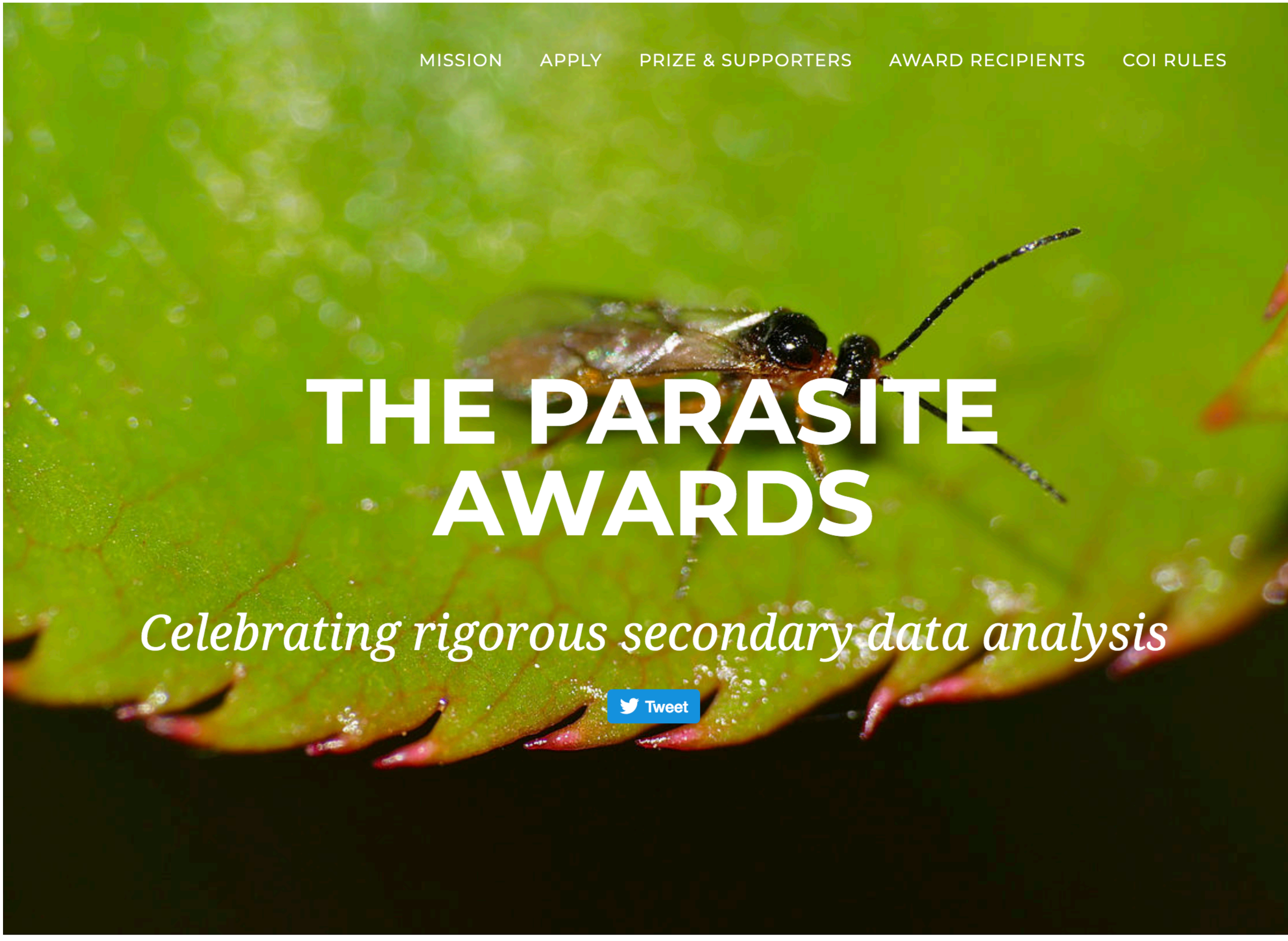
Challenges of working with controlled access datasets in human genetics

Kaur Alasoo
Research Fellow in Bioinformatics
9 April 2019

I’m a proud research parasite.

‘A second concern held by some is that a new class of research person will emerge — people who had nothing to do with the design and execution of the study but **use another group’s data for their own ends, possibly stealing from the research productivity planned by the data gatherers, or even use the data to try to disprove what the original investigators had posited**. There is concern among some front-line researchers that the system will be taken over by what some researchers have characterized as “research parasites.”’

https://www.nejm.org/doi/full/10.1056/NEJMe1516564



THE PARASITE AWARDS

Celebrating rigorous secondary data analysis

Molecular biology has a strong culture of open data sharing

- Many open access databases
    - European Nucleotide Archive (ENA)
    - Short Read Archive (SRA)
    - Gene Expression Omnibus (GEO)
    - ArrayExpress
- Submitting data is *usually* a pre-requisite for publication

The following data sets were generated

1 **European Nucleotide Archive**
K Alasoo, J Rodrigues, J Danesh, DF Freitag, DS Paul, DJ Gaffney (2017)
Genetic effects on promoter usage are highly context-specific and contribute to complex traits.

2 **European Genome-phenome Archive**
K Alasoo, J Rodrigues, J Danesh, DF Freitag (2017)
Genetic effects on promoter usage are highly context-specific and contribute to complex traits.

The following previously published data sets were used

1 **European Nucleotide Archive**
K Alasoo, J Rodrigues, S Mukhopadhyay, AJ Knights, AL Mann, K Kundu, C Hale, G Dougan, DJ Gaffney (2017)
Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response.

2 **European Genome-phenome Archive**
K Alasoo, J Rodrigues, S Mukhopadhyay, AJ Knights, AL Mann, K Kundu, C Hale, G Dougan, DJ Gaffney (2017)
Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response.

3 **European Nucleotide Archive**
H Kilpinen, A Goncalves (2017)
Common genetic variation drives molecular heterogeneity in human iPSCs.

4 **European Genome-phenome Archive**
H Kilpinen, A Goncalves (2017)
Common genetic variation drives molecular heterogeneity in human iPSCs.

…but most human genetic data cannot be shared openly

- Genetic data cannot be anonymised, it is always possible to re-identify individuals based on their genotype

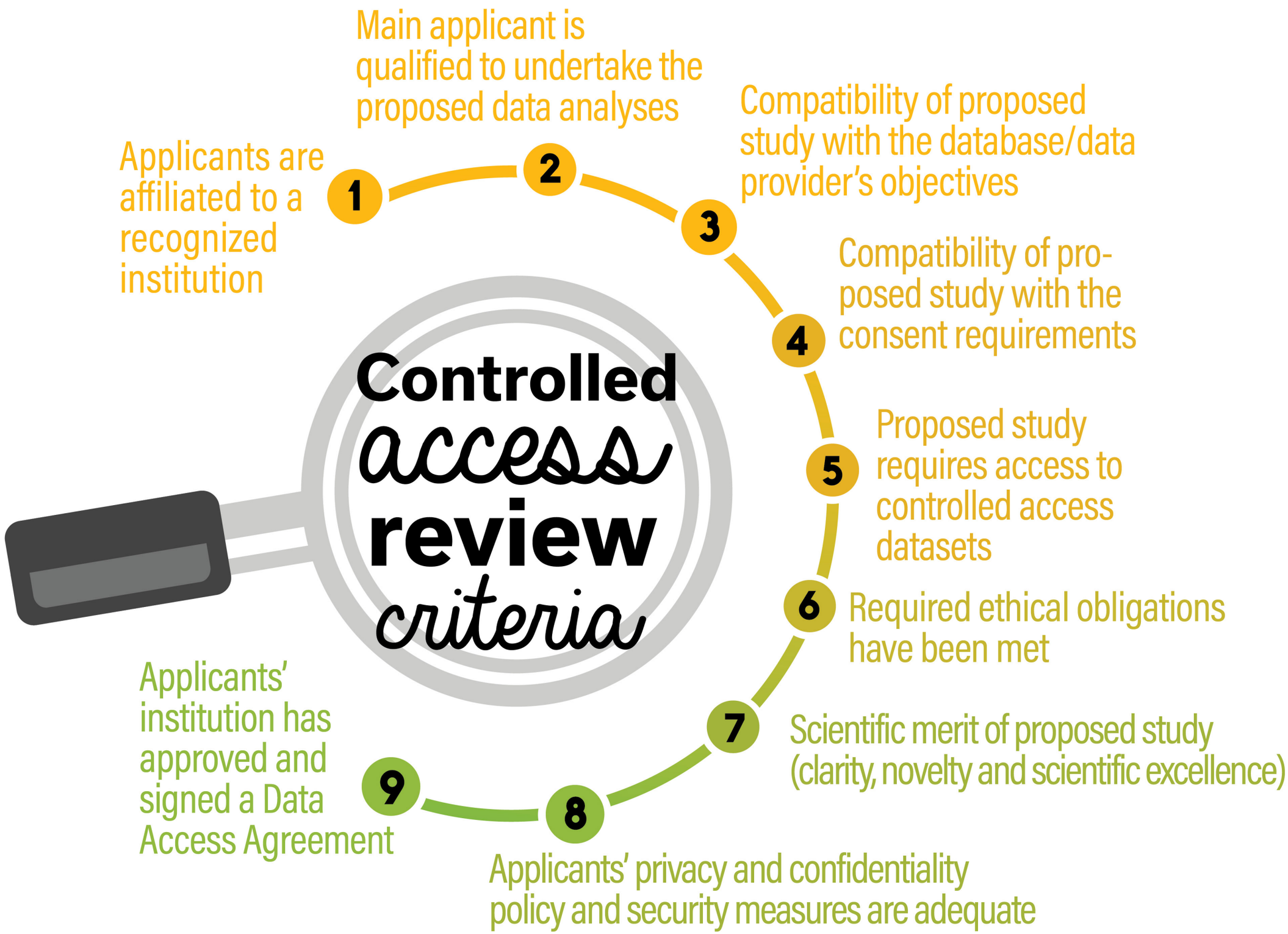
- 2.5/28 datasets have released their raw data publicly, the rest are “controlled access”

- Data sharing policies depend on the consent obtained from participants, need to be evaluated manually by the Data Access Committee.

- Some datasets can never be shared due to consent restrictions.

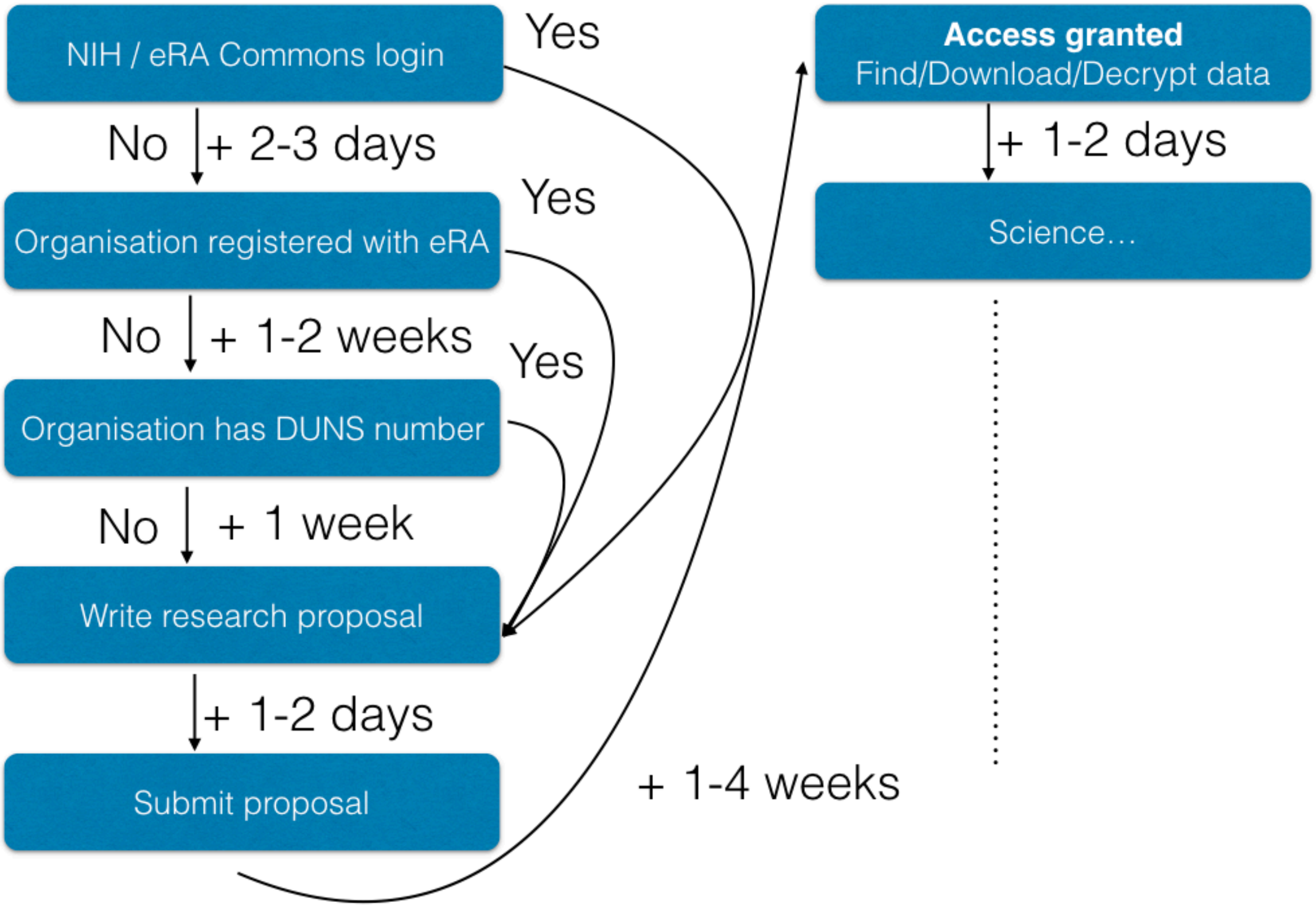
Genetic data repositories

- NCBI Database of Genotypes and Phenotypes (dbGaP)
    - Single application process
- European Genome-phenome Archive (EGA)
    - Each data owner has their own Data Access Committee and application process.
- NIMH Repository and Genomics Resources (NRGR)
    - Cover letter, biosketches, funding information, resources



SOM Dyke, *et al*. Registered Access: A "Triple-A" Approach. *EJHG,* 24 (12): 1676-80.

dbGaP Application Process



https://repositive.io/blog/post/how-to-successfully-apply-for-access-to-dbgap

https://repositive.io/blog/post/watching-paint-dry-in-the-21st-century-or-applying-for-data-from-dbgap

https://repositive.io/blog/post/accessing-dbgap-a-bureaucratic-oddesey-part-2

Rejected data access requests

	Name	Sample size	Cell type	Reference	Reason
2	Fadista_2014	89	pancreatic islets	Fadista, João, et al. "Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism." Proceedings of the National Academy of Sciences 111.38 (2014): 13924-13929.	Genotype data cannot be share (country restrictions)
3	STARNET	566	mammary artery (MAM), aortic root (AOR), subcutaneous fat (SF), visceral abdominal fat (VAF), skeletal muscle (SKLM), and liver (LIV)	Franzén, Oscar, et al. "Cardiometabolic risk loci share downstream cis-and trans-gene regulation across tissues and diseases." Science 353.6301 (2016): 827-830.	Summary statistics cannot be shared
4	MESA	AFA = 233, HIS = 352, CAU = 578	monocytes	Mogil, Lauren S., et al. "Genetic architecture of gene expression traits across diverse populations." bioRxiv (2018): 245761.	Summary statistics cannot be shared
5	Pashos_2017	91	hepatocyte like cells (HLCs)	Pashos, Evanthia E., et al. "Large, diverse population cohorts of hiPSCs and derived hepatocyte-like cells reveal functional genetic variation at blood lipid-associated loci." Cell Stem Cell20.4 (2017): 558-570.	Summary statistics cannot be shared
6	DeBoever_2017	215	iPSCs	DeBoever, Christopher, et al. "Large-scale profiling reveals the influence of genetic variation on gene expression in human induced pluripotent stem cells." Cell Stem Cell 20.4 (2017): 533-546.	Summary statistics cannot be shared
7	Ishigaki_2017	100	CD4, CD8, Mono, NK, B	Ishigaki, Kazuyoshi, et al. "Polygenic burdens on cell-specific pathway	Genotype data cannot be shared
8	Peng_2017	159	placenta	Peng, Shouneng, et al. "Expression quantitative trait loci (eQTLs) in human placentas suggest developmental origins of complex diseases." Human molecular genetics 26.17 (2017): 3432-3441.	"The genotype data is not publicly available at this time."

Rejected data access requests

	Name	Sample size	Cell type	Reference	Reason
2	Fadista_2014	89	pancreatic islets	Fadista, João, et al. "Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism." Proceedings of the National Academy of Sciences 111.38 (2014): 13924-13929.	Genotype data cannot be share (country restrictions)
3	STARNET	566	mammary artery (MAM), aortic root (AOR), subcutaneous fat (SF), visceral abdominal fat (VAF), skeletal muscle (SKLM), and liver (LIV)	Franzén, Oscar, et al. "Cardiometabolic risk loci share downstream cis-and trans-gene regulation across tissues and diseases." Science 353.6301 (2016): 827-830.	Summary statistics cannot be shared
4	MESA	AFA = 233, HIS = 352, CAU = 578	monocytes	Mogil, Lauren S., et al. "Genetic architecture of gene expression traits across diverse populations." bioRxiv (2018): 245761.	Summary statistics cannot be shared
5	Pashos_2017	91	hepatocyte like cells (HLCs)	Pashos, Evanthia E., et al. "Large, diverse population cohorts of hiPSCs and derived hepatocyte-like cells reveal functional genetic variation at blood lipid-associated loci." Cell Stem Cell20.4 (2017): 558-570.	Summary statistics cannot be shared
6	DeBoever_2017	215	iPSCs	DeBoever, Christopher, et al. "Large-scale profiling reveals the influence of genetic variation on gene expression in human induced pluripotent stem cells." Cell Stem Cell 20.4 (2017): 533-546.	Summary statistics cannot be shared
7	Ishigaki_2017	100	CD4, CD8, Mono, NK, B	Ishigaki, Kazuyoshi, et al. "Polygenic burdens on cell-specific pathway	Genotype data cannot be shared
8	Peng_2017	159	placenta	Peng, Shouneng, et al. "Expression quantitative trait loci (eQTLs) in human placentas suggest developmental origins of complex diseases." Human molecular genetics 26.17 (2017): 3432-3441.	"The genotype data is not publicly available at this time."

New NIH policy comes into effect in May 2019

Rejected data access requests

	Name	Sample size	Cell type	Reference	Reason
2	Fadista_2014	89	pancreatic islets	Fadista, João, et al. "Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism." Proceedings of the National Academy of Sciences 111.38 (2014): 13924-13929.	Genotype data cannot be share (country restrictions)
3	STARNET	566	mammary artery (MAM), aortic root (AOR), subcutaneous fat (SF), visceral abdominal fat (VAF), skeletal muscle (SKLM), and liver (LIV)	Franzén, Oscar, et al. "Cardiometabolic risk loci share downstream cis-and trans-gene regulation across tissues and diseases." Science 353.6301 (2016): 827-830.	Summary statistics cannot be shared
4	MESA	AFA = 233, HIS = 352, CAU = 578	monocytes	Mogil, Lauren S., et al. "Genetic architecture of gene expression traits across diverse populations." bioRxiv (2018): 245761.	Summary statistics cannot be shared
5	Pashos_2017	91	hepatocyte like cells (HLCs)	Pashos, Evanthia E., et al. "Large, diverse population cohorts of hiPSCs and derived hepatocyte-like cells reveal functional genetic variation at blood lipid-associated loci." Cell Stem Cell20.4 (2017): 558-570.	Summary statistics cannot be shared
6	DeBoever_2017	215	iPSCs	DeBoever, Christopher, et al. "Large-scale profiling reveals the influence of genetic variation on gene expression in human induced pluripotent stem cells." Cell Stem Cell 20.4 (2017): 533-546.	Summary statistics cannot be shared
7	Ishigaki_2017	100	CD4, CD8, Mono, NK, B	Ishigaki, Kazuyoshi, et al. "Polygenic burdens on cell-specific pathway	Genotype data cannot be shared
8	Peng_2017	159	placenta	Peng, Shouneng, et al. "Expression quantitative trait loci (eQTLs) in human placentas suggest developmental origins of complex diseases." Human molecular genetics 26.17 (2017): 3432-3441.	"The genotype data is not publicly available at this time."

New NIH policy comes into effect in May 2019

From the rejection letter: "*Please note that NIH is updating its data management procedures under the NIH Genomic Data Sharing Policy to allow unrestricted access to GSR from most studies in NIH-designated data repositories (on/about May 1, 2019): https://osp.od.nih.gov/scientific-sharing/genomic-data-sharing/*"

What is FAIR DATA?

Data and supplementary materials have sufficiently rich metadata and a unique and persistent identifier.

FINDABLE

Metadata and data are understandable to humans and machines. Data is deposited in a trusted repository.

ACCESSIBLE

Metadata use a formal, accessible, shared, and broadly applicable language for knowledge representation.

INTEROPERABLE

Data and collections have a clear usage licenses and provide accurate information on provenance.

REUSABLE

Good

Could be better

https://en.wikipedia.org/wiki/File:Implementing_FAIR_Data_Principles_-_The_Role_of_Libraries.pdf

What is FAIR DATA?

Data and supplementary materials have sufficiently rich metadata and a unique and persistent identifier.

FINDABLE

Metadata and data are understandable to humans and machines. Data is deposited in a trusted repository.

ACCESSIBLE

Metadata use a formal, accessible, shared, and broadly applicable language for knowledge representation.

INTEROPERABLE

Data and collections have a clear usage licenses and provide accurate information on provenance.

REUSABLE

Good

Could be better

https://en.wikipedia.org/wiki/File:Implementing_FAIR_Data_Principles_-_The_Role_of_Libraries.pdf

What is FAIR DATA?

Data and supplementary materials have sufficiently rich metadata and a unique and persistent identifier.

FINDABLE

Metadata and data are understandable to humans and machines. Data is deposited in a trusted repository.

ACCESSIBLE

Metadata use a formal, accessible, shared, and broadly applicable language for knowledge representation.

INTEROPERABLE

Data and collections have a clear usage licenses and provide accurate information on provenance.

REUSABLE

Good

Could be better

https://en.wikipedia.org/wiki/File:Implementing_FAIR_Data_Principles_-_The_Role_of_Libraries.pdf

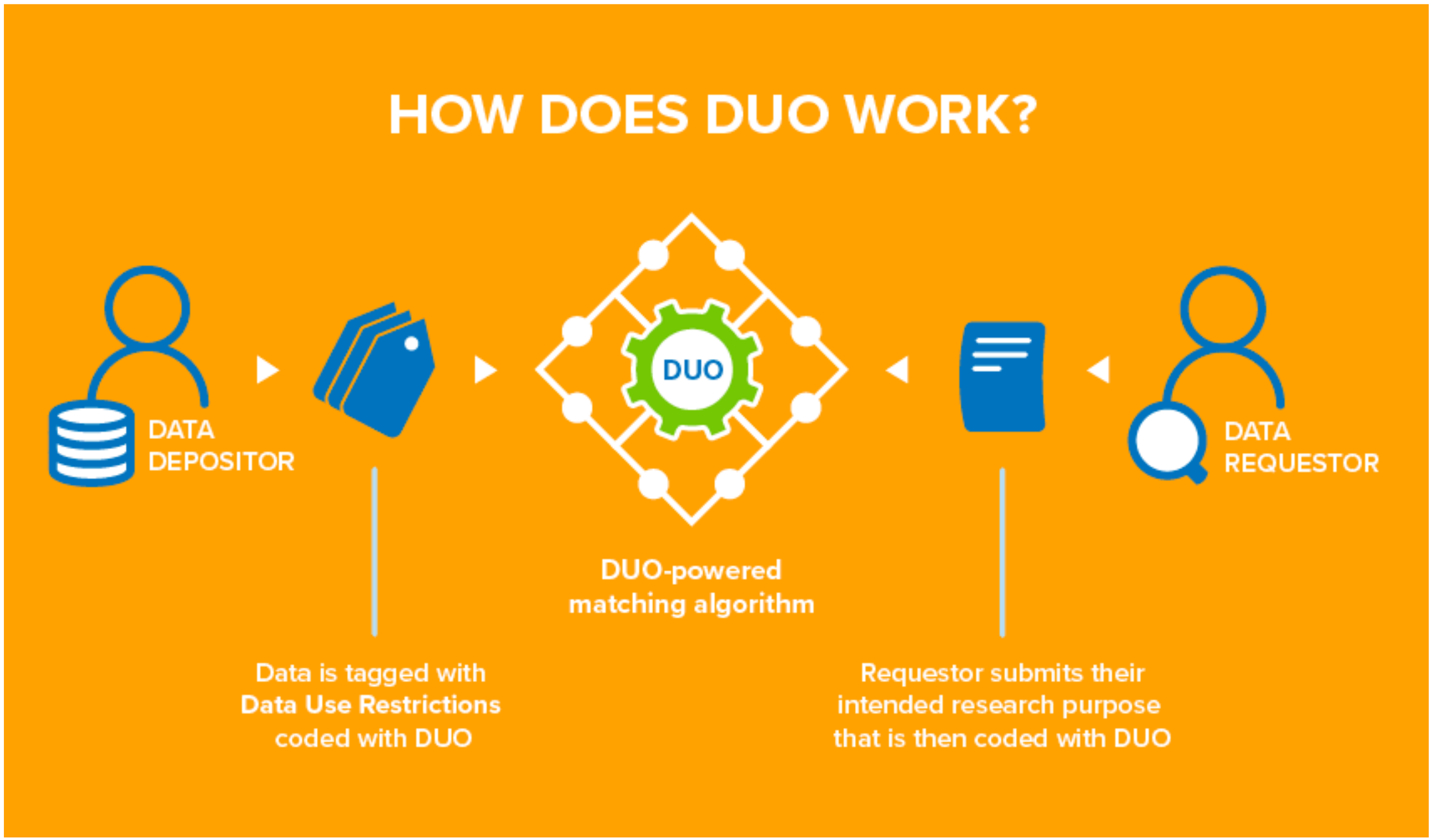
How can we improve?

Community metadata

- Metadata can always be improved upon.
- *Devil's advocate*: not all datasets are worth to be extensively annotated by the data owners. Better to share poorly annotated data than not share at all.

- Currently no good mechanism to link community-generated annotations and metadata to existing datasets.

Data Use Ontology



https://www.ga4gh.org/news/data-use-ontology-approved-as-a-ga4gh-technical-standard/

Registered access

Competence

Applicants are bona fide researchers/ clinical care professionals

The applicant's name, title, position, affiliation, institutional email, phone number, address, website and mailing address.

"I am a bona fide researcher, that is, I am involved in biological/health research and will use these data for those purposes only."

OR

"I am a bona fide clinical care professional working in genetic/ genomic medicine and will use these data for those purposes only.

Their home institution confirms

OR

They have a physician or other clinical care professional license

OR

A registered researcher corroborates their researcher status (as a reference)

Ethics

GA4GH Data Sharing Framework

"My use of the data will be consistent with the GA4GH Framework for Responsible Sharing of Genomic and Health-Related Data (https://genomicsandhealth.org)."

Meet IRB/REC requirements

"I will comply with all ethical and legal regulatory requirements applicable in my institution and country/region in my use of the data."

Security / Confidentiality

Do not re-identify data

"I agree to forego any attempt to identify individuals represented in the dataset, except by prior written permission from the provider's sponsoring institution."

Keep data confidential

"I will treat the data as confidential and I will not share it with others not specifically authorized."

Keep data secure

"I will protect confidential data against unauthorized access, and will delete all copies of the data when I no longer require the data or the permission period has expired."

Respect consent restrictions

"I will only use the data for the purposes allowed by the provider and I will abide by any consent conditions expressed as Consent Codes."

Acknowledgements

- Grant Office at University of Tartu

- Eleri Vako

- Taivo Raud