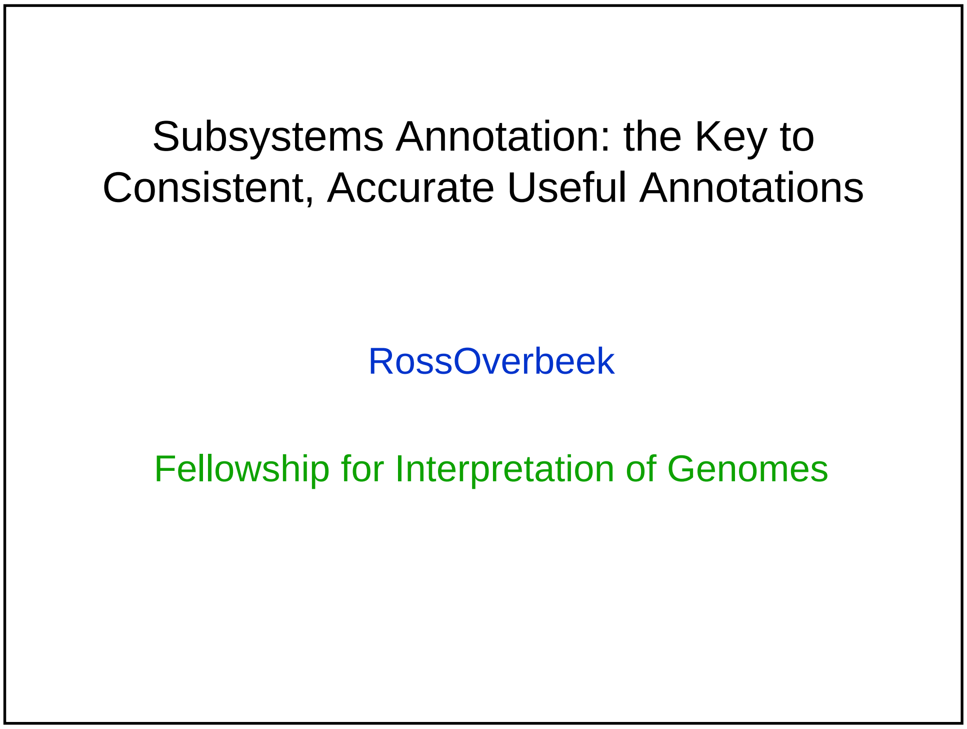
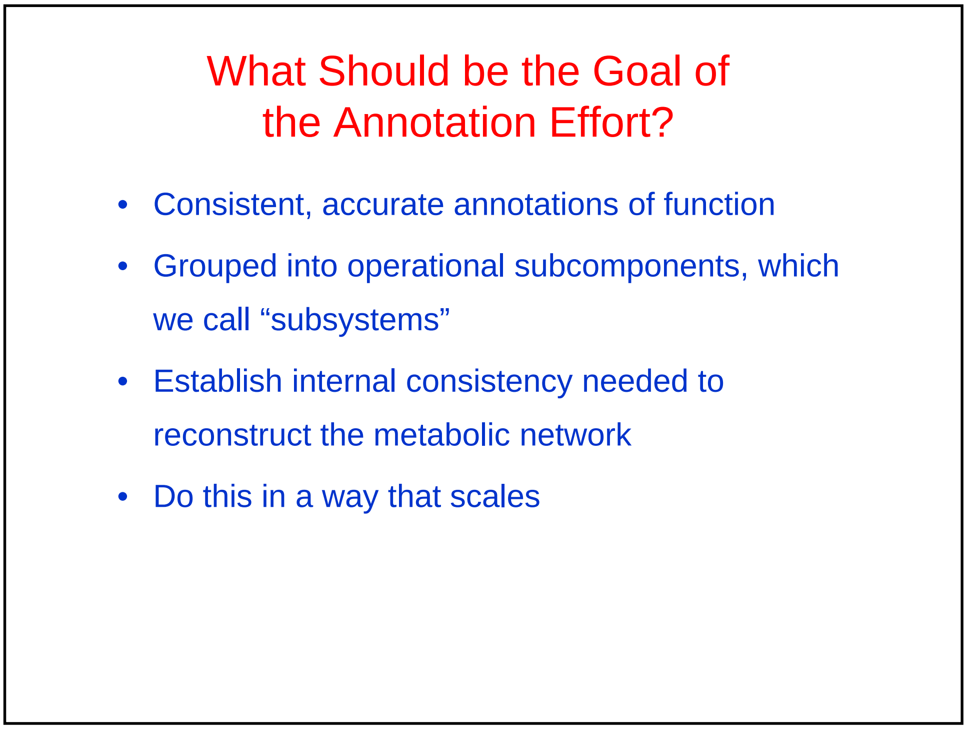


Subsystems Annotation: the Key to Consistent, Accurate Useful Annotations

RossOverbeek

Fellowship for Interpretation of Genomes

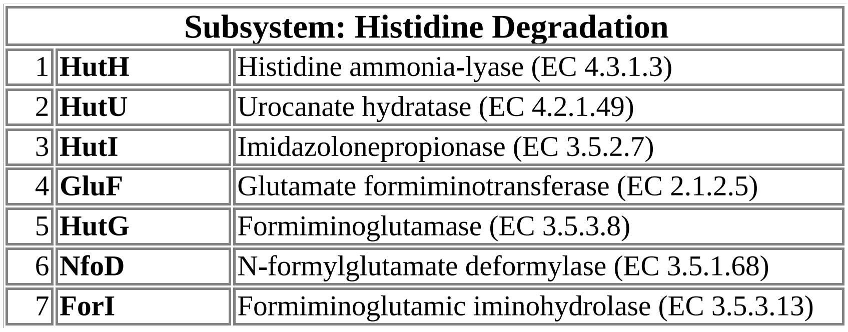


What Should be the Goal of the Annotation Effort?

- Consistent, accurate annotations of function
- Grouped into operational subcomponents, which we call "subsystems"
- Establish internal consistency needed to reconstruct the metabolic network
- Do this in a way that scales

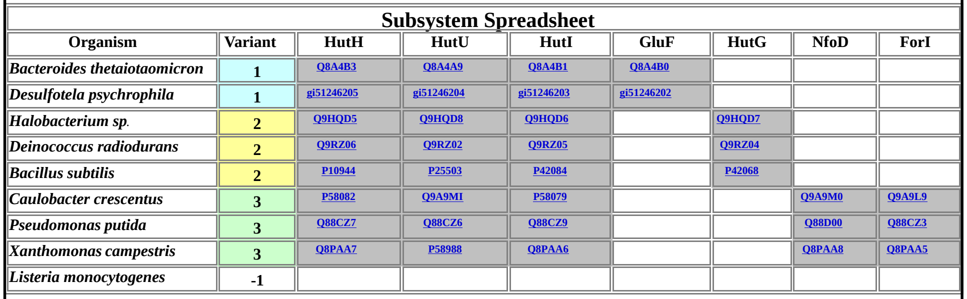
Example Subsystem: Histidine Degradation

- Conversion of histidine to glutamate
- Functional roles defined in table
- Inclusion in subsystem is *only* by functional role
- Controlled vocabulary …



Subsystem: Histidine Degradation		
1	**HutH**	Histidine ammonia-lyase (EC 4.3.1.3)
2	**HutU**	Urocanate hydratase (EC 4.2.1.49)
3	**HutI**	Imidazolonepropionase (EC 3.5.2.7)
4	**GluF**	Glutamate formiminotransferase (EC 2.1.2.5)
5	**HutG**	Formiminoglutamase (EC 3.5.3.8)
6	**NfoD**	N-formylglutamate deformylase (EC 3.5.1.68)
7	**ForI**	Formiminoglutamic iminohydrolase (EC 3.5.3.13)

Subsystem Spreadsheet

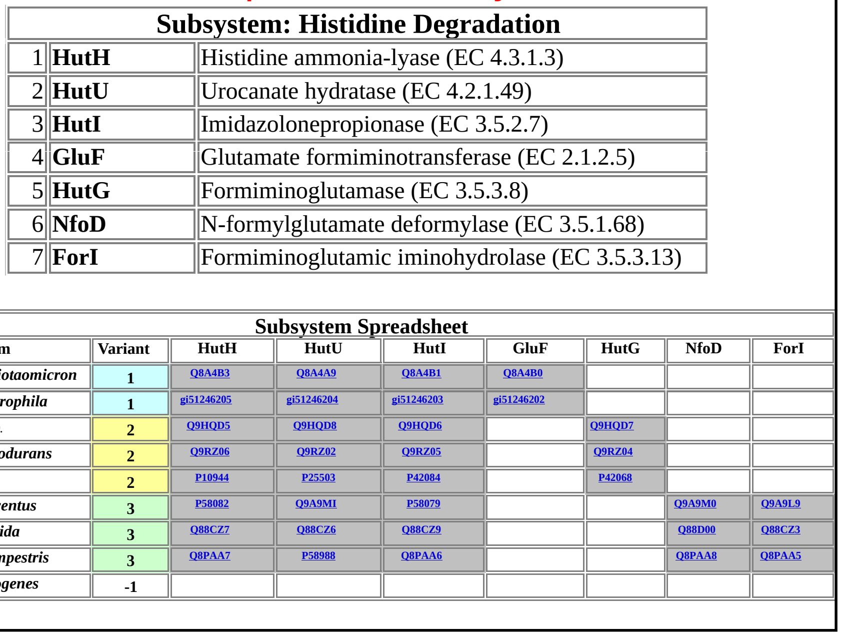


Subsystem Spreadsheet

Organism	Variant	HutH	HutU	HutI	GluF	HutG	NfoD	ForI
Bacteroides thetaiotaomicron	1	Q8A4B3	Q8A4A9	Q8A4B1	Q8A4B0			
Desulfotela psychrophila	1	gi51246205	gi51246204	gi51246203	gi51246202			
Halobacterium sp.	2	Q9HQD5	Q9HQD8	Q9HQD6		Q9HQD7		
Deinococcus radiodurans	2	Q9RZ06	Q9RZ02	Q9RZ05		Q9RZ04		
Bacillus subtilis	2	P10944	P25503	P42084		P42068		
Caulobacter crescentus	3	P58082	Q9A9MI	P58079			Q9A9M0	Q9A9L9
Pseudomonas putida	3	Q88CZ7	Q88CZ6	Q88CZ9			Q88D00	Q88CZ3
Xanthomonas campestris	3	Q8PAA7	P58988	Q8PAA6			Q8PAA8	Q8PAA5
Listeria monocytogenes	-1							

- Column headers taken from table of functional roles
- Rows are selected genomes or organisms
- Cells are populated with specific, annotated genes
- Functional variants defined by the annotated roles
- Variant code -1 indicates subsystem is not functional
- Clustering shown by color

“The Populated Subsystem”



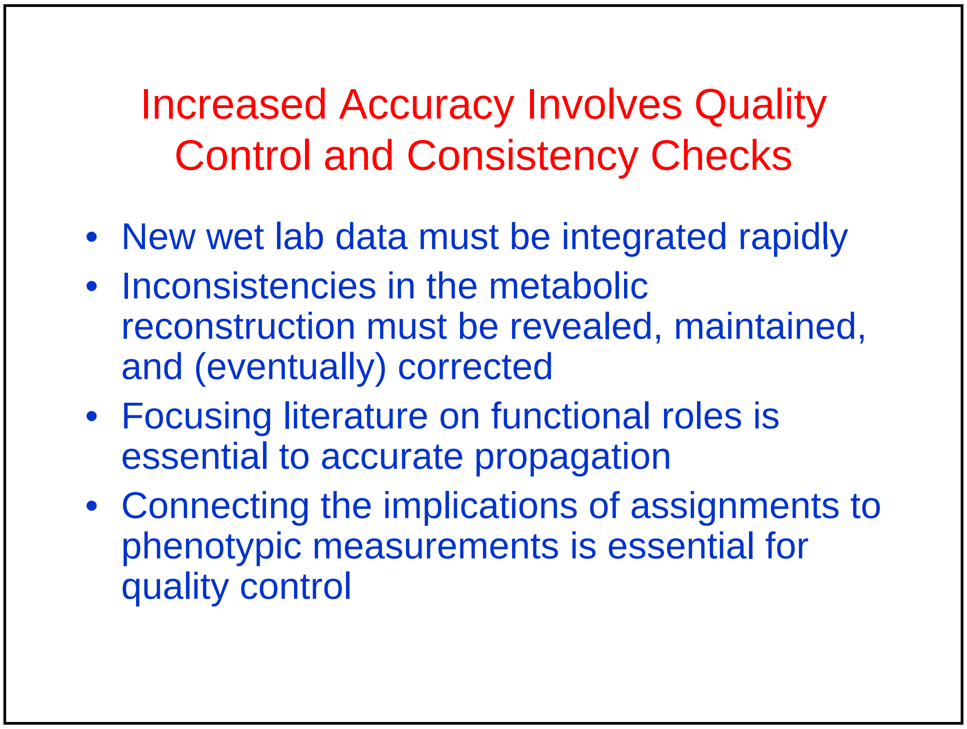
Subsystem: Histidine Degradation		
1	**HutH**	Histidine ammonia-lyase (EC 4.3.1.3)
2	**HutU**	Urocanate hydratase (EC 4.2.1.49)
3	**HutI**	Imidazolonepropionase (EC 3.5.2.7)
4	**GluF**	Glutamate formiminotransferase (EC 2.1.2.5)
5	**HutG**	Formiminoglutamase (EC 3.5.3.8)
6	**NfoD**	N-formylglutamate deformylase (EC 3.5.1.68)
7	**ForI**	Formiminoglutamic iminohydrolase (EC 3.5.3.13)

Subsystem Spreadsheet

Organism	Variant	HutH	HutU	HutI	GluF	HutG	NfoD	ForI
Bacteroides thetaiotaomicron	1	Q8A4B3	Q8A4A9	Q8A4B1	Q8A4B0			
Desulfotela psychrophila	1	gi51246205	gi51246204	gi51246203	gi51246202			
Halobacterium sp.	2	Q9HQD5	Q9HQD8	Q9HQD6		Q9HQD7		
Deinococcus radiodurans	2	Q9RZ06	Q9RZ02	Q9RZ05		Q9RZ04		
Bacillus subtilis	2	P10944	P25503	P42084		P42068		
Caulobacter crescentus	3	P58082	Q9A9MI	P58079			Q9A9M0	Q9A9L9
Pseudomonas putida	3	Q88CZ7	Q88CZ6	Q88CZ9			Q88D00	Q88CZ3
Xanthomonas campestris	3	Q8PAA7	P58988	Q8PAA6			Q8PAA8	Q8PAA5
Listeria monocytogenes	-1							

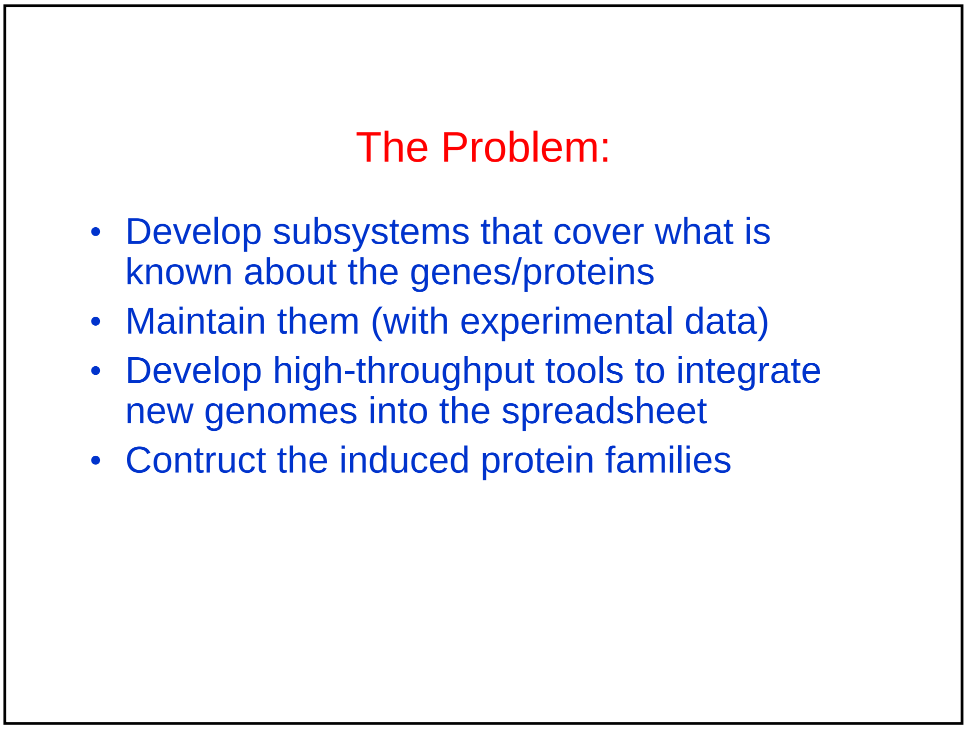
Subsystems Can Be Made as “Stand-Alone” Data Objects

- This only requires conversion of gene/protein IDs
- It establishes a consistent, controlled vocabulary
- You can annotate the functional roles (with reactions, GO terms, literature) and achieve propagation to genes



Increased Accuracy Involves Quality Control and Consistency Checks

- New wet lab data must be integrated rapidly
- Inconsistencies in the metabolic reconstruction must be revealed, maintained, and (eventually) corrected
- Focusing literature on functional roles is essential to accurate propagation
- Connecting the implications of assignments to phenotypic measurements is essential for quality control



The Problem:

- Develop subsystems that cover what is known about the genes/proteins
- Maintain them (with experimental data)
- Develop high-throughput tools to integrate new genomes into the spreadsheet
- Contruct the induced protein families