AgBioData

Data reuse working group

First steps and goals

Chair: James Koltes
Members: Alenka Hafner (co-chair), Chris Elsik, Boas Pucker, Cecilia Deng, Peter Harrison, Ted Kalbfleisch, Elsa Herminia Quezada Rodríguez, Victoria DeLeo, Bruna Petry, Anne Thessen

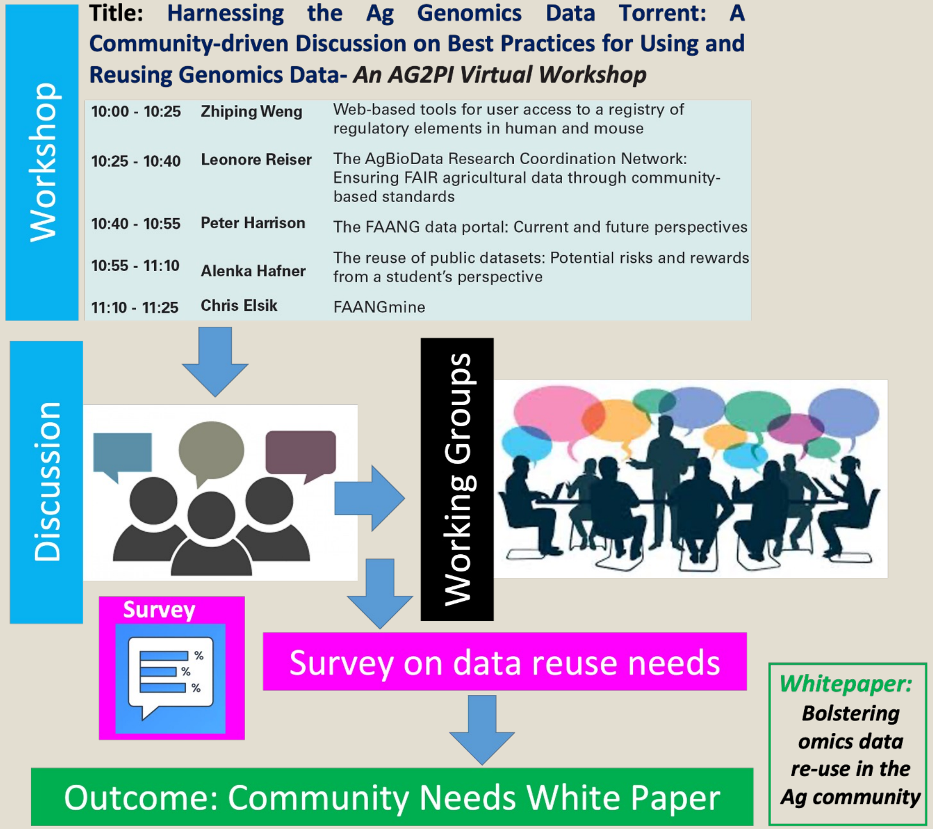
Motivation and beginnings:

- AG2PI seed grant in 2021
- Exponential growth of publicly available genomics and epigenetics data
- Data are often underutilized due to insufficient metadata and other challenges related to the FAIR data standards
- Intermediate details from supplemental files and initial data are frequently not captured
- **Objective of WG: to identify bottlenecks in data reuse and critical needs to propose solutions**

AG2PI

Agricultural Genome to Phenome Initiative

Figure 1: Overview of Project Activities



Title: Harnessing the Ag Genomics Data Torrent: A Community-driven Discussion on Best Practices for Using and Reusing Genomics Data- *An AG2PI Virtual Workshop*

10:00 - 10:25	Zhiping Weng	Web-based tools for user access to a registry of regulatory elements in human and mouse
10:25 - 10:40	Leonore Reiser	The AgBioData Research Coordination Network: Ensuring FAIR agricultural data through community-based standards
10:40 - 10:55	Peter Harrison	The FAANG data portal: Current and future perspectives
10:55 - 11:10	Alenka Hafner	The reuse of public datasets: Potential risks and rewards from a student's perspective
11:10 - 11:25	Chris Elsik	FAANGmine

Workshop
Discussion
Working Groups
Survey
Survey on data reuse needs
Outcome: Community Needs White Paper
Whitepaper: *Bolstering omics data re-use in the Ag community*

Credit: James Koltes

AgBioData

AgBioData Data Reuse WG

- Met at PAG, first meeting on Feb 2nd, 2023
- **This working group will:**
  - Identify **bottlenecks** in data reuse in livestock and plant communities based on discussions and community surveys.
  - Identify **critical needs** that will help improve data reuse and promote more FAIR data.
  - Identify **missed opportunities** to improve data sharing
  - Publish a **white paper** describing these bottlenecks and needs.

Working group members

James Koltes (chair)	Iowa State	Assistant Professor
Alenka Hafner (co-chair)	Penn State	PhD candidate in Plant Biology
Boas Pucker	TU Braunschweig	Assistant Professor
Cecilia Deng	The New Zealand Institute for Plant and Food Research Limited	Senior scientist
Peter Harrison	EMBL-European Bioinformatics Institute	Genome Analysis Team Leader
Chris Elsik	University of Missouri	Professor
Ted Kalbfleisch	University of Kentucky	Associate Professor
Elsa Herminia Quezada Rodríguez	Universidad Autónoma Metropolitana - México	Postdoc
Victoria DeLeo	Bowery Farming	Plant Scientist
Bruna Petry	Iowa State University	Postdoc Research Associate
Anne Thessen	University of Colorado Anschutz	Assoc Prof

White paper draft sections

1. Introduction

2. Barriers/risks/limitations of reuse

a. Data standards

I. Data quality

ii. Experimental standards

b. Metadata and ontologies

c. Data interoperability

d. Data ownership propriety

e. User skill level

f. Resource availability

3. Data availability

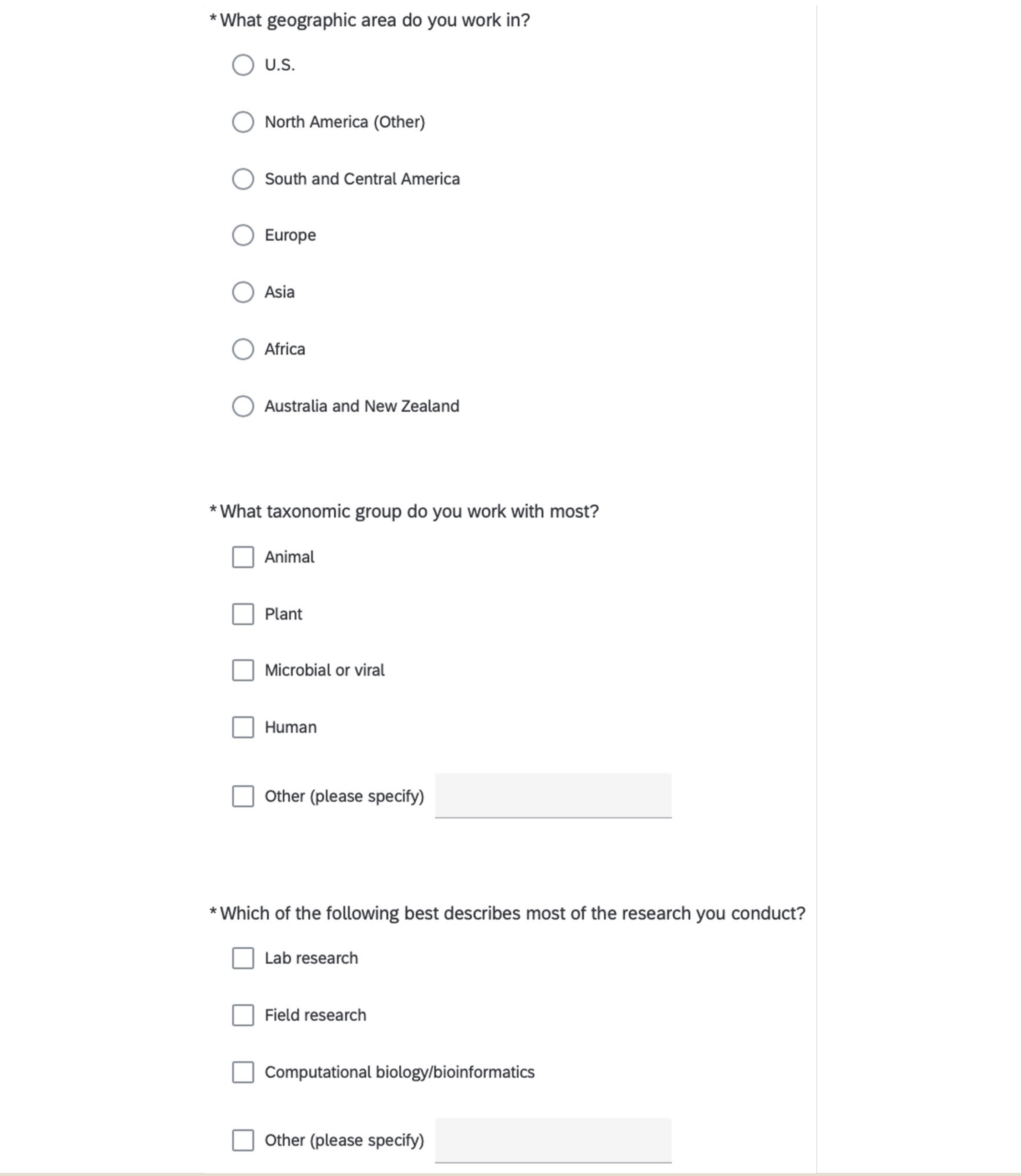
4. Other/emerging data types

5. Genotype to phenotype

6. Equity and inclusion in data reuse

7. Future of data reuse

COMMUNITY SURVEY



*What geographic area do you work in?

○ U.S.

○ North America (Other)

○ South and Central America

○ Europe

○ Asia

○ Africa

○ Australia and New Zealand

*What taxonomic group do you work with most?

□ Animal

□ Plant

□ Microbial or viral

□ Human

□ Other (please specify)

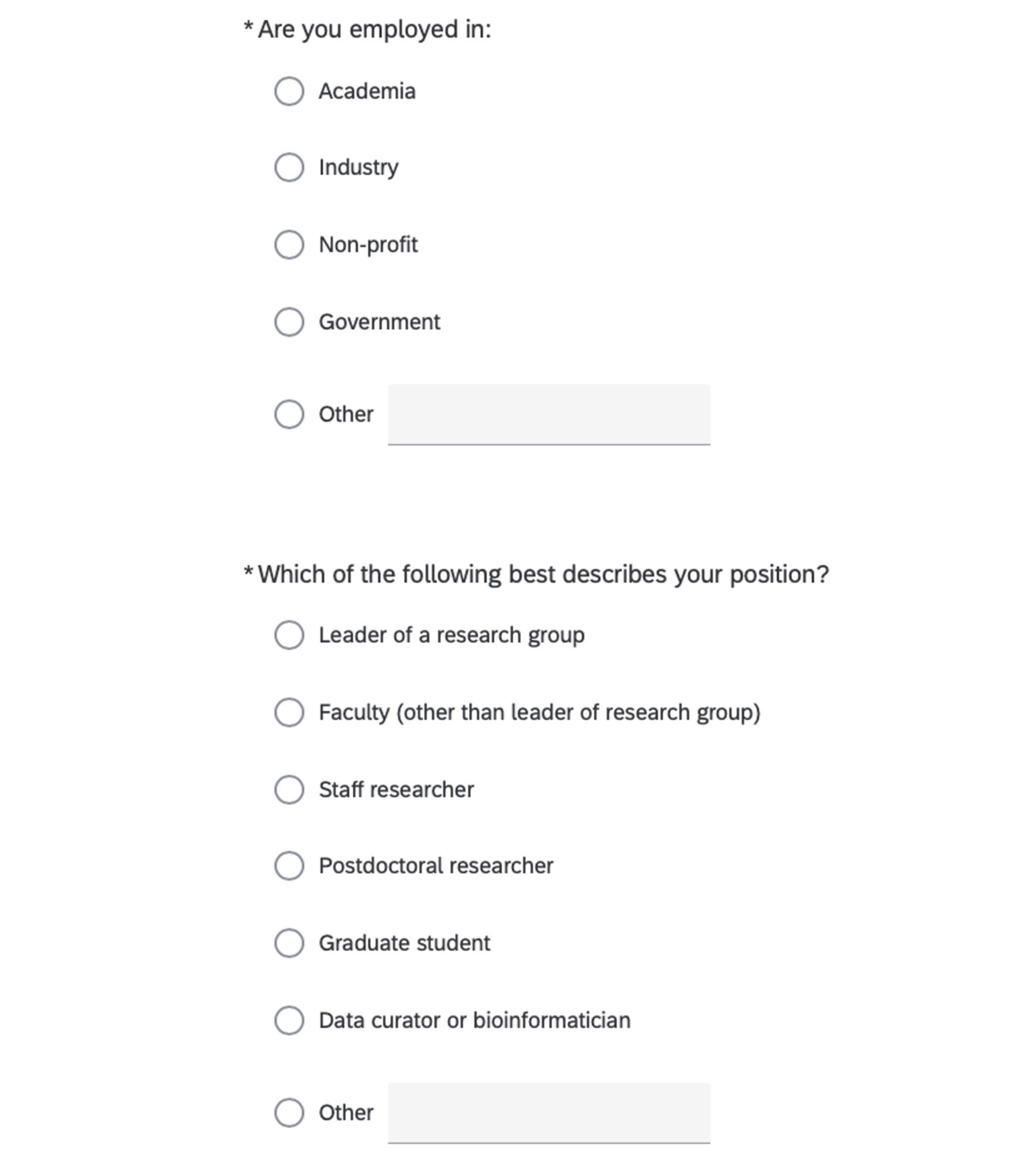
*Which of the following best describes most of the research you conduct?

□ Lab research

□ Field research

□ Computational biology/bioinformatics

□ Other (please specify)



*Are you employed in:

○ Academia

○ Industry

○ Non-profit

○ Government

○ Other

*Which of the following best describes your position?

○ Leader of a research group

○ Faculty (other than leader of research group)

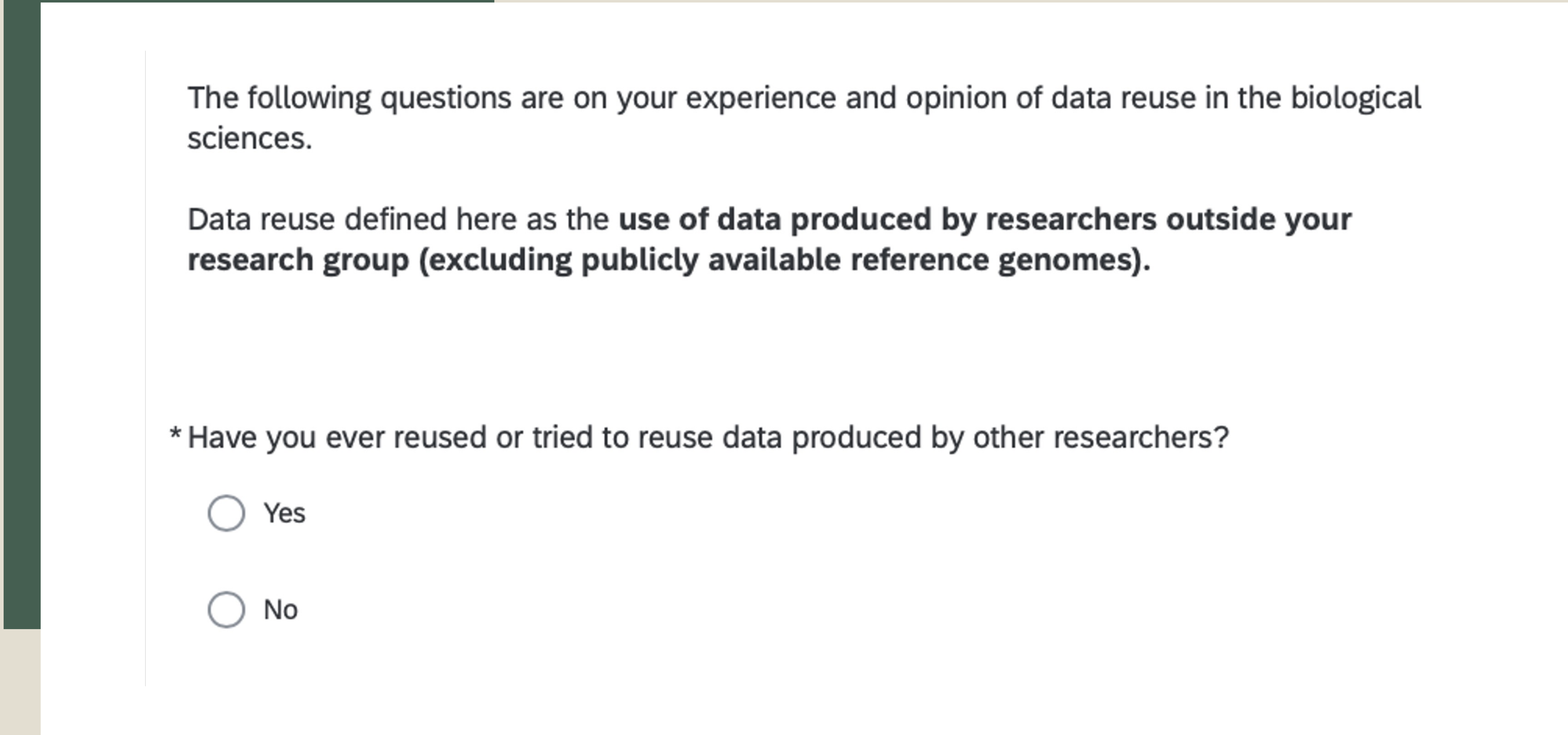
○ Staff researcher

○ Postdoctoral researcher

○ Graduate student

○ Data curator or bioinformatician

○ Other



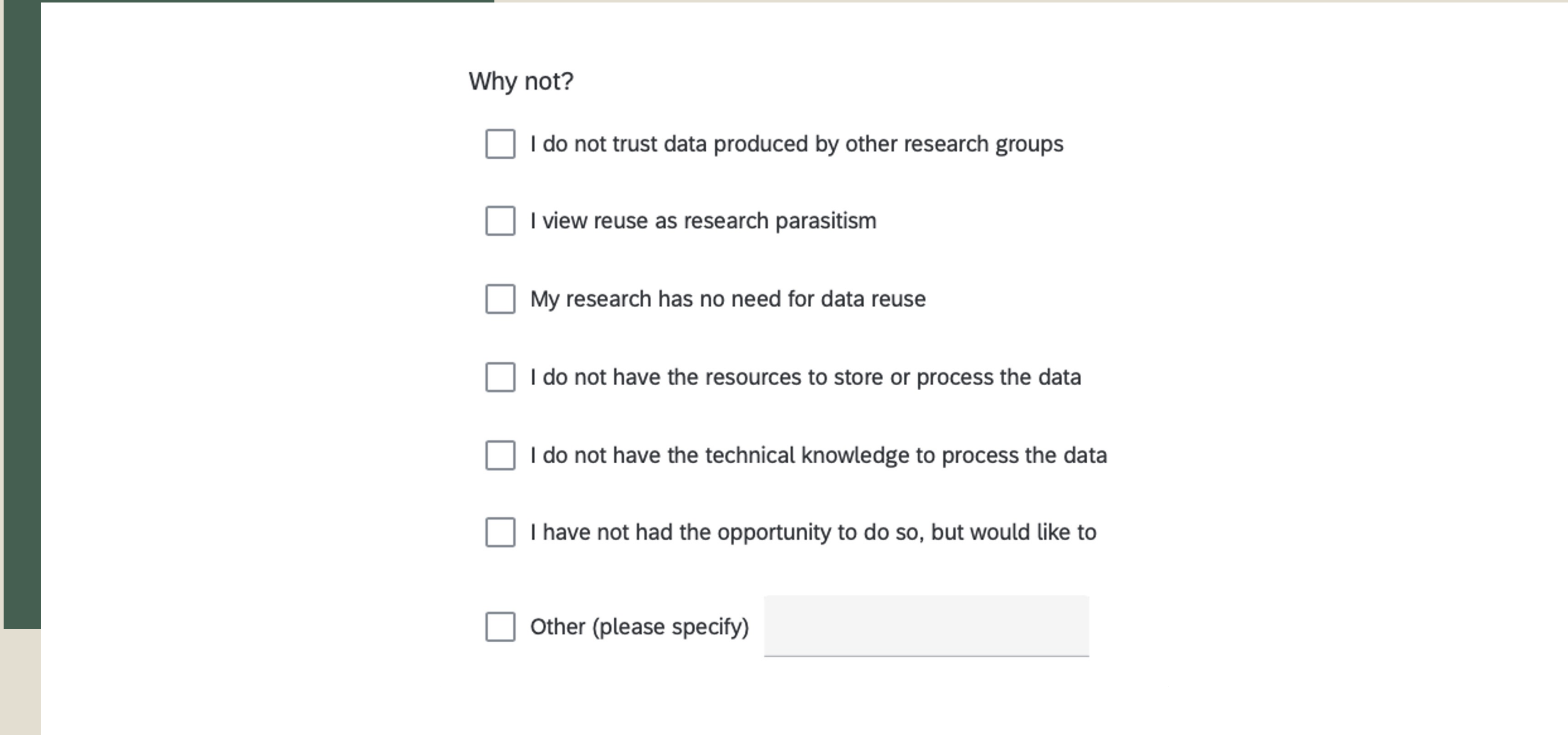
The following questions are on your experience and opinion of data reuse in the biological sciences.

Data reuse defined here as the **use of data produced by researchers outside your research group (excluding publicly available reference genomes).**

* Have you ever reused or tried to reuse data produced by other researchers?

○ Yes

○ No



Why not?

□ I do not trust data produced by other research groups

□ I view reuse as research parasitism

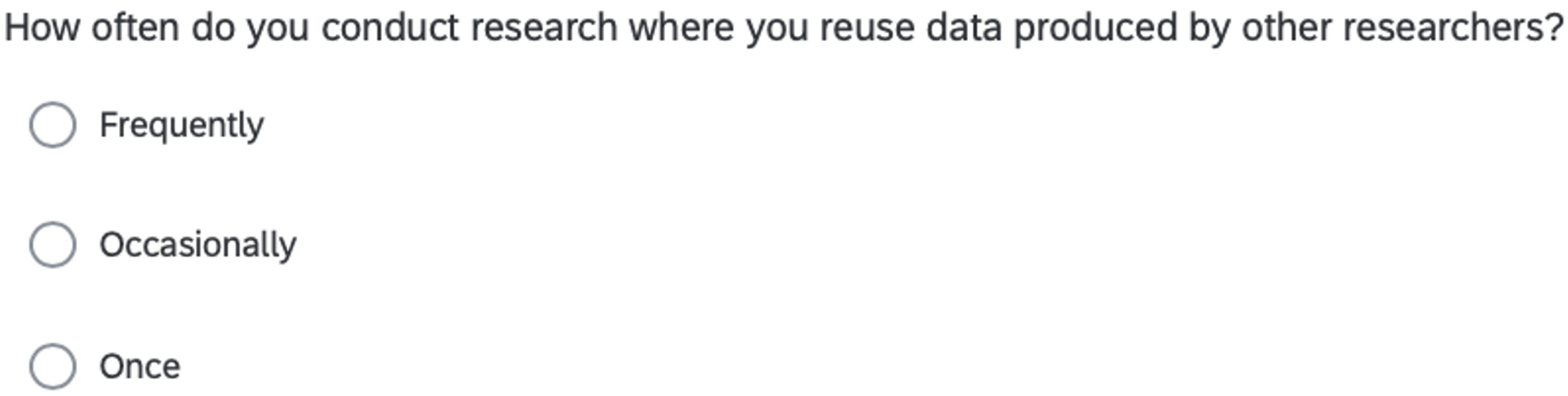
□ My research has no need for data reuse

□ I do not have the resources to store or process the data

□ I do not have the technical knowledge to process the data

□ I have not had the opportunity to do so, but would like to

□ Other (please specify)

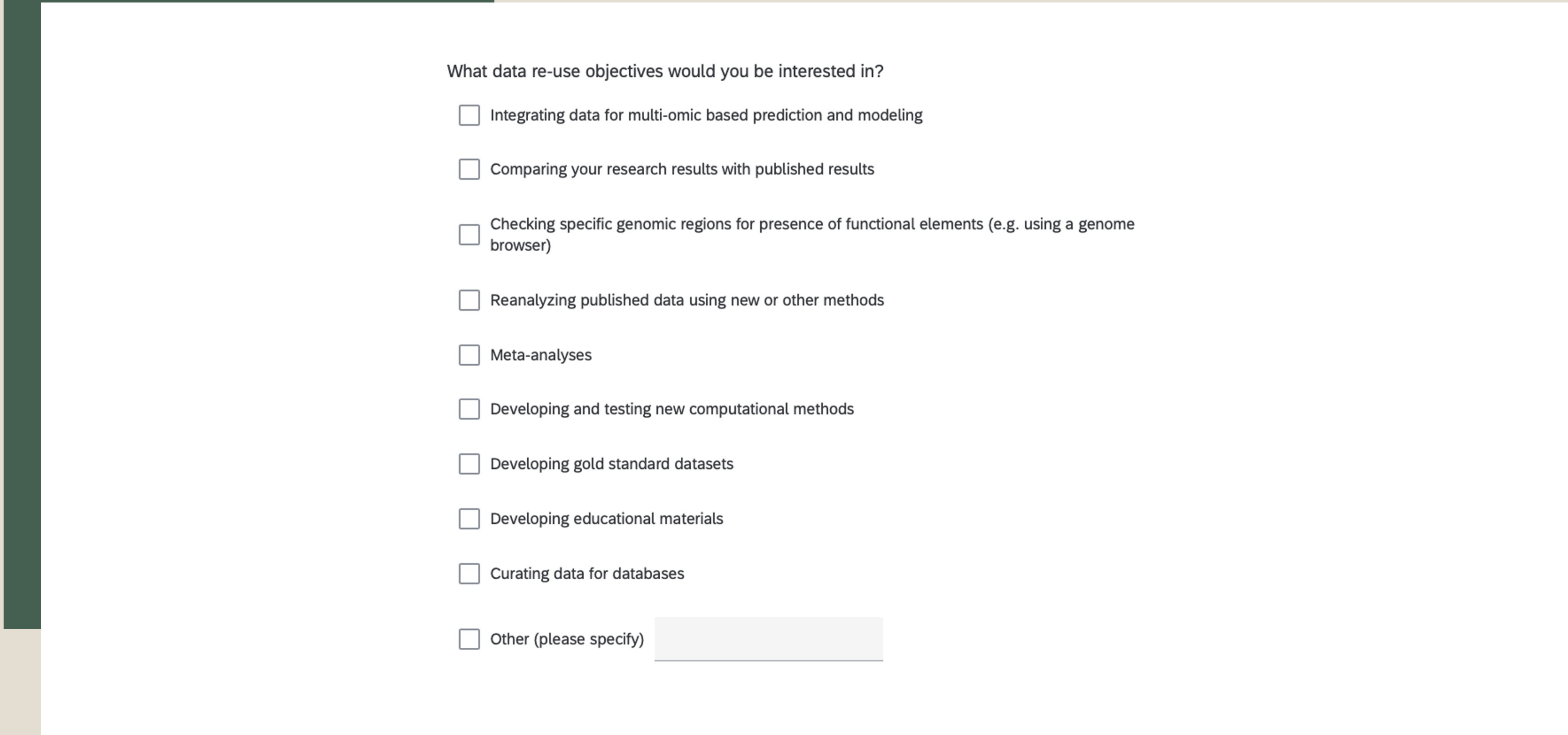


How often do you conduct research where you reuse data produced by other researchers?

○ Frequently

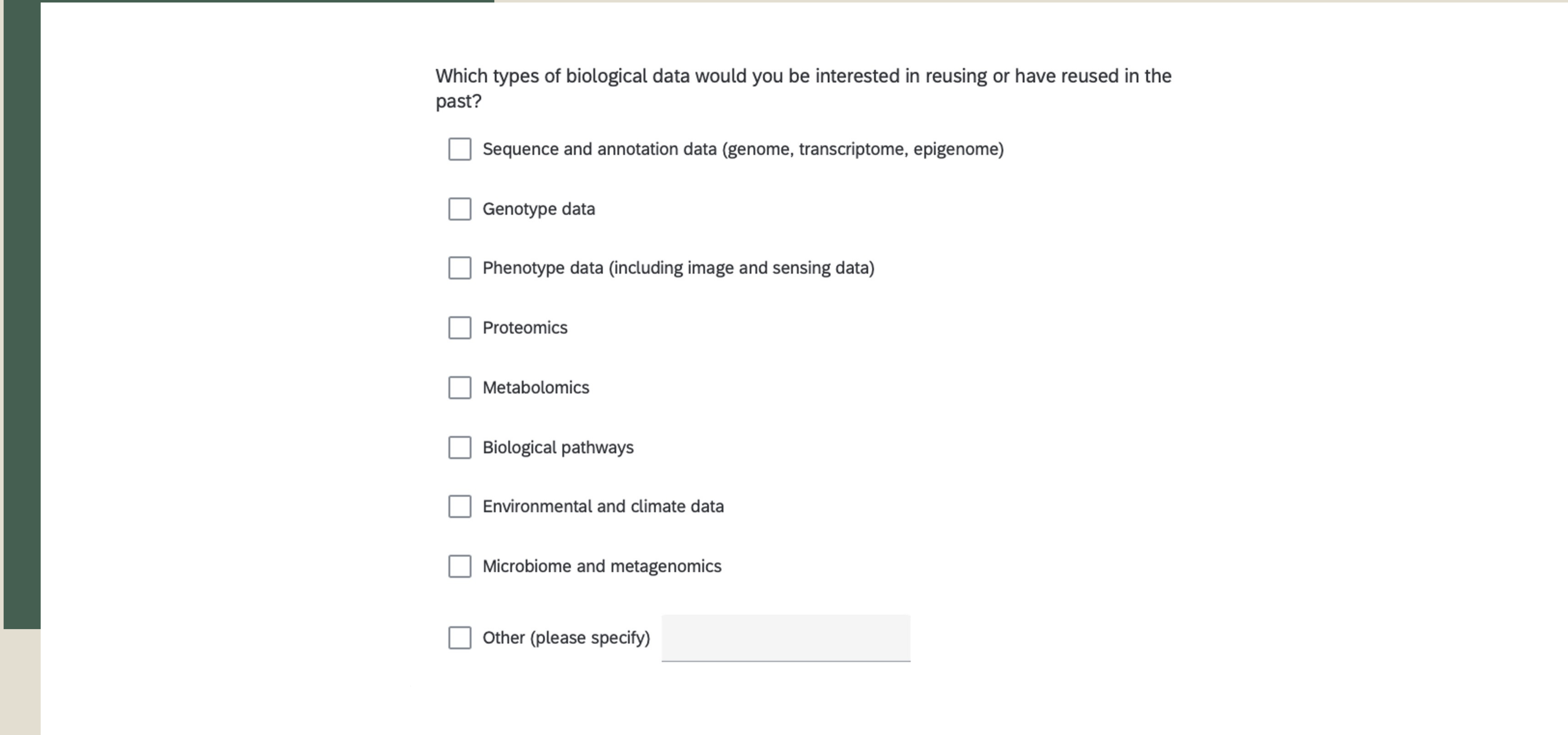
○ Occasionally

○ Once



What data re-use objectives would you be interested in?

- □ Integrating data for multi-omic based prediction and modeling
- □ Comparing your research results with published results
- □ Checking specific genomic regions for presence of functional elements (e.g. using a genome browser)
- □ Reanalyzing published data using new or other methods
- □ Meta-analyses
- □ Developing and testing new computational methods
- □ Developing gold standard datasets
- □ Developing educational materials
- □ Curating data for databases
- □ Other (please specify)



Which types of biological data would you be interested in reusing or have reused in the past?

□ Sequence and annotation data (genome, transcriptome, epigenome)

□ Genotype data

□ Phenotype data (including image and sensing data)

□ Proteomics

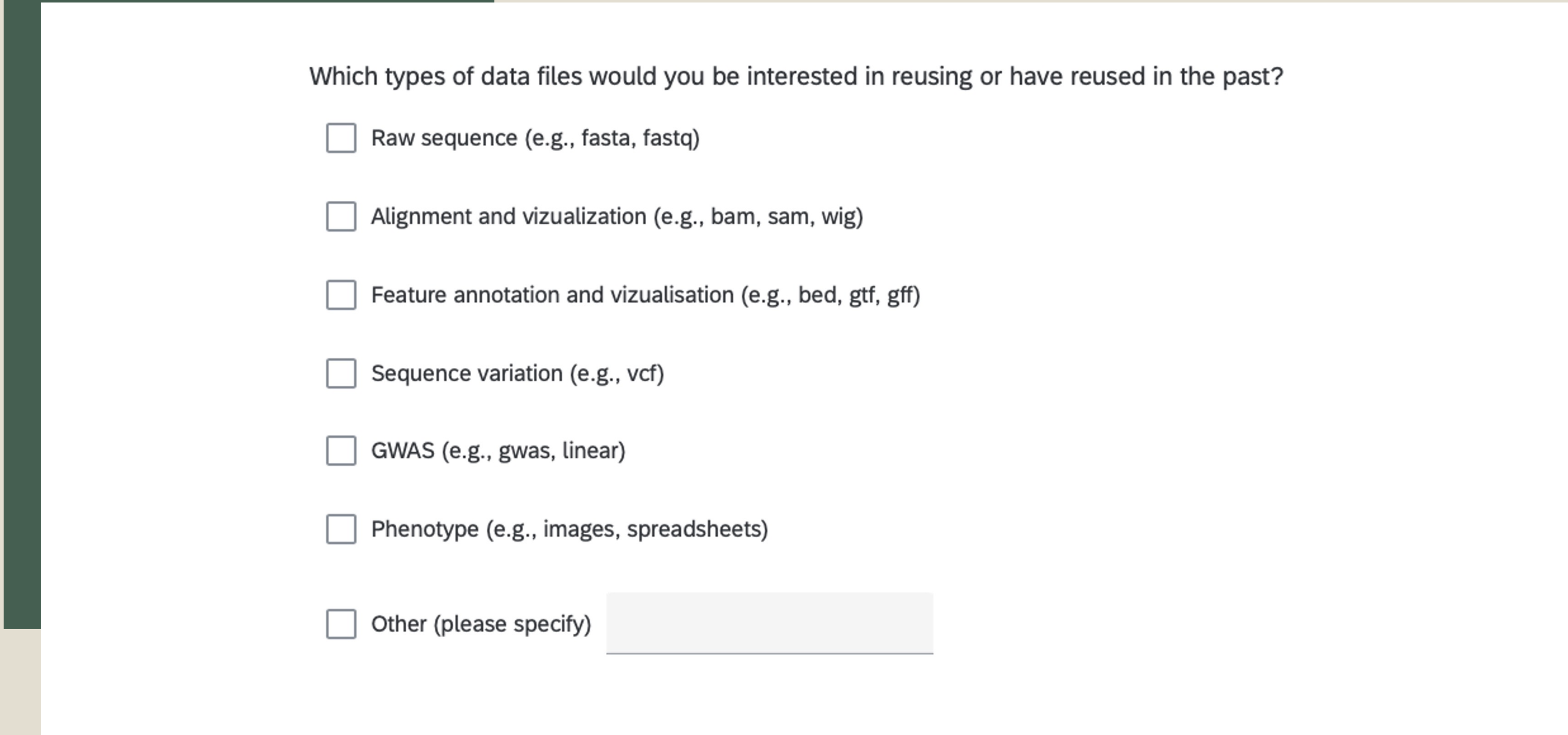
□ Metabolomics

□ Biological pathways

□ Environmental and climate data

□ Microbiome and metagenomics

□ Other (please specify) ____________



Which types of data files would you be interested in reusing or have reused in the past?

□ Raw sequence (e.g., fasta, fastq)

□ Alignment and vizualization (e.g., bam, sam, wig)

□ Feature annotation and vizualisation (e.g., bed, gtf, gff)

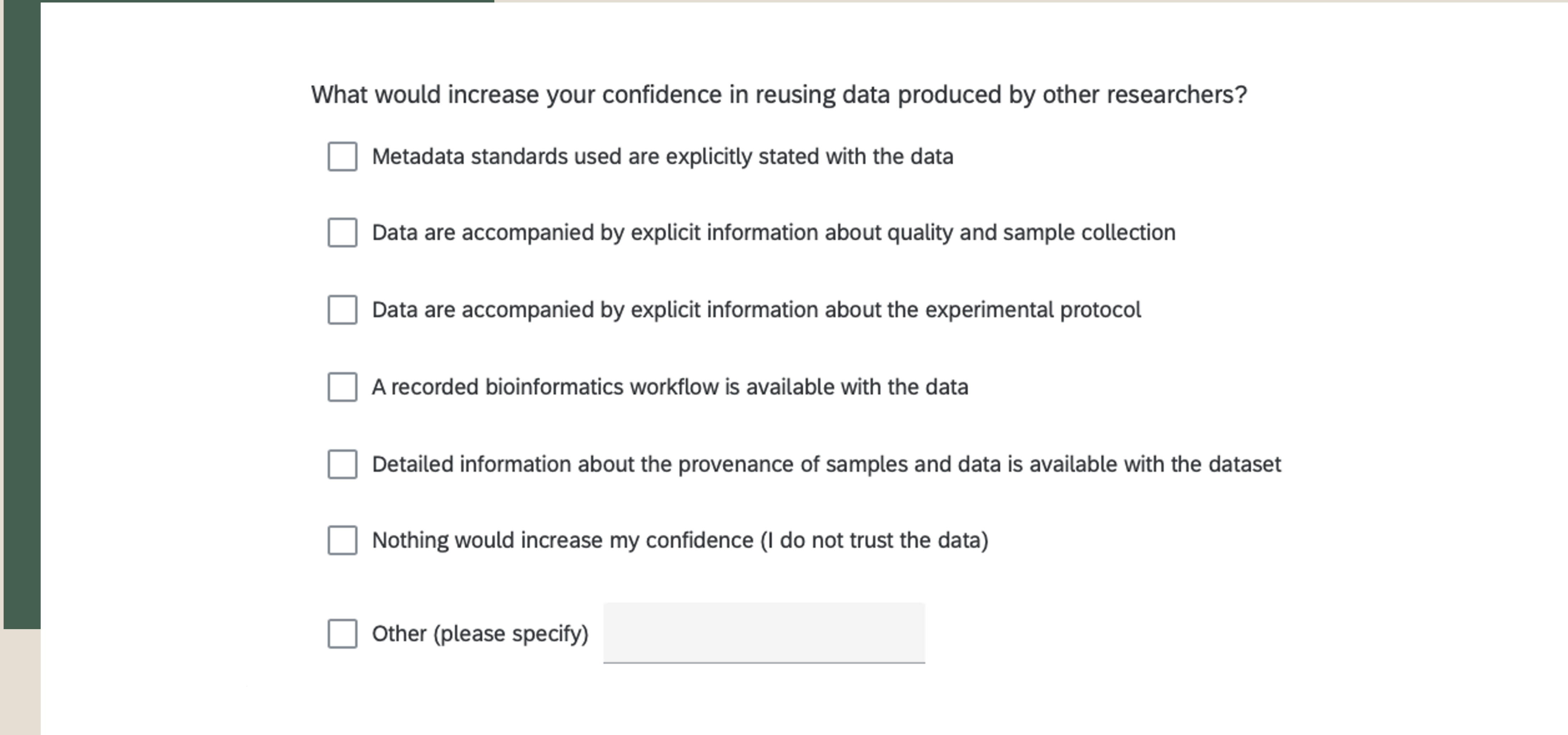
□ Sequence variation (e.g., vcf)

□ GWAS (e.g., gwas, linear)

□ Phenotype (e.g., images, spreadsheets)

□ Other (please specify)

AgBioData



What would increase your confidence in reusing data produced by other researchers?

☐ Metadata standards used are explicitly stated with the data

☐ Data are accompanied by explicit information about quality and sample collection

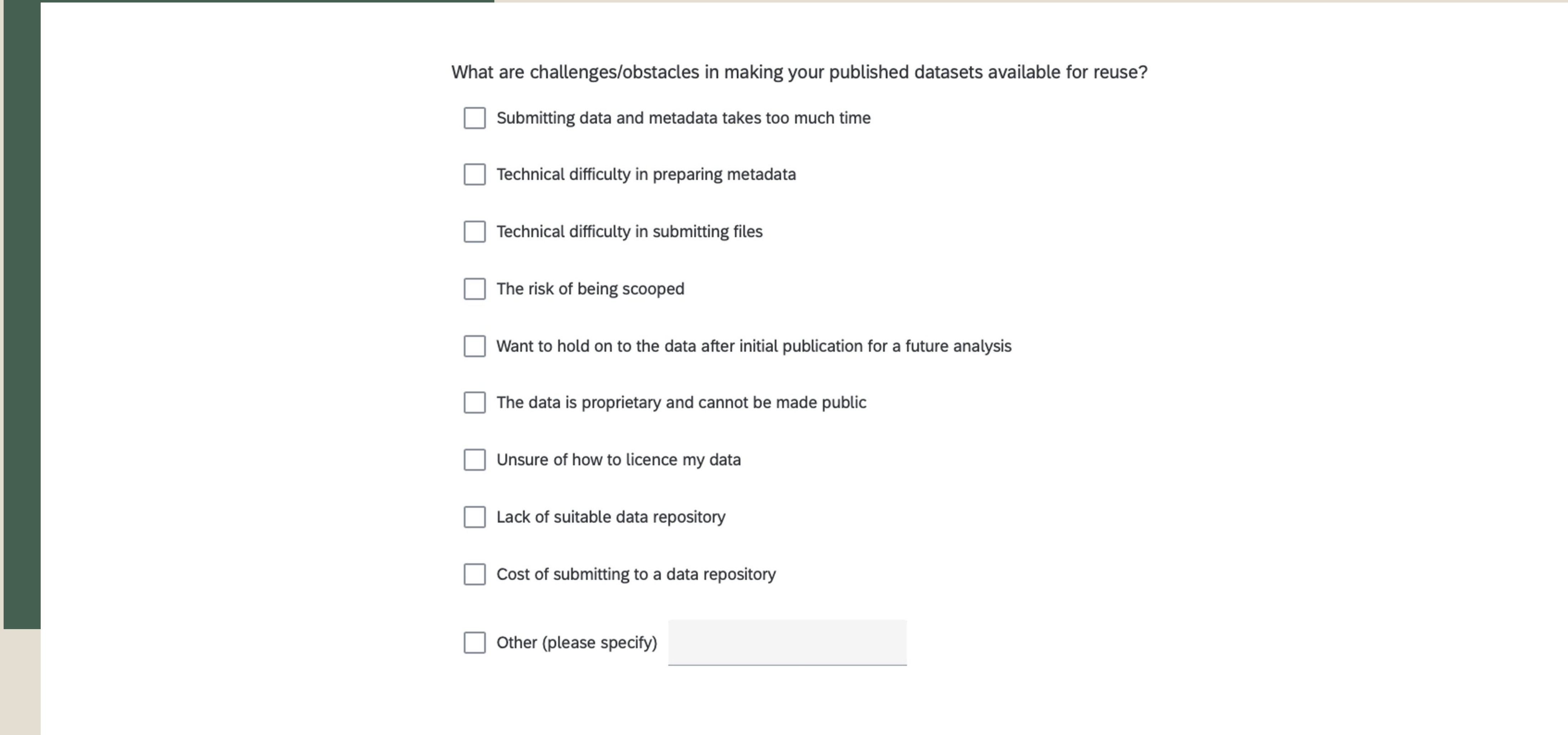
☐ Data are accompanied by explicit information about the experimental protocol

☐ A recorded bioinformatics workflow is available with the data

☐ Detailed information about the provenance of samples and data is available with the dataset

☐ Nothing would increase my confidence (I do not trust the data)

☐ Other (please specify) ____________



What are challenges/obstacles in making your published datasets available for reuse?

□ Submitting data and metadata takes too much time

□ Technical difficulty in preparing metadata

□ Technical difficulty in submitting files

□ The risk of being scooped

□ Want to hold on to the data after initial publication for a future analysis

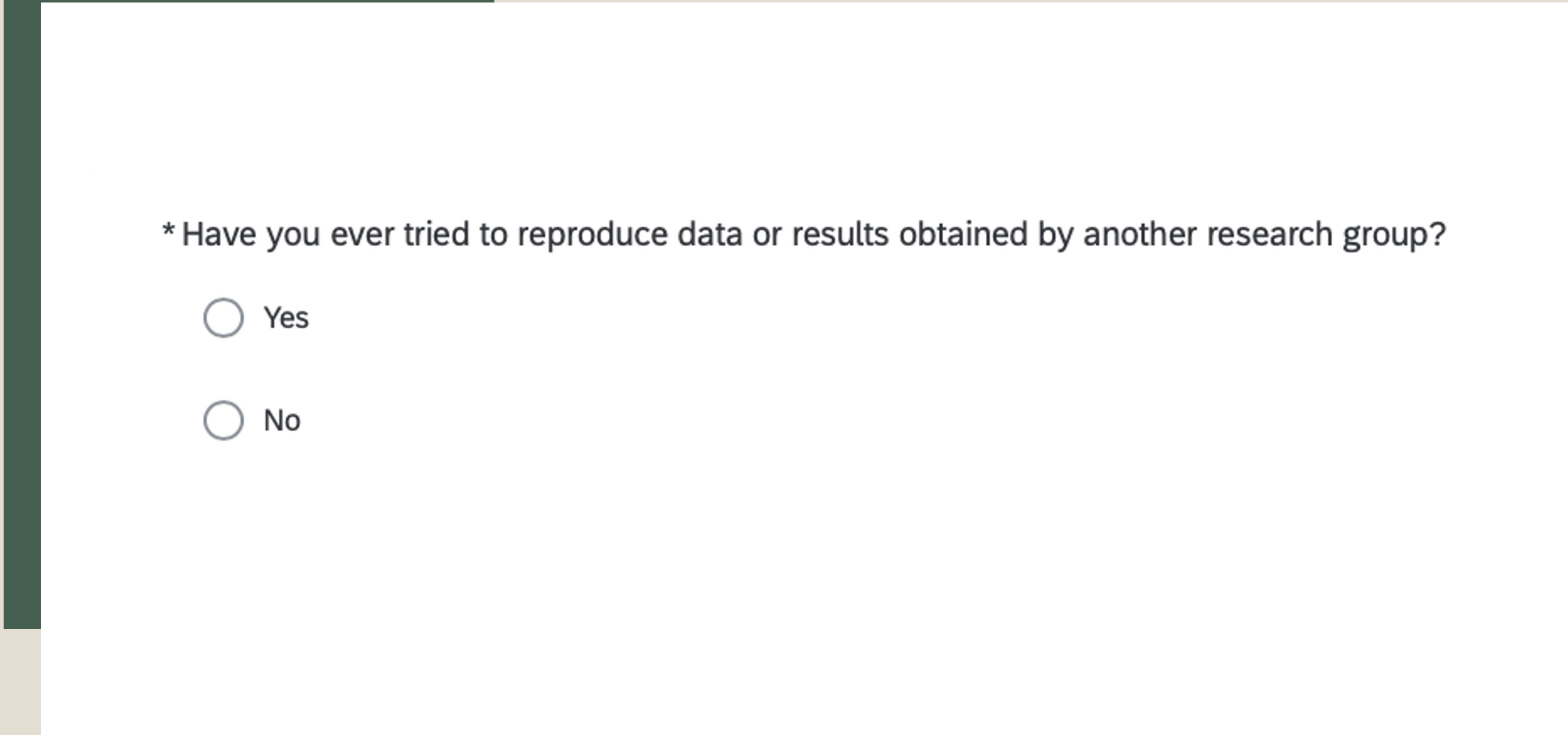
□ The data is proprietary and cannot be made public

□ Unsure of how to licence my data

□ Lack of suitable data repository

□ Cost of submitting to a data repository

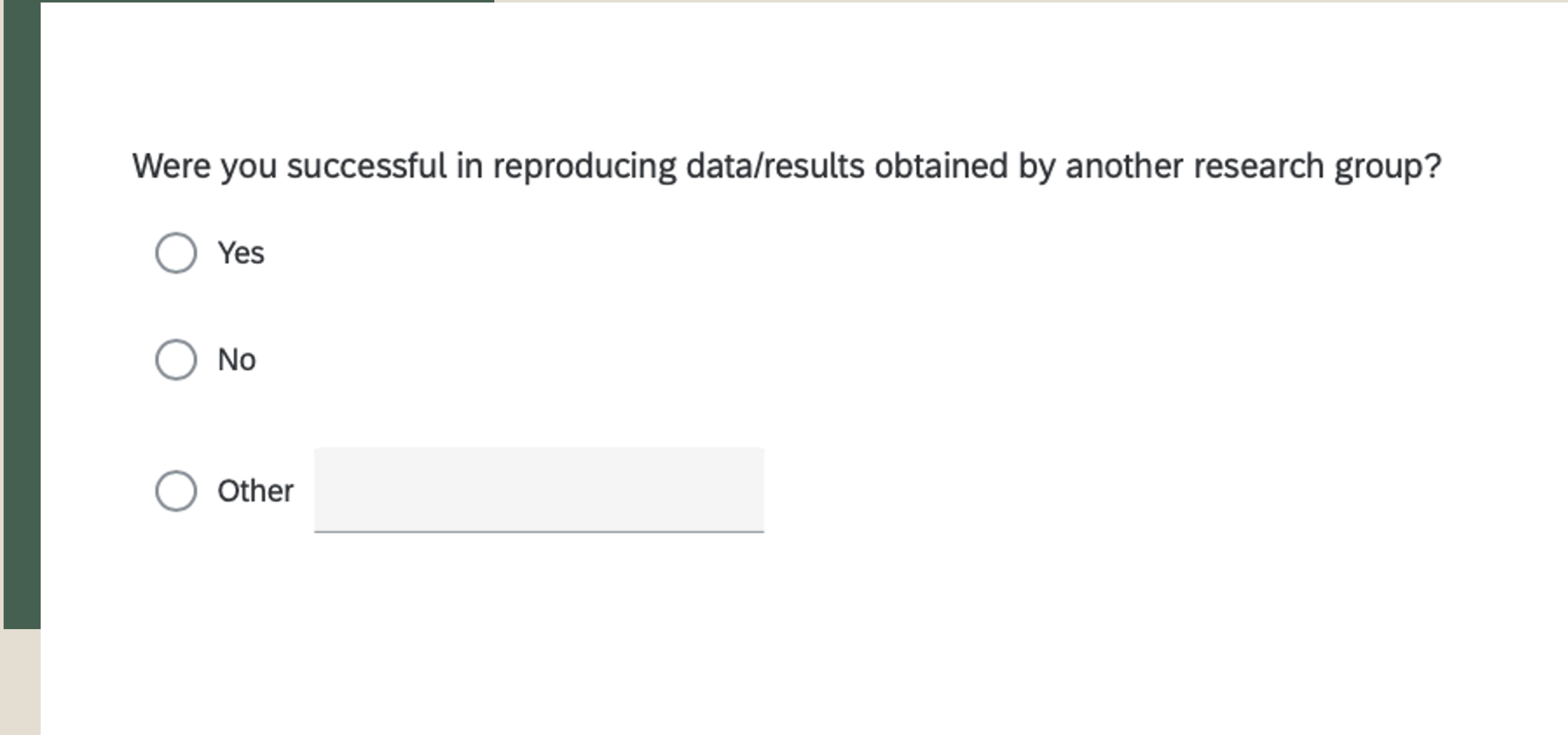
□ Other (please specify)



* Have you ever tried to reproduce data or results obtained by another research group?

○ Yes

○ No

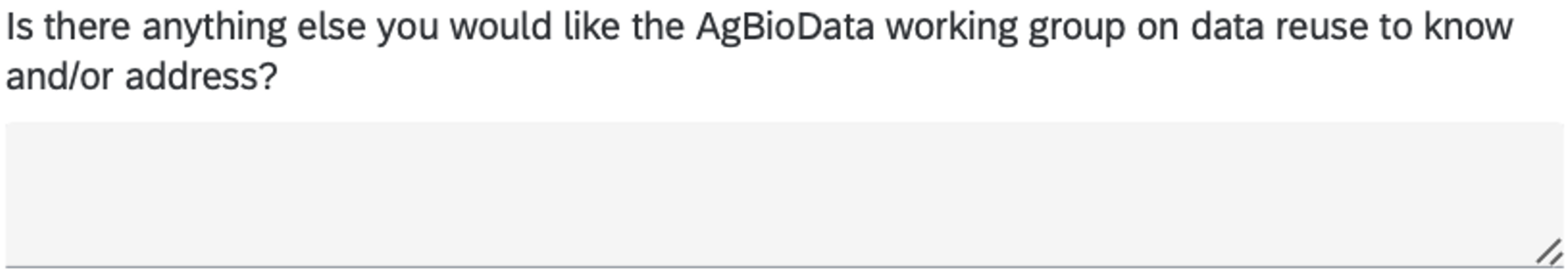


Were you successful in reproducing data/results obtained by another research group?

○ Yes

○ No

○ Other



Is there anything else you would like the AgBioData working group on data reuse to know and/or address?

Future directions

- Phenomics
- Emerging data types and methods
- Machine learning - assisted metadata curation
- Other ideas?

Interested in joining? Have questions or ideas?

Contact me (ahafner@psu.edu) or James (jekoltes@iastate.edu)!

AgBioData