QIIMP:
Microbiome Metadata Made Easy

AMANDA BIRMINGHAM

SENIOR BIOINFORMATICS ENGINEER

CENTER FOR COMPUTATIONAL BIOLOGY & BIOINFORMATICS, UCSD

Metadata Are Critical

- Metadata: Information about each sample that was sequenced
  - *e.g.* host organism, treatment, date of sampling
- Hypotheses to test usually depend on metadata
  - "Are the gut microbial communities more diverse in treated mice than untreated ones?"

Metadata Are Critical–And Often Poor

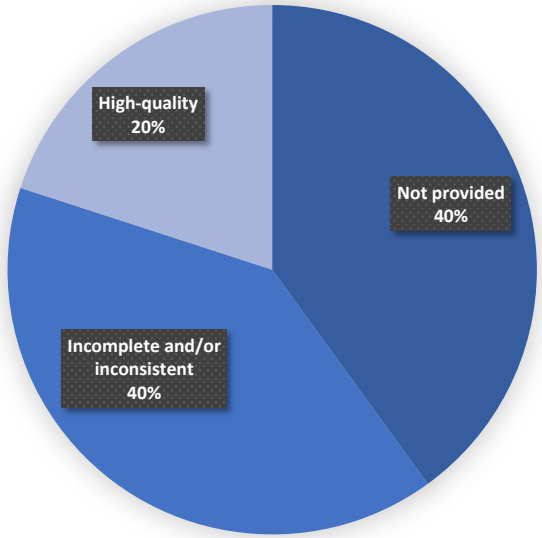
- Metadata: Information about each sample that was sequenced
  - *e.g.* host organism, treatment, date of sampling
- Hypotheses to test usually depend on metadata
  - "Are the gut microbial communities more diverse in treated mice than untreated ones?"
- "Here is my mapping file. Please analyze my samples by treatment."

```
#SampleID BarcodeSequence LinkerPrimerSequence Description
PC.354 AGCACGAGCCTA YATGCTGCCTCCCGTAGGAGT mouse__I.D._354
PC.355 AACTCGTCGATG YATGCTGCCTCCCGTAGGAGT mouse__I.D._355
PC.356 ACAGACCACTCA YATGCTGCCTCCCGTAGGAGT mouse__I.D._356
PC.481 ACCAGCGACTAG YATGCTGCCTCCCGTAGGAGT mouse__I.D._481
PC.593 AGCAGCACTTGT YATGCTGCCTCCCGTAGGAGT mouse__I.D._593
PC.607 AACTGTGCGTAC YATGCTGCCTCCCGTAGGAGT mouse__I.D._607
PC.634 ACAGAGTCGGCT YATGCTGCCTCCCGTAGGAGT mouse__I.D._634
PC.635 ACCGCAGAGTCA YATGCTGCCTCCCGTAGGAGT mouse__I.D._635
PC.636 ACGGTGAGTGTC YATGCTGCCTCCCGTAGGAGT mouse__I.D._636
```

Metadata Are Critical–And Often Poor

- Metadata: Information about each sample that was sequenced
  - *e.g.* host organism, treatment, date of sampling
- Hypotheses to test usually depend on metadata
  - "Are the gut microbial communities more diverse in treated mice than untreated ones?"
- "Here is my mapping file. Please analyze my samples by treatment."

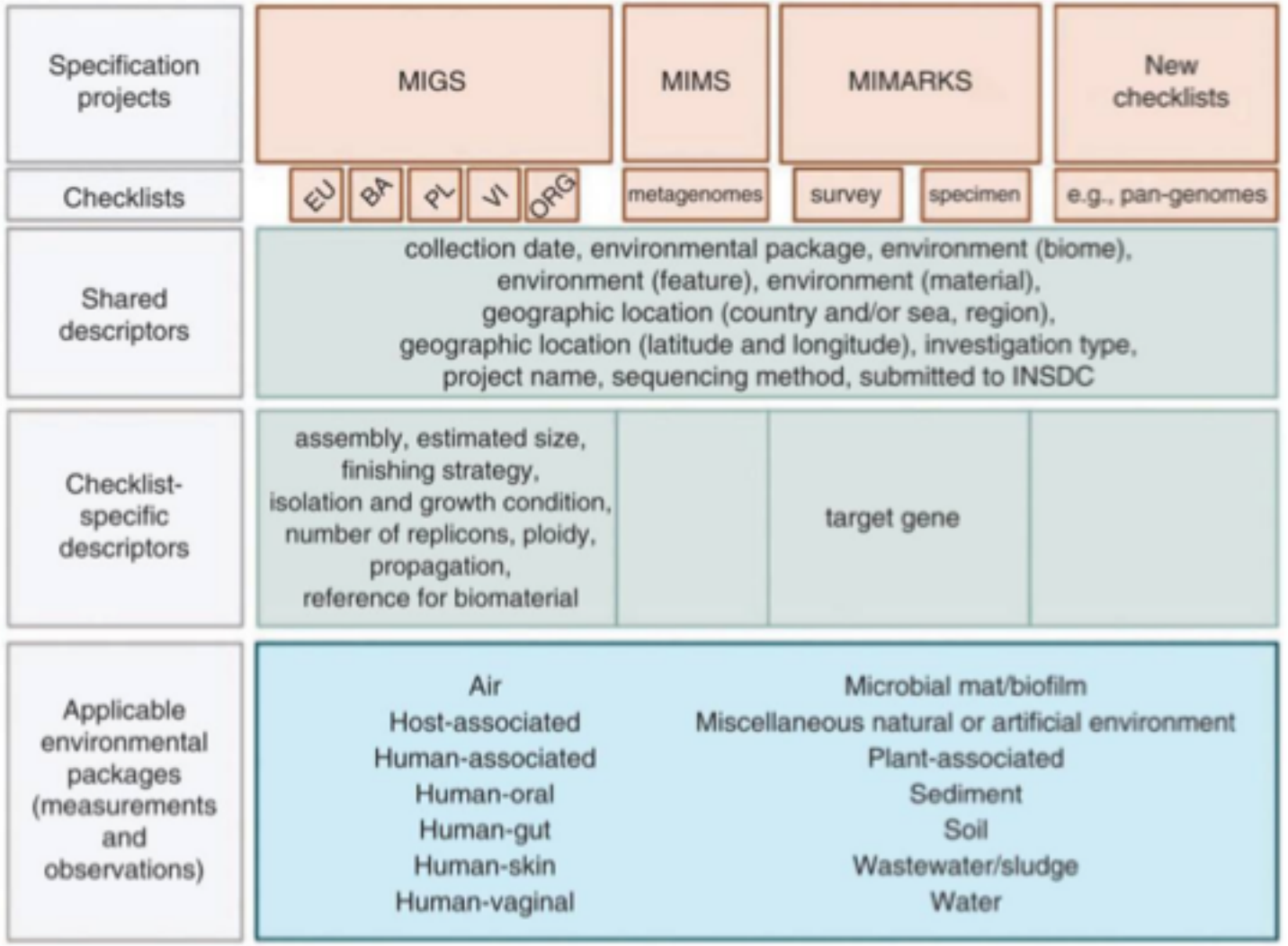
```
#SampleID BarcodeSequence LinkerPrimerSequence Description
PC.354 AGCACGAGCCTA YATGCTGCCTCCCGTAGGAGT mouse__I.D._354
PC.355 AACTCGTCGATG YATGCTGCCTCCCGTAGGAGT mouse__I.D._355
PC.356 ACAGACCACTCA YATGCTGCCTCCCGTAGGAGT mouse__I.D._356
PC.481 ACCAGCGACTAG YATGCTGCCTCCCGTAGGAGT mouse__I.D._481
PC.593 AGCAGCACTTGT YATGCTGCCTCCCGTAGGAGT mouse__I.D._593
PC.607 AACTGTGCGTAC YATGCTGCCTCCCGTAGGAGT mouse__I.D._607
PC.634 ACAGAGTCGGCT YATGCTGCCTCCCGTAGGAGT mouse__I.D._634
PC.635 ACCGCAGAGTCA YATGCTGCCTCCCGTAGGAGT mouse__I.D._635
PC.636 ACGGTGAGTGTC YATGCTGCCTCCCGTAGGAGT mouse__I.D._636
```



Researchers' Metadata

Standardized Metadata Are Complicated

- To support cross-experiment comparisons, metadata must be *standardized*
- From Genomic Standards Consortium (GSC) :
  - Minimum Information about any (x) Sequence (MIxS)
    - MIGS for genomes
    - MIMS for metagenomes
    - MIMARKS for marker genes



Yilmaz et al, Nat Biotechnol. 2011

Standardized Metadata Are Complicated

- Environmental packages add more detail—and yet more complexity

	A	B	C	E	F	O	P	R
1	Structured comment name	Item	Definition	Expected value	Section	Value syntax	Occurence	Preferred unit
57	alt	altitude	the altitude of the sample is the vertical distance between Earth's surface above sea level and the sampled position in the air	measurement value	host-associated	{float} m	1	meter
58	depth	depth	depth is defined as the vertical distance below local surface, e.g. for sediment or soil samples depth is measured from sediment or soil surface, respectivly. Depth can be reported as an interval for subsurface samples	measurement value	host-associated	{float} m	1	
59	elev	elevation	the elevation of the sampling site as measured by the vertical distance from mean sea level	measurement value	host-associated	{float} {unit}	1	meter
60	host_common_name	host common name	common name of the host, e.g. human	common name	host-associated	{text}	1	
61	host_taxid	host taxid	NCBI taxon id of the host, e.g. 9606	NCBI taxon identifier	host-associated	{integer}	1	
62	host_subject_id	host subject id	a unique identifier by which each subject can be refered to, de-identified, e.g. #131	unique identifier	host-associated	{text}	1	
63	host_age	host age	age of host at the time of sampling; relevant scale depends on species and study, e.g. could be seconds for amoebae or centuries for trees	value	host-associated	{float} {unit}	1	year, day, hour
64	host_life_stage	host life stage	description of life stage of host	stage	host-associated	{text}	1	
65	host_sex	host sex	physical sex of the host	enumeration	host-associated	{male|female|neuter|hermaphrodite|not determined|not applicable|not collected|not provided|restricted access}	1	
66	host_disease_stat	host disease status	list of diseases with which the host has been diagnosed; can include multiple diagnoses. the value of the field depends on host; for humans the terms should be chosen from DO (Disease Ontology) at http://www.disease-ontology.org, other hosts are free text	disease name or DO	host-associated	{term}	m	
67	chem_administration	chemical administration	list of chemical compounds administered to the host or site where sampling occurred, and when (e.g. antibiotics, N fertilizer, air filter); can include multiple compounds. For Chemical Entities of Biological Interest ontology (CHEBI) (v 111), http://purl.bioontology.org/ontology/CHEBI	CHEBI;timestamp	host-associated	{term}; {timestamp}	m	
68	host_body_habitat	host body habitat	original body habitat where the sample was obtained from	FMA	host-associated	{term}	1	
69	host_body_site	host body site	name of body site where the sample was obtained from	FMA	host-associated	{term}	1	
70	host_body_product	host body product	substance produced by the body, e.g. stool, mucus, where the sample was obtained from	FMA	host-associated	{term}	1	
71	host_tot_mass	host total mass	total mass of the host at collection, the unit depends on host	measurement value	host-associated	{float} {unit}	1	kilogram, gram
72	host_height	host height	the height of subject	measurement value	host-associated	{float} {unit}	1	centimeter, millimeter, meter
73	host_length	host length	the length of subject	measurement value	host-associated	{float} {unit}	1	centimeter, millimeter, meter

README | MIxS-host associated | +

MIxShostassoc_210514.xls by core MIxS team

Excel Wins (Nearly) Every Time

- Metadata are perfect for a database
  - Enforce relational integrity, minimize redundancy through normalization, track modifications, etc
  - But most biologists won't set up and maintain one

Excel Wins (Nearly) Every Time

- Metadata are perfect for a database
  - Enforce relational integrity, minimize redundancy through normalization, track modifications, etc
  - But most biologists won't set up and maintain one
- (At least one) lovely metadata software suite exists
  - Based on Investigation/Study/Assay model
  - Offers GUI for custom field configuration, data entry
  - Outputs ISA-TAB, submits to European Nucleotide Archive
  - But most biologists won't install and register it, let alone use it

isatools

http://isa-tools.org/

Excel Wins (Nearly) Every Time

- Metadata are perfect for a database
  - Enforce relational integrity, minimize redundancy through normalization, track modifications, etc
  - But most biologists won't set up and maintain one
- (At least one) lovely metadata software suite exists
  - Based on Investigation/Study/Assay model
  - Offers GUI for custom field configuration, data entry
  - Outputs ISA-TAB, submits to European Nucleotide Archive
  - But most biologists won't install and register it, let alone use it

http://isa-tools.org/

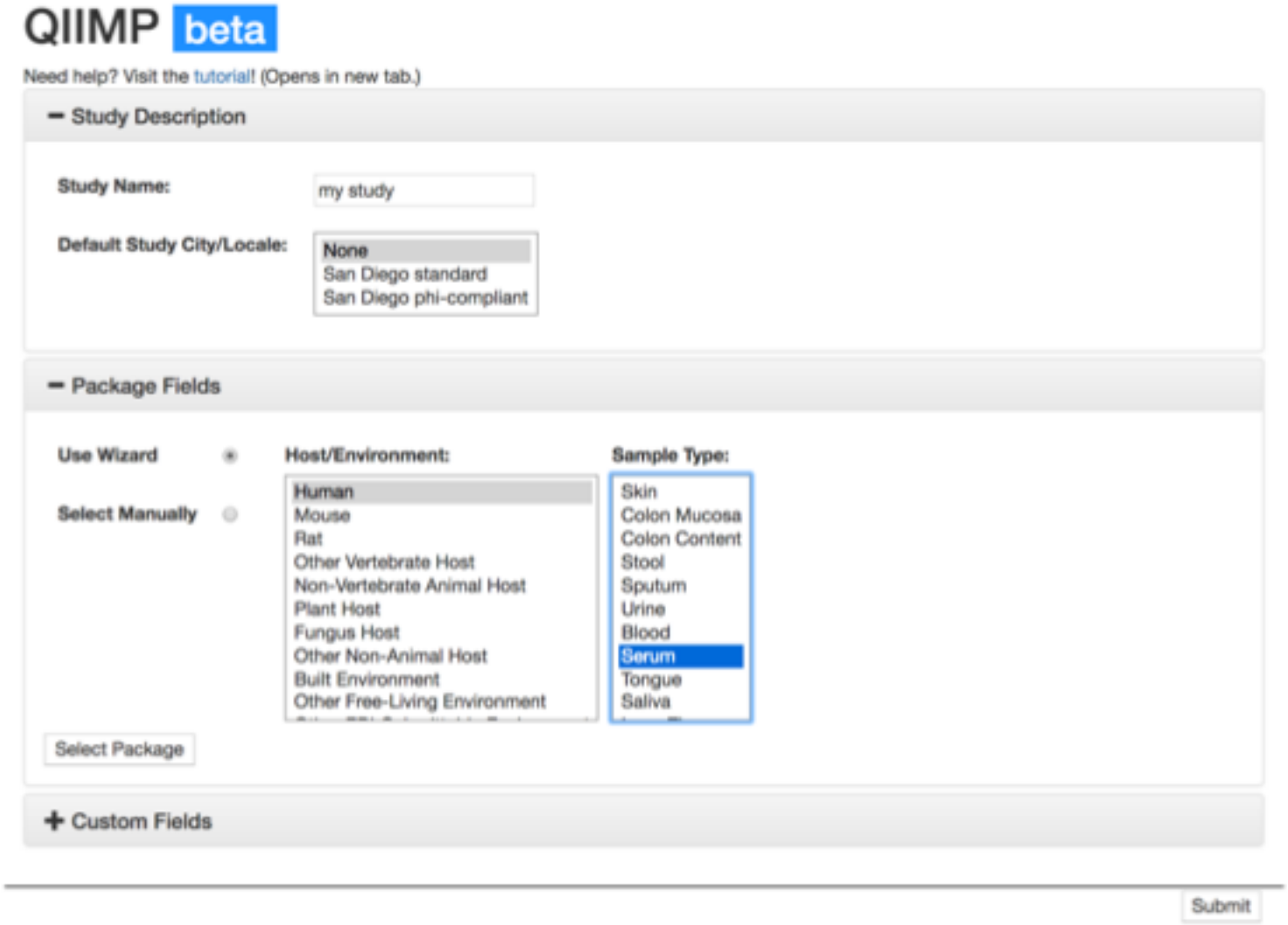
- What most biologists WILL and DO use is Excel
- For them, data analysts and computationalists have to decide:

Do we want exquisite tools or do we want quality metadata?

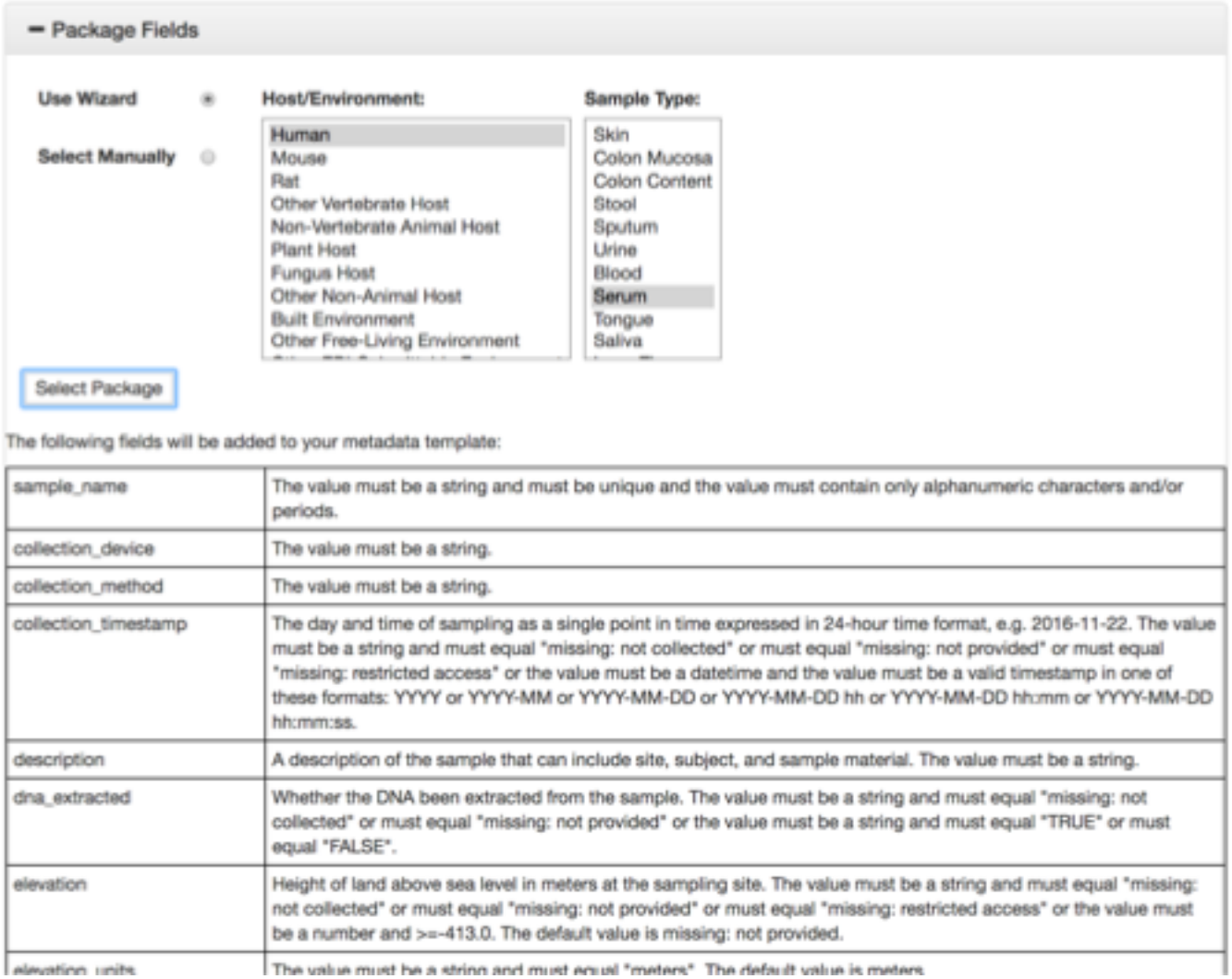
QIIMP Meets Users Where They Are

- The Quick and Intuitive Interactive Metadata Portal (QIIMP)
  - Pronounced 'chimp'—part of the QIIME & Qiita family
  - Asks minimal questions to determine appropriate standards package
  - Invites easy but structured custom field definition
  - Creates a *macro-free*, *cross-platform* Excel file that validates entered metadata

One-Page Web Interface (QIIMP)

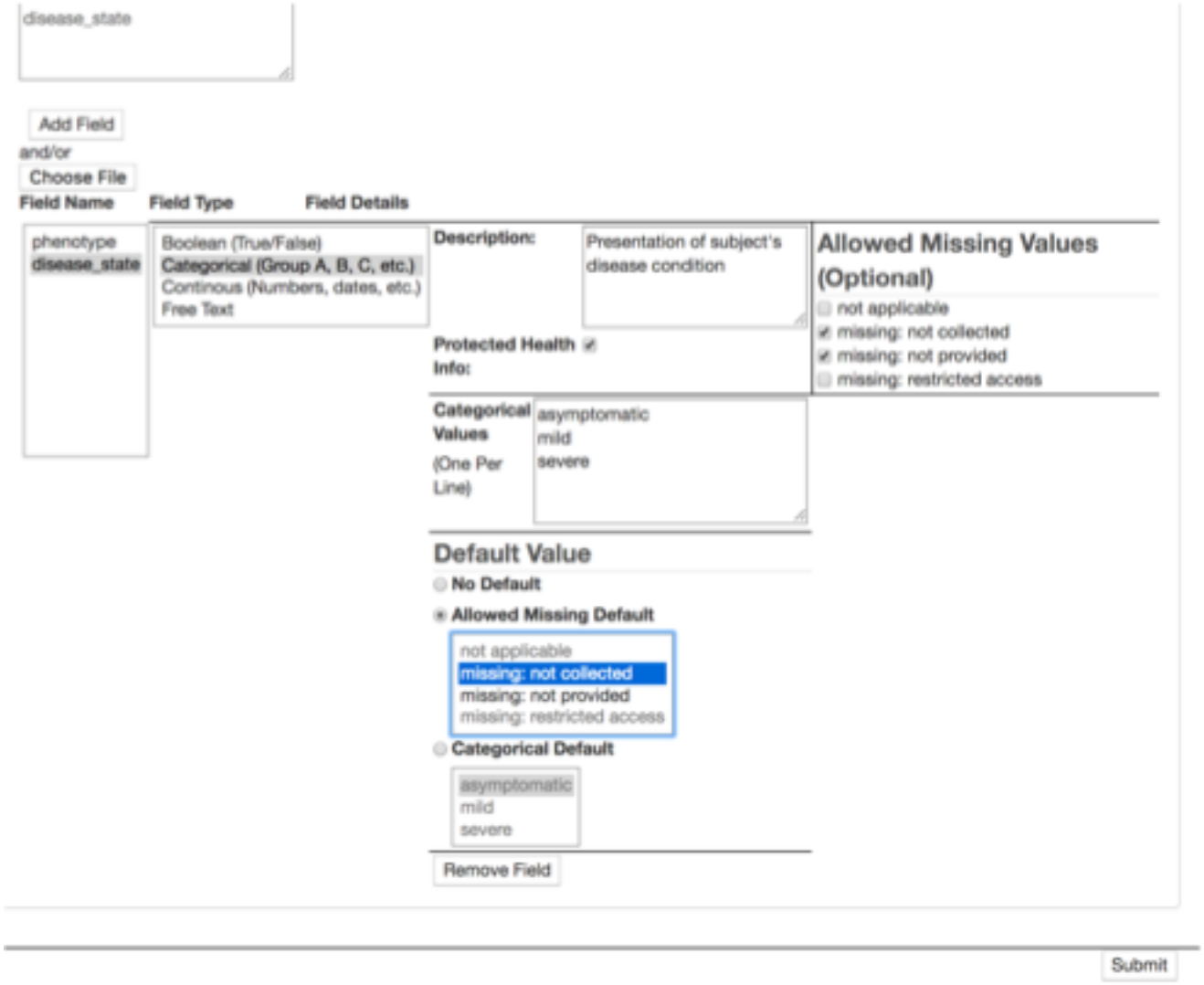


One-Page Web Interface (QIIMP)

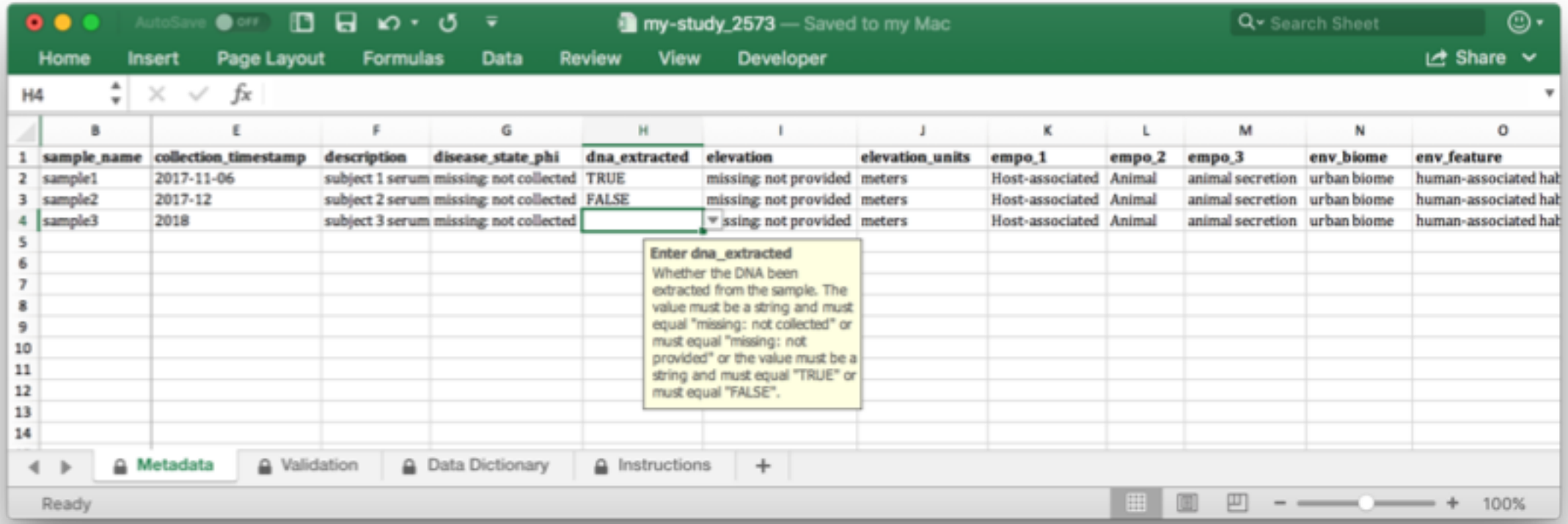


sample_name	The value must be a string and must be unique and the value must contain only alphanumeric characters and/or periods.
collection_device	The value must be a string.
collection_method	The value must be a string.
collection_timestamp	The day and time of sampling as a single point in time expressed in 24-hour time format, e.g. 2016-11-22. The value must be a string and must equal "missing: not collected" or must equal "missing: not provided" or must equal "missing: restricted access" or the value must be a datetime and the value must be a valid timestamp in one of these formats: YYYY or YYYY-MM or YYYY-MM-DD or YYYY-MM-DD hh or YYYY-MM-DD hh:mm or YYYY-MM-DD hh:mm:ss.
description	A description of the sample that can include site, subject, and sample material. The value must be a string.
dna_extracted	Whether the DNA been extracted from the sample. The value must be a string and must equal "missing: not collected" or must equal "missing: not provided" or the value must be a string and must equal "TRUE" or must equal "FALSE".
elevation	Height of land above sea level in meters at the sampling site. The value must be a string and must equal "missing: not collected" or must equal "missing: not provided" or must equal "missing: restricted access" or the value must be a number and >=-413.0. The default value is missing: not provided.
elevation_units	The value must be a string and must equal "meters". The default value is meters.

One-Page Web Interface (QIIMP)

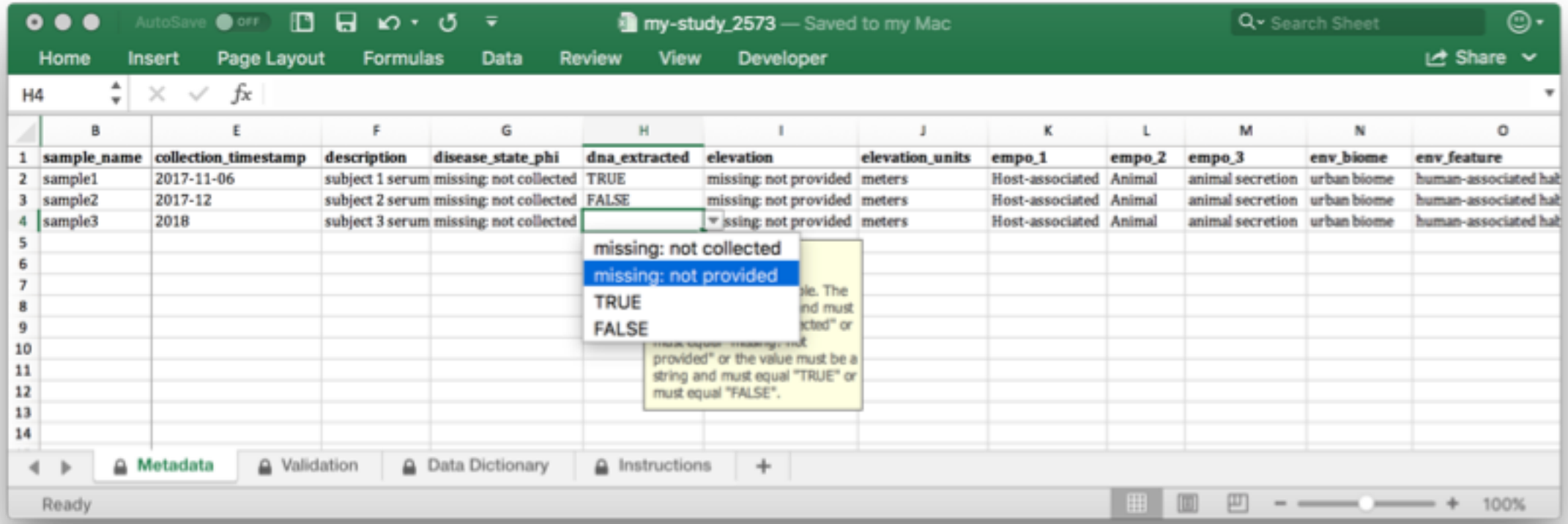


Customized Excel Template (QIIMP)



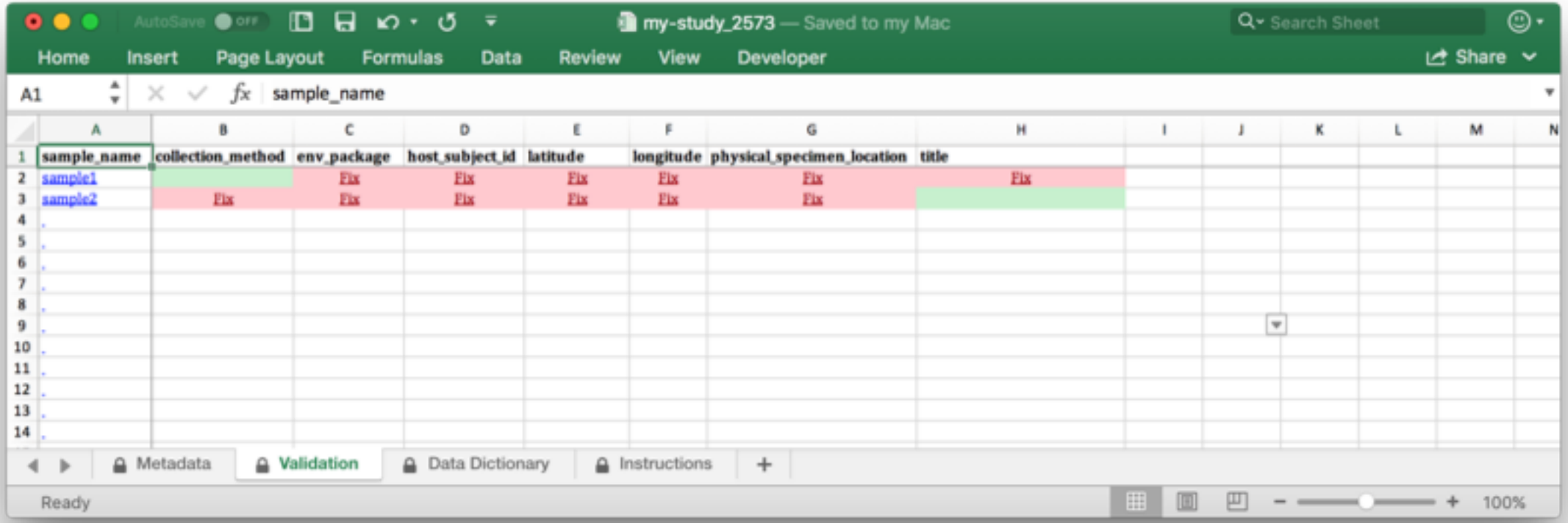
	B	E	F	G	H	I	J	K	L	M	N	O
1	sample_name	collection_timestamp	description	disease_state_phi	dna_extracted	elevation	elevation_units	empo_1	empo_2	empo_3	env_biome	env_feature
2	sample1	2017-11-06	subject 1 serum	missing: not collected	TRUE	missing: not provided	meters	Host-associated	Animal	animal secretion	urban biome	human-associated hab
3	sample2	2017-12	subject 2 serum	missing: not collected	FALSE	missing: not provided	meters	Host-associated	Animal	animal secretion	urban biome	human-associated hab
4	sample3	2018	subject 3 serum	missing: not collected		ssing: not provided	meters	Host-associated	Animal	animal secretion	urban biome	human-associated hab
5												
6												
7												
8												
9												
10												
11												
12												
13												
14												

Customized Excel Template (QIIMP)



	B	E	F	G	H	I	J	K	L	M	N	O
1	sample_name	collection_timestamp	description	disease_state_phi	dna_extracted	elevation	elevation_units	empo_1	empo_2	empo_3	env_biome	env_feature
2	sample1	2017-11-06	subject 1 serum	missing: not collected	TRUE	missing: not provided	meters	Host-associated	Animal	animal secretion	urban biome	human-associated hab
3	sample2	2017-12	subject 2 serum	missing: not collected	FALSE	missing: not provided	meters	Host-associated	Animal	animal secretion	urban biome	human-associated hab
4	sample3	2018	subject 3 serum	missing: not collected		ssing: not provided	meters	Host-associated	Animal	animal secretion	urban biome	human-associated hab
5												
6												
7												
8												
9												
10												
11												
12												
13												
14												

Customized Excel Template (QIIMP)



Conclusions & Acknowledgments

- Improving metadata management is as important as improving laboratory protocols
- Several excellent tools exist for capturing standardized, high-quality metadata
  - But a significant segment of researchers simply won't USE them
- *"The Customer Is Always Right!"*

Conclusions & Acknowledgments

- Improving metadata management is as important as improving laboratory protocols
- Several excellent tools exist for capturing standardized, high-quality metadata
  - But a significant segment of researchers simply won't USE them
- *"The Customer Is Always Right!"*
  - Well, no—the customer is often wrong, but it is no good telling them so
  - Just have to find a way to make them happy!
- QIIMP provides simple but rigorous Excel-based metadata collection
- QIIMP launches **July 23**!
  - July 24 St. Petersburg time :)
  - Part of the Qiita open-source microbial study management platform
  - https://qiita.ucsd.edu/qiimp
- Acknowledgements
  - Gail Ackermann, Jeff Dereus, Antonio Gonzalez-Pena
  - Austin Swafford, Katherine Fisch