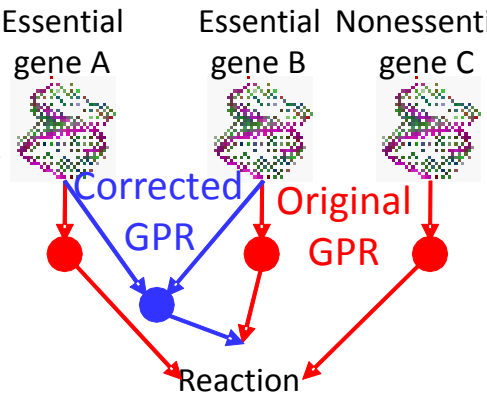


Predicted gene essentiality

Predicted growth media

Predicted phenotypes

Correction for 202 annotations inconsistent with essentiality data

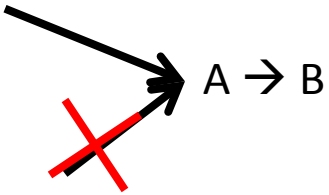


Annotation Consistency Analysis

Essentiality data

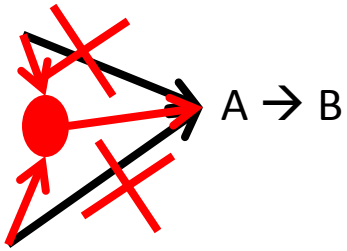
•Reconciling annotation inconsistent with essentiality data

Essential gene



Nonessential gene

Essential gene A



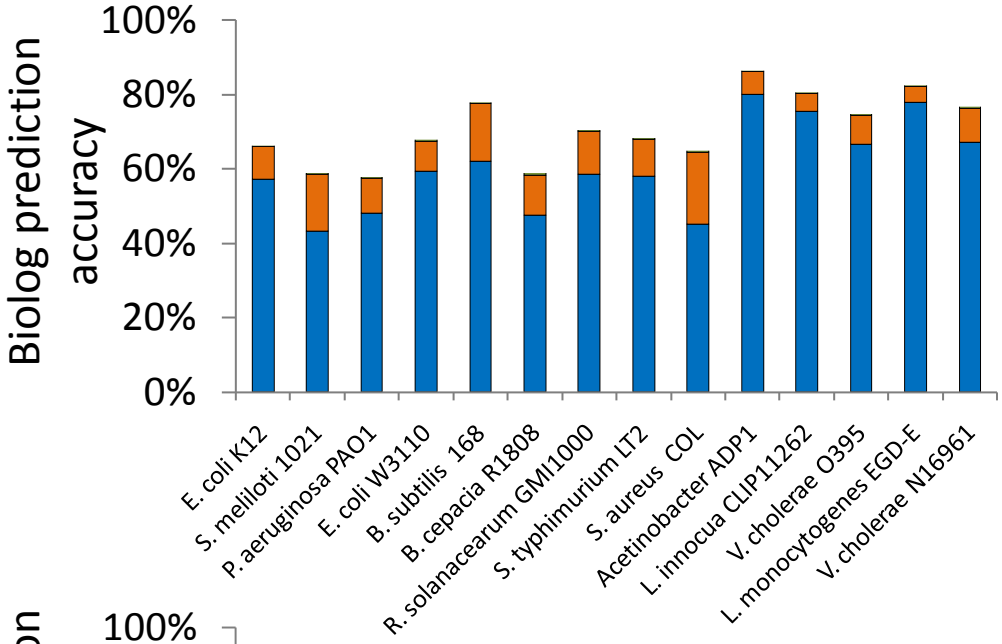
Essential gene B

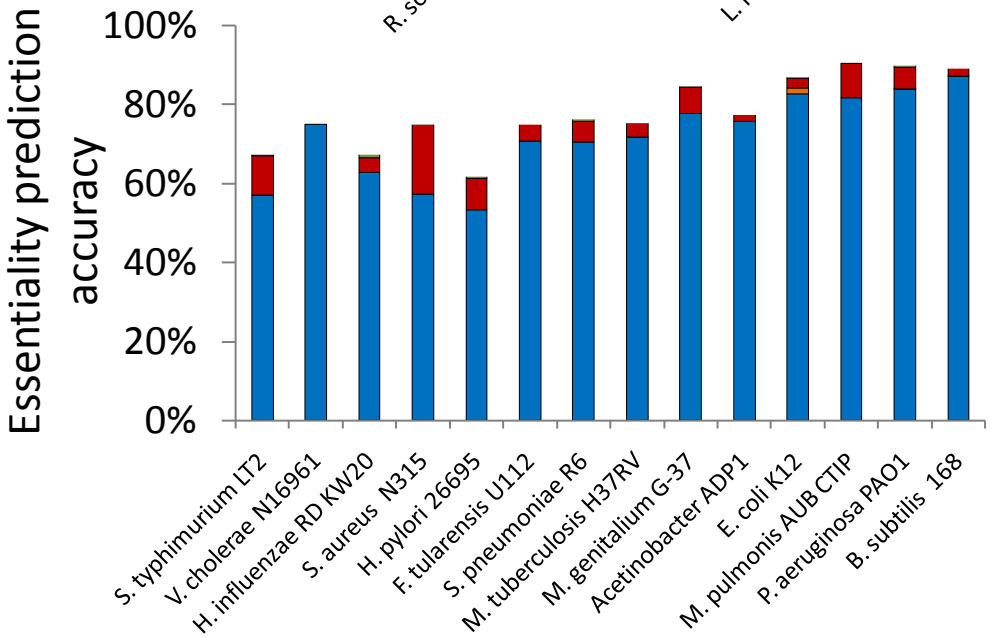
•Accuracy 78%

Biolog phenotype data

•Accuracy unchanged: 70%

Overall accuracy: 75%





www.theseed.org/models/

Model SEED: Converting Annotated Genomes into Genome-scale Metabolic Models

RAST annotation server

Annotated genome in SEED

Preliminary reconstruction

Auto-completion

Analysis-ready models

Biolog consistency analysis

71%

Gene essentiality consistency analysis

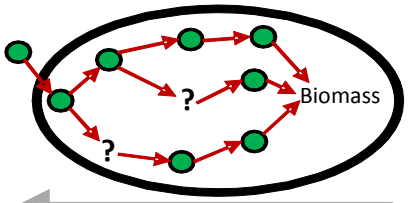
74%

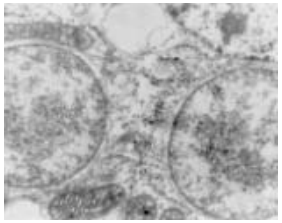
Model opt: GapFill

82%

Model accuracy 66%

Predicted 56 missing metabolic functions/ model

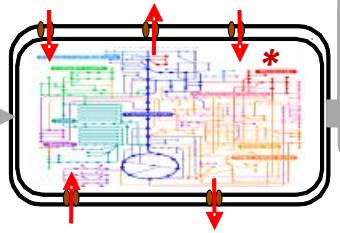




Predicted cell-host interactions

130 new metabolic models

- *965 reactions*
- *688 genes*
- *876 metabolites*

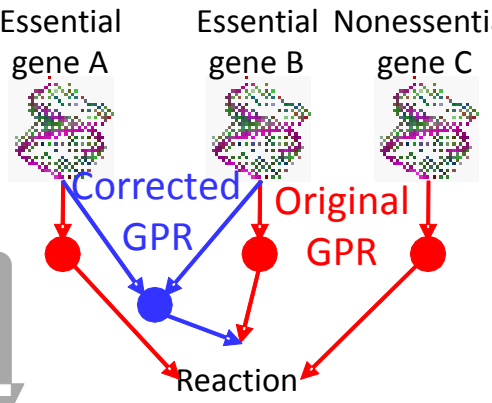


Predicted gene essentiality

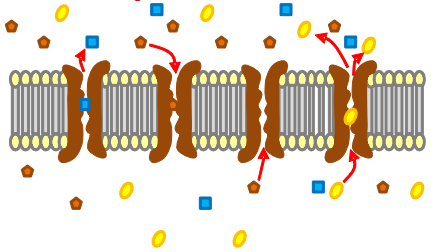
Predicted growth media

Predicted phenotypes

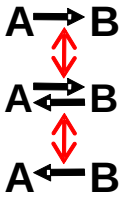
Correction for 202 annotations inconsistent with essentiality data



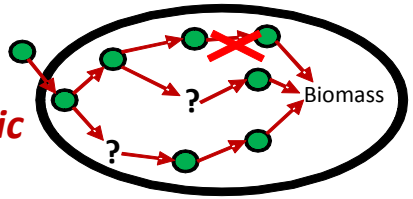
Predicting 69 missing transporters/model



Correcting reversibility constraints

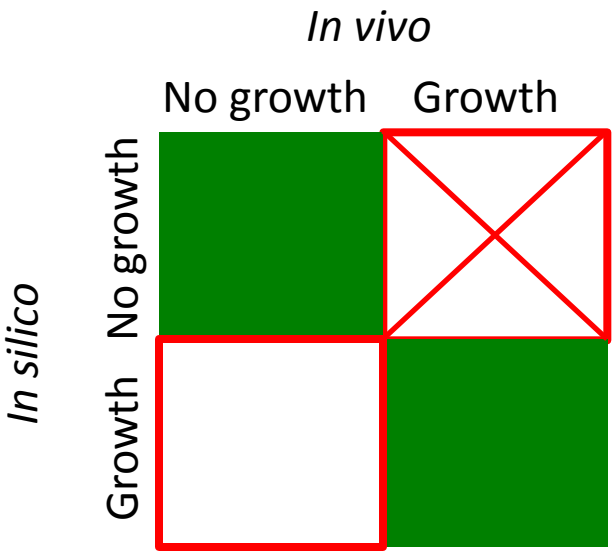


Predicted missing and extra metabolic functions



Model Optimization: Gap Filling

Additional gap filling:



•Fix false negative predictions by adding reactions to models

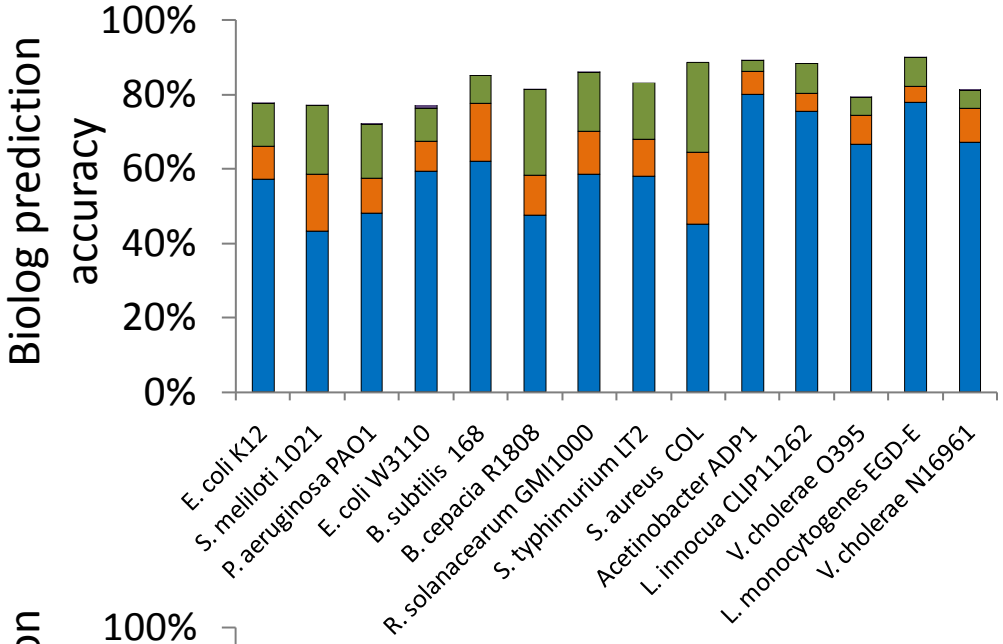
Biolog accuracy

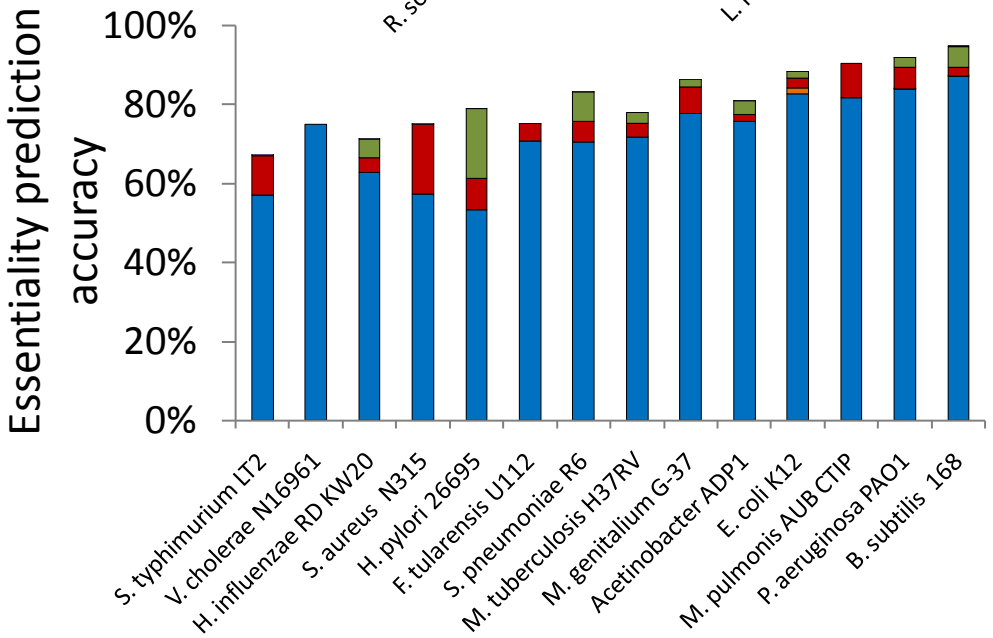
•Average accuracy: 83%

Essentiality accuracy

•Average accuracy: 81%

Overall accuracy: 82%





www.theseed.org/models/

Model SEED: Converting Annotated Genomes into Genome-scale Metabolic Models

RAST annotation server

Annotated genome in SEED

Preliminary reconstruction

Auto-completion

Analysis-ready models

Biolog consistency analysis

Gene essentiality consistency analysis

Model opt: GapFill

Model opt: GapGen

Model accuracy

66%

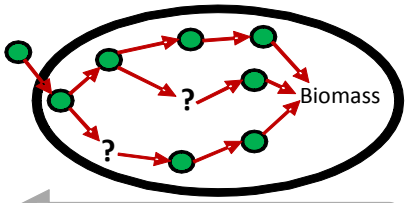
71%

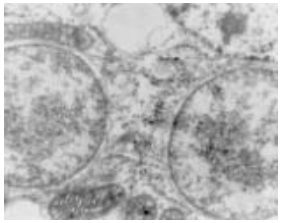
74%

82%

87%

Predicted 56 missing metabolic functions/ model

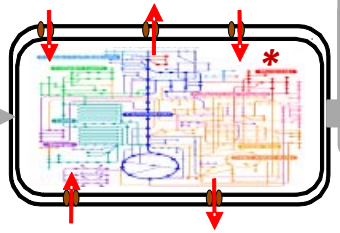




Predicted cell-host interactions

130 new metabolic models

- *965 reactions*
- *688 genes*
- *876 metabolites*

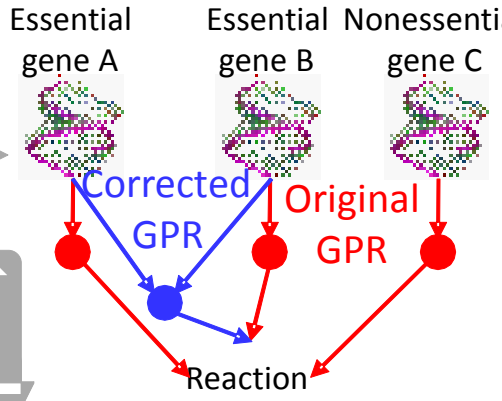


Predicted gene essentiality

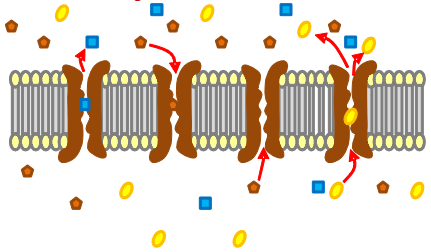
Predicted growth media

Predicted phenotypes

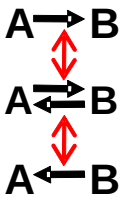
Correction for 202 annotations inconsistent with essentiality data



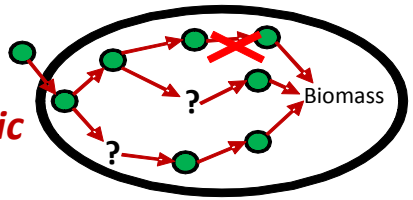
Predicting 69 missing transporters/model



Correcting reversibility constraints

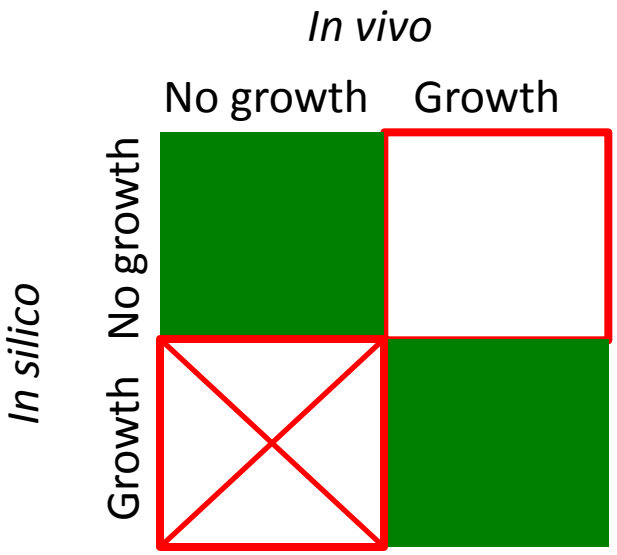


Predicted missing and extra metabolic functions



Model Optimization: Gap Generation

Additional gap filling:



•Fix false positive predictions by removing reactions from models

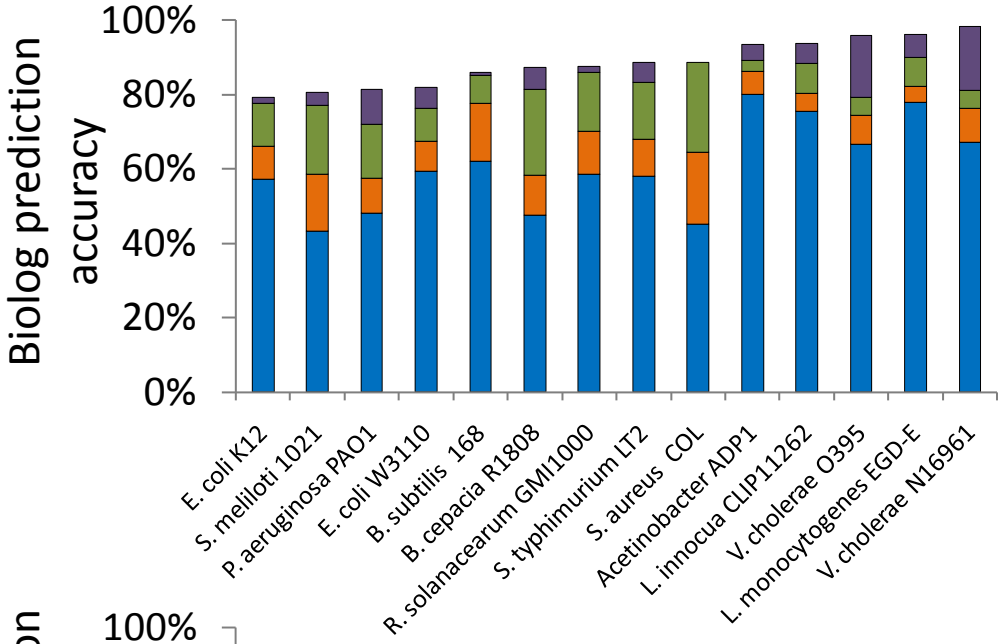
Biolog accuracy

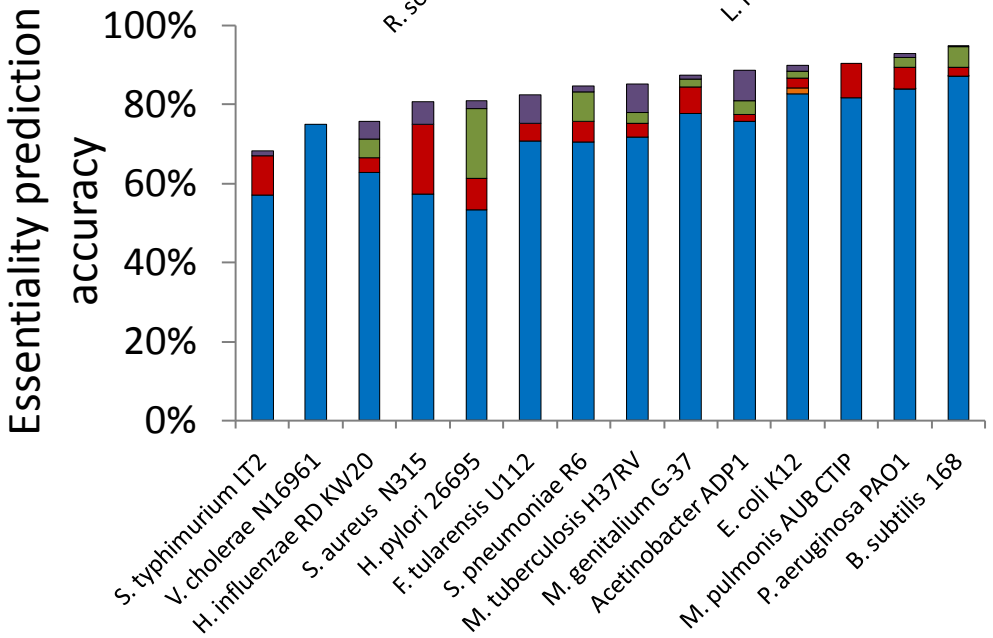
•Average accuracy: 88%

Essentiality accuracy

•Average accuracy: 85%

Overall accuracy: 87%





www.theseed.org/models/

Model SEED: Converting Annotated Genomes into Genome-scale Metabolic Models

RAST annotation server

Annotated genome in SEED

Preliminary reconstruction

Auto-completion

Analysis-ready models

Biolog consistency analysis

71%

Gene essentiality consistency analysis

74%

Model opt: GapFill

82%

Model opt: GapGen

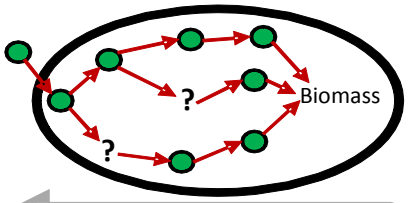
87%

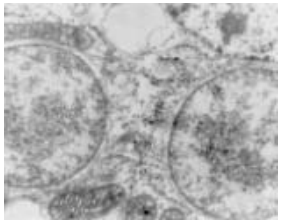
Optimized models

22 optimized models

Model accuracy 66%

Predicted 56 missing metabolic functions/ model

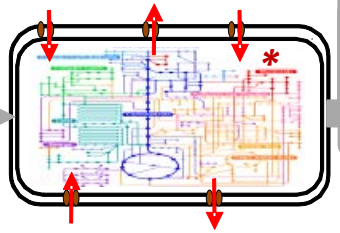




Predicted cell-host interactions

130 new metabolic models

- *965 reactions*
- *688 genes*
- *876 metabolites*

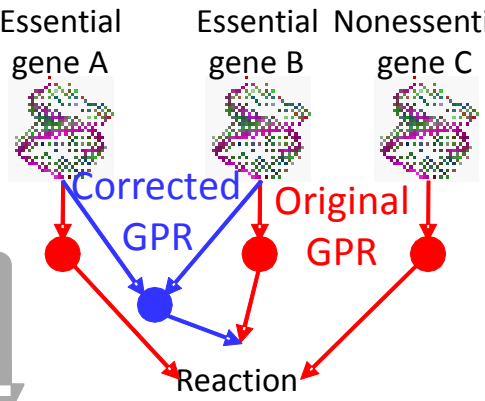


Predicted gene essentiality

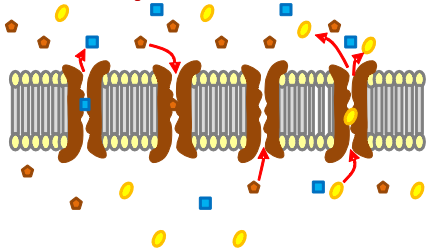
Predicted growth media

Predicted phenotypes

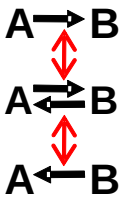
Correction for 202 annotations inconsistent with essentiality data



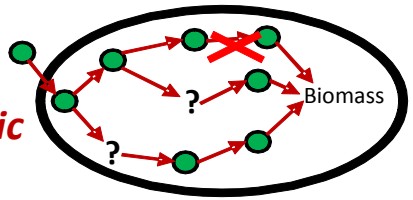
Predicting 69 missing transporters/model



Correcting reversibility constraints



Predicted missing and extra metabolic functions



Seed Models vs Published Models

•Single-genome Seed models compare favorably with published single genome models

Organism name	Published model	Published reactions	SEED Reactions	Published genes	SEED genes
Acinetobacter	iAbaylyiv4	868	1196	775	785
B. subtilis	iYO844	1020	1463	844	1041
C. acetobutylicum	iJL432	502	989	432	721
E. coli	iAF1260	2013	1529	1261	1083
G. sulfurreducens	iRM588	523	721	588	468
H. influenzae	iCS400	461	969	400	575
H. pylori	iIT341	476	731	341	421
L. plantarum	iBT721	643	908	721	699
L. lactis	iAO358	621	965	358	646
M. succiniciproducens	iTK425	686	1048	425	659
M. tuberculosis	iNJ661	939	1021	661	728
M. genitalium	iPS189	264	294	189	214
N. meningitidis	iGB555	496	903	555	560
P. gingivalis	iVM679	679	744	0*	399
P. aeruginosa	iMO1056	883	1386	1056	1094
P. putida	iNJ746	950	1261	746	1053
R. etli	iOR363	387	1264	363	1242
S. aureus	iSB619	641	1115	619	770
S. coelicolor	iIB700	700	1159	700	987

www.theseed.org/models/

Assessing Subsystem Annotations From Auto-completion

•We identify how *complete* the annotations are for each of the Seed subsystems by calculating the following ratio:

$$\frac{\text{auto-completion reactions in subsystem}}{\text{total reactions in subsystem}} = \text{Fraction of subsystem reactions with missing genes}$$

•Highest scoring subsystems:

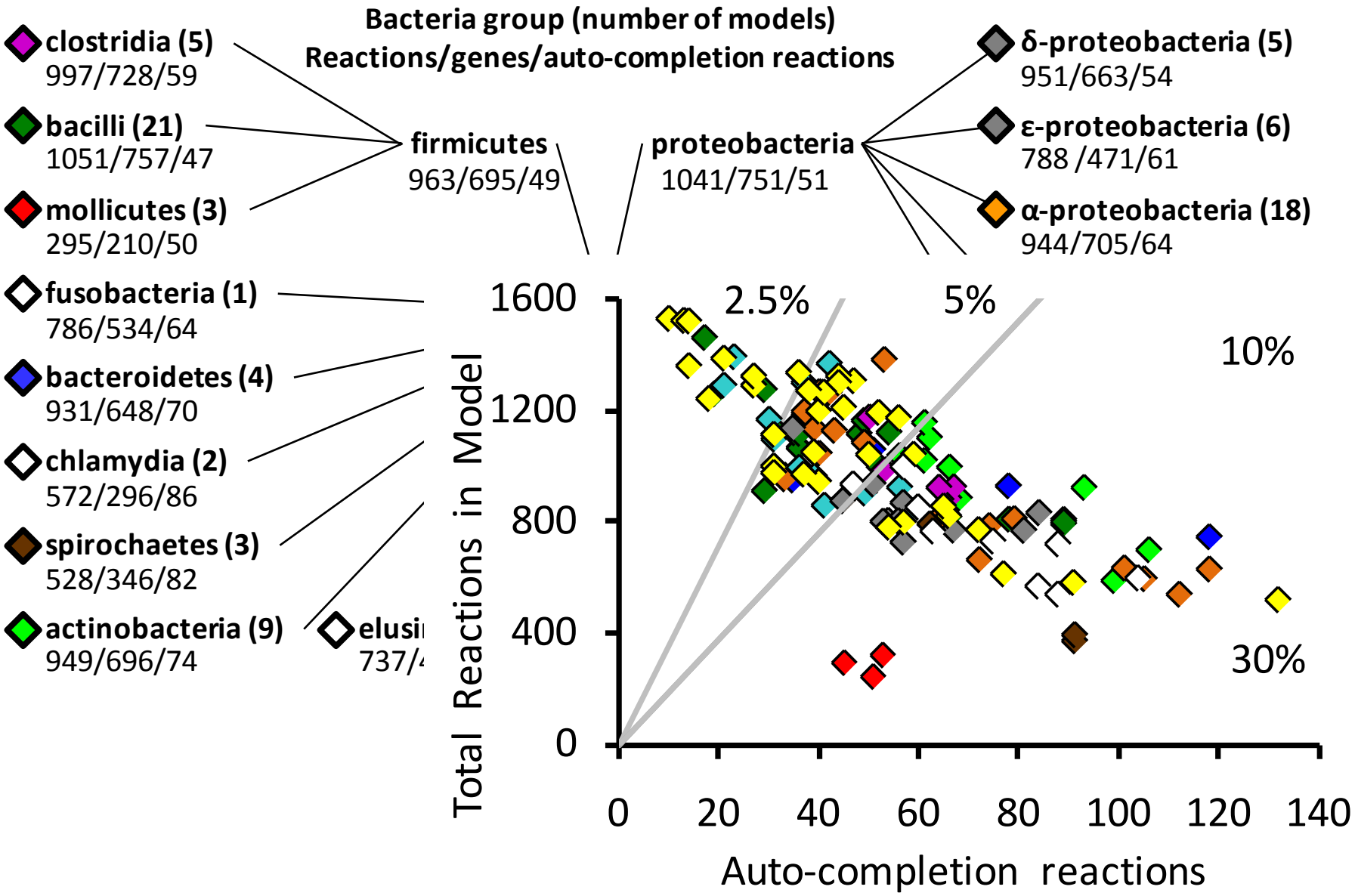
•*Cell Wall and Capsule Biosynthesis (15%)*
- •21 reactions per model added during auto-completion
- •*LOS Core Oligosaccharide Biosynthesis* (Gram negative)
- •*Teichoic and Lipoteichoic Acids Biosynthesis* (Gram positive)
- •*KDO2-Lipid A Biosynthesis*

•*Cofactors, Vitamins, and Prosthetic Group Biosynthesis (5%)*
- •10 reaction per model added during auto-completion
- •*Ubiquinone Biosynthesis*
- •*Menaquinone and Phylloquinone Biosynthesis*
- •*Thiamin Biosynthesis*

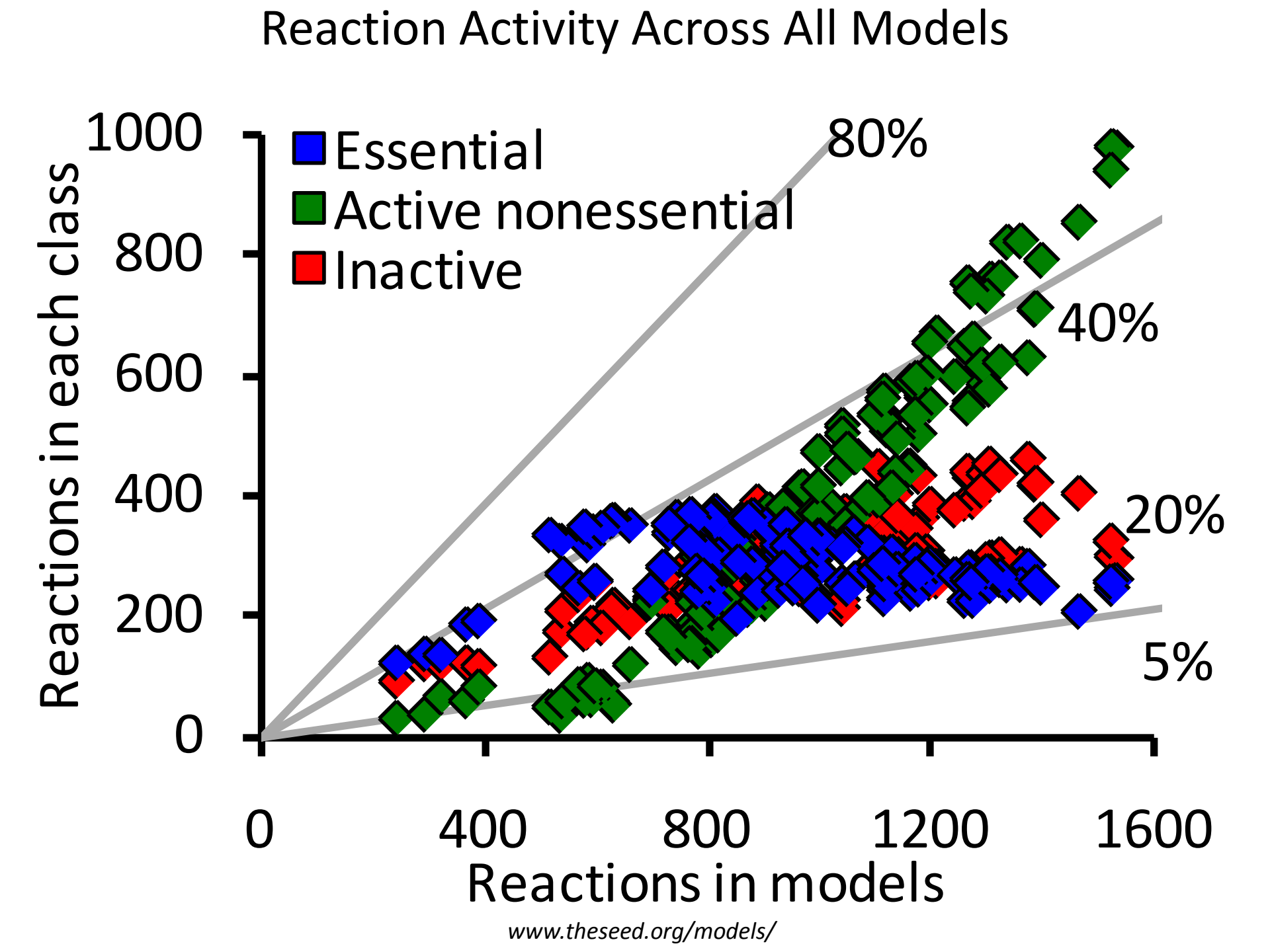
•Six subsystems account for 31/56 reactions added to each model during the auto-completion process

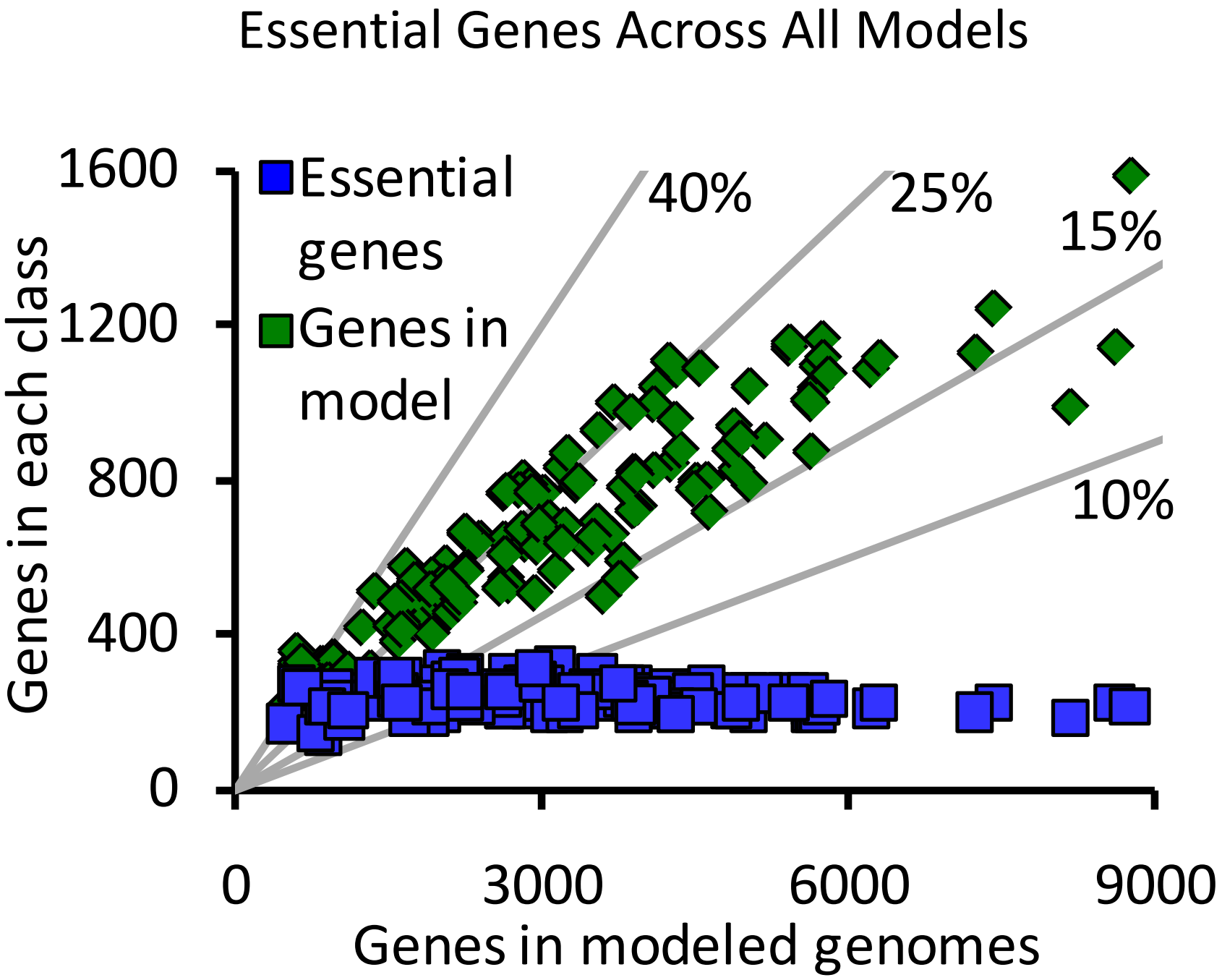
www.theseed.org/models/

Model statistics across the phylogenetic tree

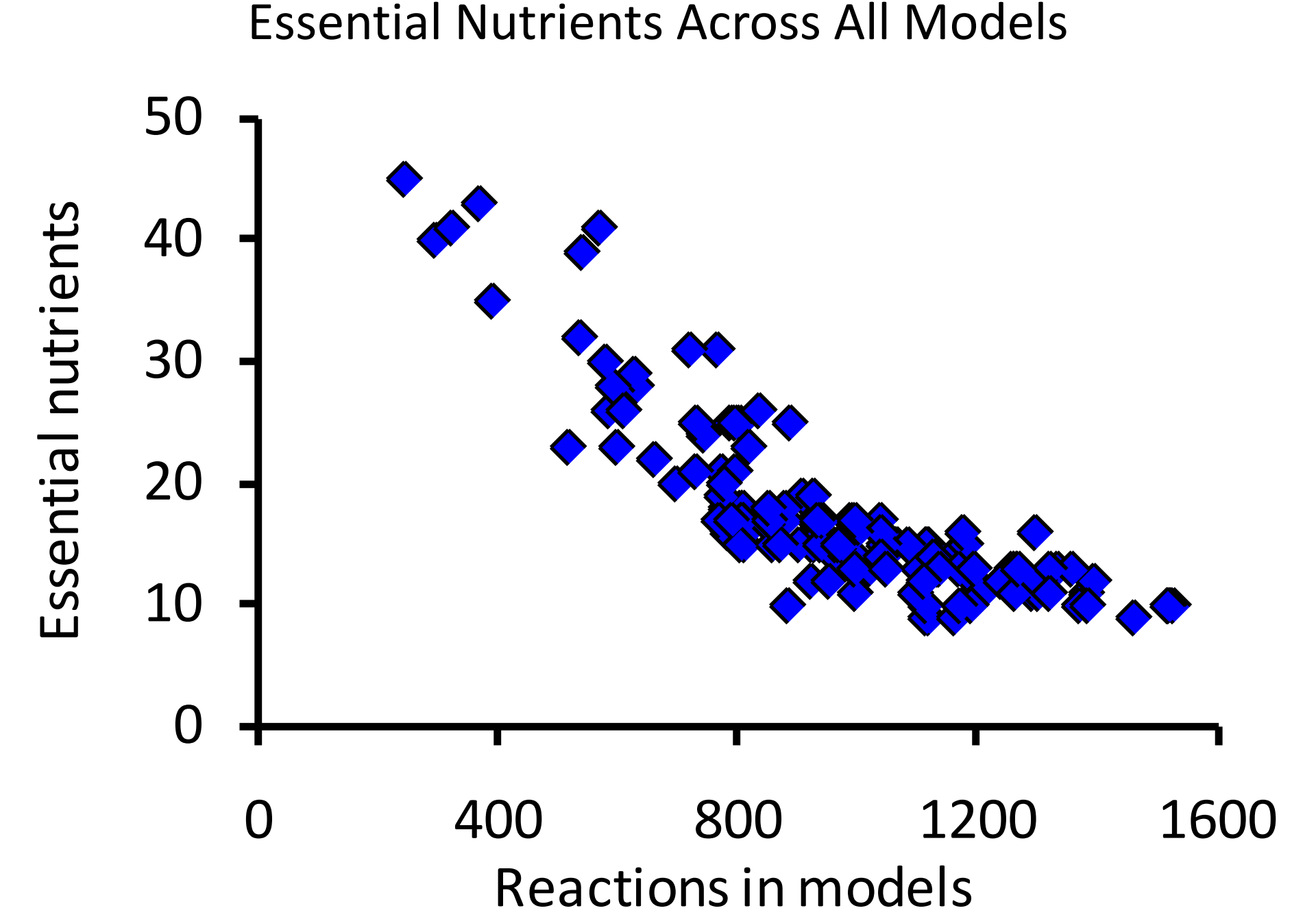


www.theseed.org/models/


Reaction Activity Across All Models
Essential
Active nonessential
Inactive
80%
40%
20%
5%
Reactions in each class
0
200
400
600
800
1000
0
400
800
1200
1600
Reactions in models
www.theseed.org/models/



www.theseed.org/models/



www.theseed.org/models/

Acknowledgements

ANL/U. Chicago Team

- Robert Olson
- Terry Disz
- Daniela Bartels
- Tobias Paczian
- Daniel Paarmann
- Scott Devoid
- Andreas Wilke
- Bill Mihalo
- Elizabeth Glass
- Folker Meyer
- Jared Wilkening
- Rick Stevens
- Alex Rodriguez
- Mark D'Souza
- Rob Edwards
- Christopher Henry

FIG Team

- Ross Overbeek
- Gordon Pusch
- Bruce Parello
- Veronika Vonstein
- Andrei Ostermann
- Olga Vassieva
- Olga Zagnitzko
- Svetlana Gerdes

Hope College Team

- Aaron Best
- Matt DeJongh
- Nathan Tintle
- Hope college students

Fellowship for Interpretation of Genomes

National Institute of Allergy and Infectious Diseases
National Institutes of Health