Charting a path forward: Data and Metadata

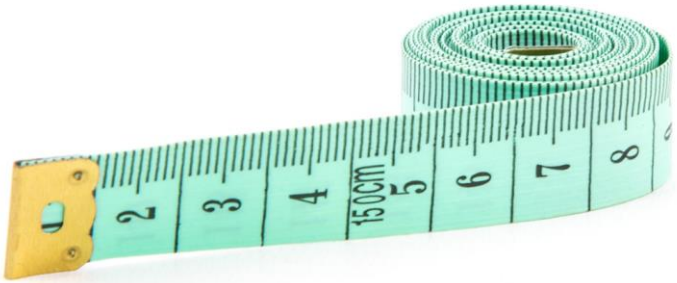
Kendell Clement
Luca Pinello Lab - http://pinellolab.org/
Massachusetts General Hospital
Harvard Medical School
kclement@mgh.harvard.edu

A need for standards

Targeted genome editing is a rapidly-developing technology

Individual entities are developing a variety of methodologies for recording and reporting genome editing results

Consistent and clear reporting of genome editing results will encourage reproducibility and ensure scientific integrity within the field and maintain positive public perception



Metadata and Data

Data: what you obtain from the measurement
Metadata: information necessary to reproduce your measurement conditions

- Design procedure
- Experimental components used
- Measurement assays performed
- Data processing steps

Data and metadata subgroup

Recording and sharing experimental data and metadata is important.

How can we make it painless?

Why metadata?

Efficiency: When performing a genome editing experiment, what elements of the experimental process would you want to record if you had to repeat the experiment?

Comprehension: In reading a published study, what information would you like to know about the steps authors took to arrive at the results?

Reproducibility: In trying to reproduce reported results, what details would be necessary?

Supervision: What biochemical or computational tools were used in a genome editing experiment? (Gov. agents, licensed products, etc.)

Cooperation: What information could be shared across and through procedures and pipelines for greater efficiency?

Genome editing workflow

Determine genome target

Design editing molecule

Assess editing performance *in silico*

Order oligos

Perform editing

Sequencing

Analysis

Storage or sharing of results

Challenges

1. **What** metadata should be stored?
2. **When** in the experimentation process will metadata be added, who adds the metadata?
3. **How** will the metadata be stored?
4. **Where** will metadata be collected and accessed?

Actions

1. Assembled a team of experts from many backgrounds
2. Held bimonthly meetings to discuss needs and metadata requirements
3. Compiled a list of metadata entries
4. Implementing a metadata storage format

Data and metadata working group planning participants

NIST

CARIBOU BIOSCIENCES

IDT INTEGRATED DNA TECHNOLOGIES

NEW YORK GENOME CENTER

UNIVERSITY OF CALIFORNIA SANTA CRUZ

CRISPR THERAPEUTICS

KBIOBOX

horizon precision genome editing

BCM Baylor College of Medicine

BROAD INSTITUTE

PACIFIC BIOSCIENCES

DESKTOPGENETICS

Intellia THERAPEUTICS

SickKids THE HOSPITAL FOR SICK CHILDREN

EMBL

illumina

HARVARD MEDICAL SCHOOL

dkfz. GERMAN CANCER RESEARCH CENTER IN THE HELMHOLTZ ASSOCIATION

MASSACHUSETTS GENERAL HOSPITAL

NYU

Stanford

Bar-Ilan University

Metadata categories

1. Editing molecule design
2. Oligo synthesis
3. Indel detection
4. Off-target detection

Metadata entries

Data and Meta Data Subgroup Metadata entries

File Edit View Insert Format Data Tools Add-ons Help All changes saved in Drive

Comparison of methods/sequencing data

	A	B	C	D
1	Metadata entry	Metadata ID	Metadata Type	Metadata example
2	Guide sequence	guide_sequence	string	ATCCGATCCGATCC
3	Targeting Strand	targeting_strand	{positive,negativ	positive
4	Guide start	guide_start	number	1500
5	Guide end	guide_end	number	1600
6	DSB position (sticky or blunt)	dsb_position	{sticky, blunt}	sticky
7	Genome build	genome_build	string	hg19
8	Chromosome	guide_chr	string	chr1
9	Target (gene vs. other genomic region)	guide_target	{gene body,enha	enhancer
10	Species	guide_species	string	human
11	PAM	guide_pam	string	AGG

Cleaning up metadata entries

Merging metadata entries that apply to different stages in the experimental process

Scoring each entry as High/Medium/Low importance

Cleaning up metadata entries

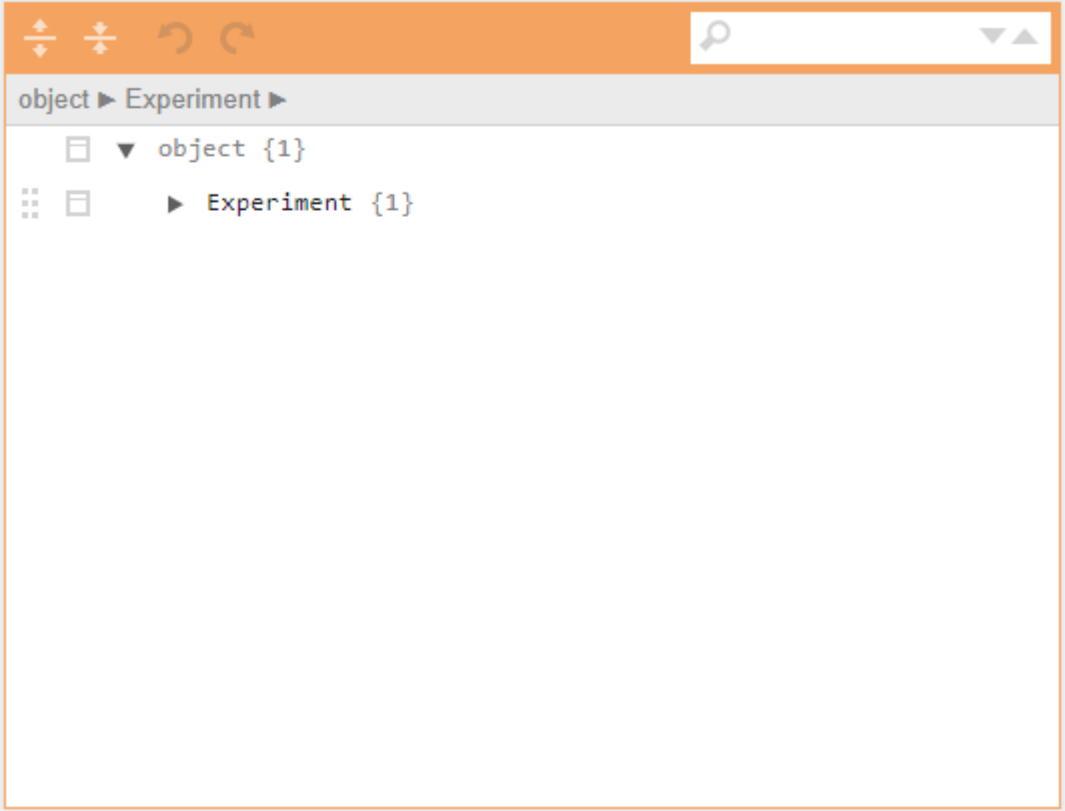
Data and Meta Data Subgroup Metadata entries

File Edit View Insert Format Data Tools Add-ons Help All changes saved in Drive

Please fill

	A	B	C	D	E	F	G	H	I
1	Metadata entry	Metadata ID	Metadata Type	Metadata example	Guide Design	Prediction of Off-Targets	Oligo Synthesis	Quantification of Indels	Discovery of Off-targets
2	Guide sequence	guide_sequence	string	ATCCGATCCGATCC	H	H			H
3	Targeting Strand	targeting_strand	{positive,negativ	positive	H	H			H
4	Guide start	guide_start	number	1500	H	H			H
5	Guide end	guide_end	number	1600	H	H			H
6	DSB position (sticky or blunt)	dsb_position	{sticky, blunt}	sticky	H	H			H
7	Genome build	genome_build	string	hg19	H	H		M	H
8	Chromosome	guide_chr	string	chr1	H	H			H
9	Target (gene vs. other genomic region)	guide_target	{gene body,enha	enhancer	L/M	M/L			M/L
10	Species	guide_species	string	human	L	H			H
11	PAM	guide_pam	string	AGG	H	H		H	H
12	Nuclease used	experiment_nuclease	string	Cas9	M	H		M	H
13	Scaffold (standard or modified, e.g. EF for Cas9)	guide_scaffold_sequenc	string	sequence...	H	M/L			M/L
14	Guide length	guide_length	number	20	M	H			H
15	Guide design Software used & version	guide_design_software	string	guide software v1.2.1	M	H			
16	Guide On-target score (efficiency)	guide_design_ontarget_s	number	87	L	H			
17	Guide Off-target score (avoidance of other genes)	guide_design_offtarget_s	number	32	L	H			
18	Guide filtering (by score or other criteria)	guid_design_filter	string	< 6 offtargets	L	H			
19	Expression system (U6 vs T7)	guide_design_expressio	string	U6	L				H

An intuitive and extensible storage prototype



An intuitive and extensible storage prototype

powered by ace

```
1  {
2    "Experiment": {
3      "Guides": {
4        "EMX1": {
5          "Guide Name": "EMX1",
6          "Reference Genome": "hg19",
7          "Guide Sequence": "AAAACCCCTTTTTGGGGAAAATTTC",
8          "Target Location": "chr2:10-20",
9          "Offtargets": [
10           "chr5:2000-2020",
11           "chr7:60-80"
12         ]
13       },
14       "VEGF2": {
15         "Guide Name": "VEGF2",
16         "Reference Genome": "hg19",
17         "Guide Sequence": "TTTACCCCTTTTTGGGGAAAATTTC",
18         "Target Location": "chr5:10-20",
19         "Offtargets": [
20           "chr2:2000-2020",
21           "chr3:60-80"
```

Ln: 26 Col: 2

Expected Outcomes

1. Comprehensive and well-vetted list of metadata entries
2. Recommendations for which point during the experimental process each metadata entry should be added
3. A common file standard for storing, accessing, and sharing data and metadata
4. A platform for benchmarking and comparing bioinformatic approaches and results

Why should I invest?

- Benchmarking and comparison of biochemical and computational approaches
- Transfer of experimental process and results
  - Within organization
  - Public databases (GEO, SRA, etc)
  - Supervisory agencies (FDA)
- Integration of products with a well-established genome editing process
- Exposure to potential customers/collaborators

Breakout topics

What metadata should be stored?

Who are key players?

How do we incentivize buy-in?

Questions?