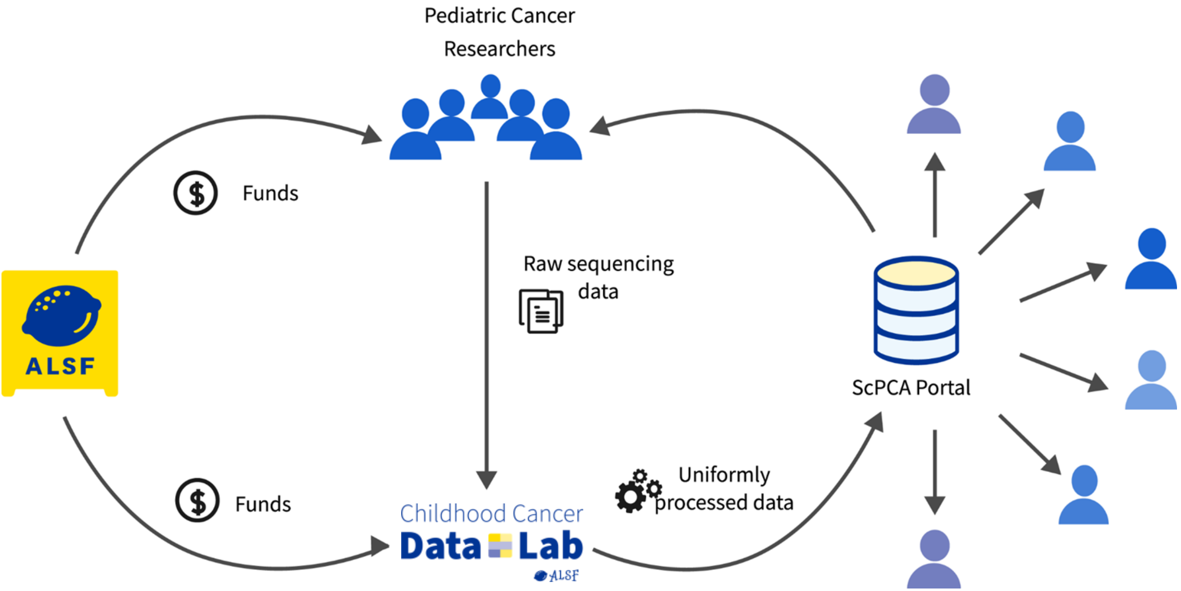
The Single-cell Pediatric Cancer Atlas Experience

Jaclyn N. Taroni, PhD
HRA Open Science Session
2022-04-20

Alex's Lemonade Stand Foundation for Childhood Cancer™

Childhood Cancer Data Lab ALSF

The Single-cell Pediatric Cancer Atlas (ScPCA) experience *from my perspective*



https://scpca.alexslemonade.org/

Highlights from the ScPCA Grant Guidelines

- A primary goal of this funding mechanism was to **produce a data resource**
  - An atlas of summarized gene expression and cell surface marker data from different pediatric cancer types and organ sites to be released in a timely manner
- Data sharing was a critical component of the application and any limitations with regards to data use and sharing per patient consent needed to be highlighted in the application
- Applicants must use a specific platform and a specific sequencing unit unless there was a scientifically justifiable reason to deviate from these specifications

RFA

Priorities and tradeoffs

Without building a data resource

- Prioritize best technology/kit for assaying samples in proposal
- Emphasize studies with the most rigorous design and highest potential for impact
- Data sharing might be less central
- Depending on your organization's policies, you may aim to fund a breadth of disease types or biological contexts

When building a data resource

- Prioritize uniformity because it makes building the resource more straightforward
- Overlapping disease types or biological contexts between projects lowers the barrier to creating validation sets
- Clarity around what will be transferred and when is extremely helpful

Excerpt from grant guidelines

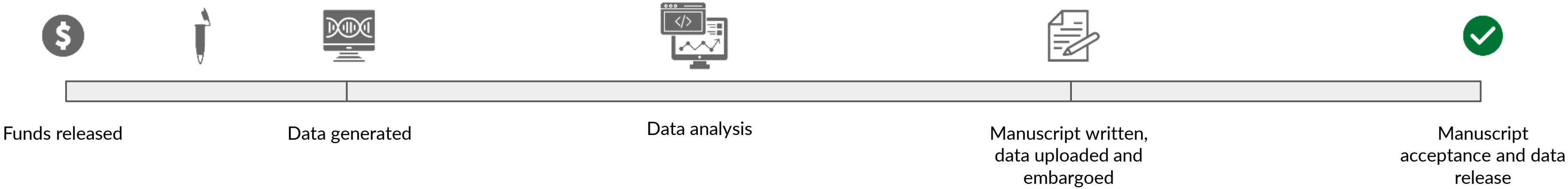
- Raw data must be transferred to Data Lab within 1 month of profiling on a rolling basis
- Agreements must allow the Data Lab to make summarized data available no more than 6 months after profiling
- Within 6 months of the end of the grant, grantees must deposit raw data in an appropriate repository

“Applicants are expected to make raw sequencing data (FASTQ files) available to the ALSF Childhood Cancer Data Lab (CCDL) within one month of profiling on a rolling basis. As part of the ScPCA, the CCDL will uniformly process the raw sequencing data through a common pipeline to estimate gene expression and where appropriate, the levels of cell surface markers. Any data transfer agreements, if required, must allow the CCDL to make gene expression and cell surface marker abundance estimates, as well as the de-identified sample-associated metadata, available without restriction no more than six months after raw sequencing data are generated. The goal of this requirement is to make sure that reusable data are released in a timely manner. The grantee is also required to deposit raw data in the appropriate repository (either NCBI SRA or EBI ENA) within six months of the conclusion of the grant.”

RFA

What does this mean for the data release timeline?

Typical process (based on existing norms)



Funds released

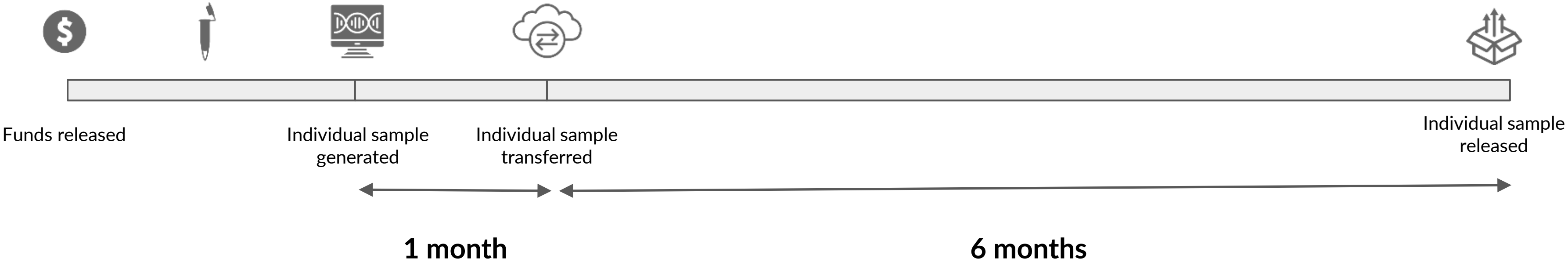
Data generated

Data analysis

Manuscript written,
data uploaded and
embargoed

Manuscript
acceptance and data
release

“Rolling release” interpretation of ScPCA guidelines



Funds released

Individual sample
generated

Individual sample
transferred

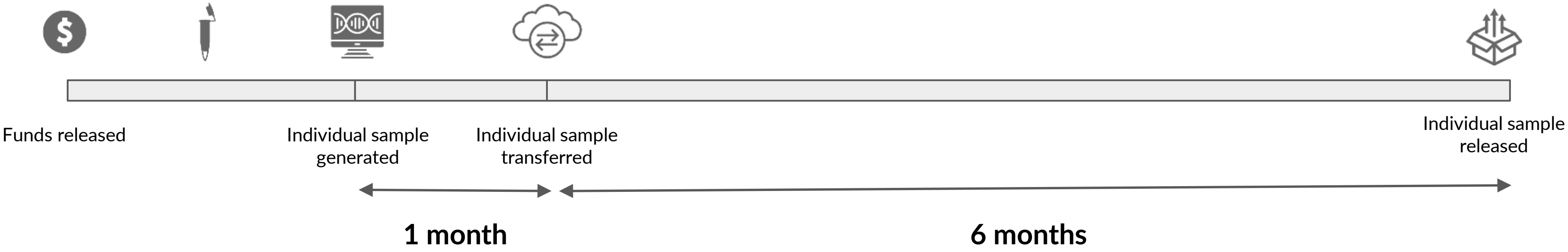
Individual sample
released

1 month

6 months

What does this mean for the data release timeline?

“Rolling release” interpretation of ScPCA guidelines



Funds released

Individual sample generated

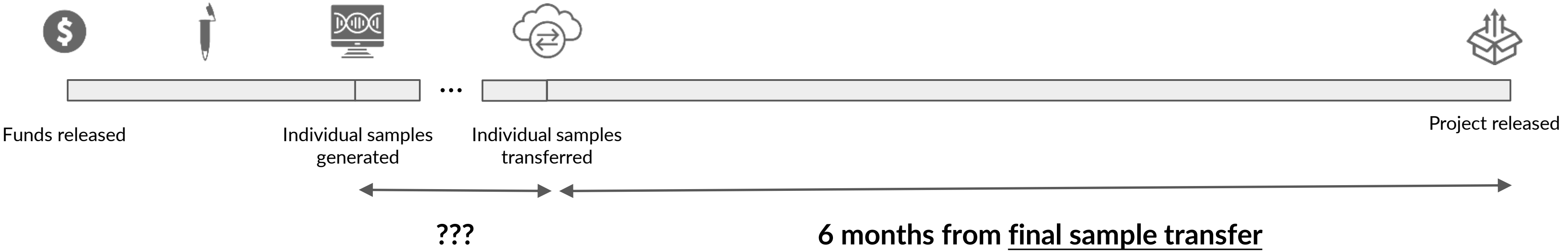
Individual sample transferred

Individual sample released

1 month

6 months

“Batched release” interpretation of ScPCA guidelines



Funds released

Individual samples generated

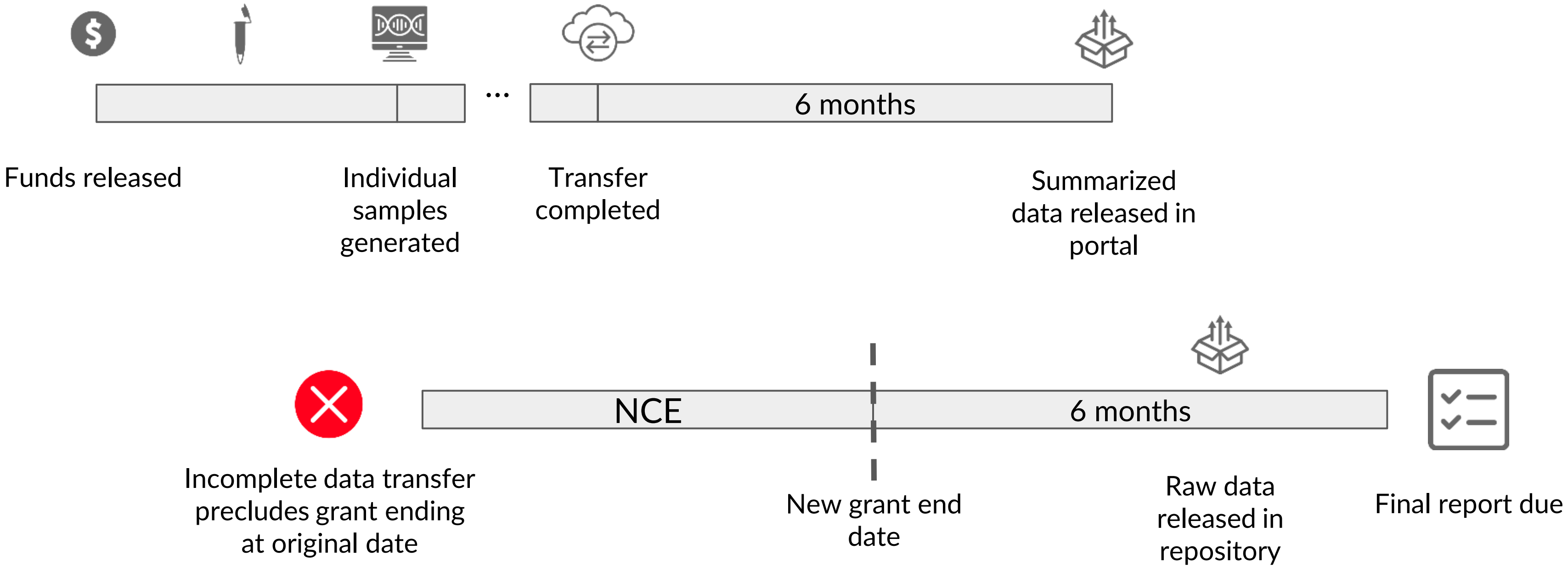
Individual samples transferred

Project released

???

6 months from final sample transfer

Grant reporting timelines are not entirely coupled to transfer and release timelines



…

6 months

Funds released

Individual samples generated

Transfer completed

Summarized data released in portal

NCE

6 months

Incomplete data transfer precludes grant ending at original date

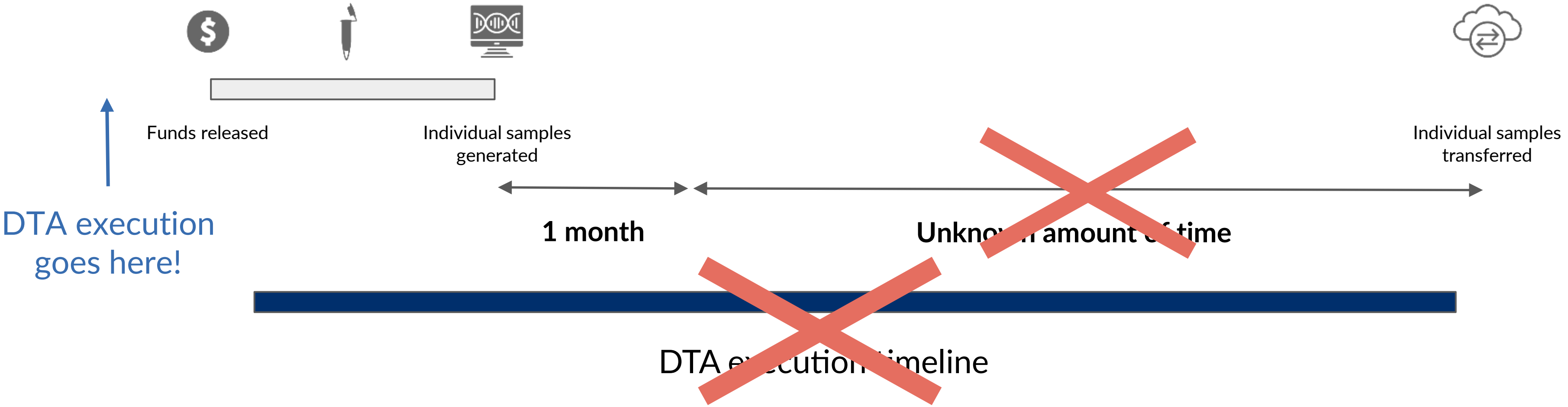
New grant end date

Raw data released in repository

Final report due

⚡ Problems and potential solutions ⚡

Problem 1: If a data transfer agreement is required, the time needed to execute it may preclude compliance with the rest of the terms.



Funds released

Individual samples generated

Individual samples transferred

1 month

~~Unknown amount of time~~

~~DTA execution timeline~~

DTA execution goes here!

Potential solution: Require data transfer agreements to be executed prior to the release of funds (and therefore data generation)

Problem 2: If you’re not used to transferring data on a rolling basis, you may not have a standard or preferred method for transferring data outside of submission to a repository at the time of publication.

We asked investigators if their institutions had a standard way of transferring raw data files of this nature (e.g., Globus). Not everyone had experience with transferring files this large or were aware of their institution’s preferred method.

Potential solution: Have *data recipients* standardize method of transfer and document it

Problem 3: How does the data recipient know what to expect or when transfer is completed?

If funded investigators don’t have samples in hand, the number of samples and characteristics of samples may change over time

Reading tables and free text that are not uniformly formatted can lead to erroneous conclusions about expected samples

Potential solution: Create a portion of the application and progress reports that is standardized and specifically meant to be consumed by data recipients

Problem 4: The type of data, and maturity of that technology, matter when we talk about sharing.

If an end goal is to release processed data, someone needs to figure out how to process it. There may be new or emerging methods for processing that data that require testing. The longer transfer of a breadth of samples takes, the more the timeline for benchmarking and therefore processing is extended.

Potential solution: I don’t have one! Just be aware of the complexity that this adds, *especially* if the people building the resource are other grantees rather than internal to your organization.

Takeaways

- Recognize that you may look for different characteristics in grants you are funding with the express purpose of creating a data resource as compared to other mechanisms
- Be as explicit as possible in communicating requirements and release timelines with investigators
- Design processes to smooth the way for agreements, transfer, and “data accounting” upfront
- Consider how the *type of data* being shared may introduce complexities into data transfer, processing, and release
- Take extra care if the folks involved in data release (e.g., building a portal) are grantees that were selected independently of the data generator selection process

Thank you!