Phenotype Harmonization Guidelines

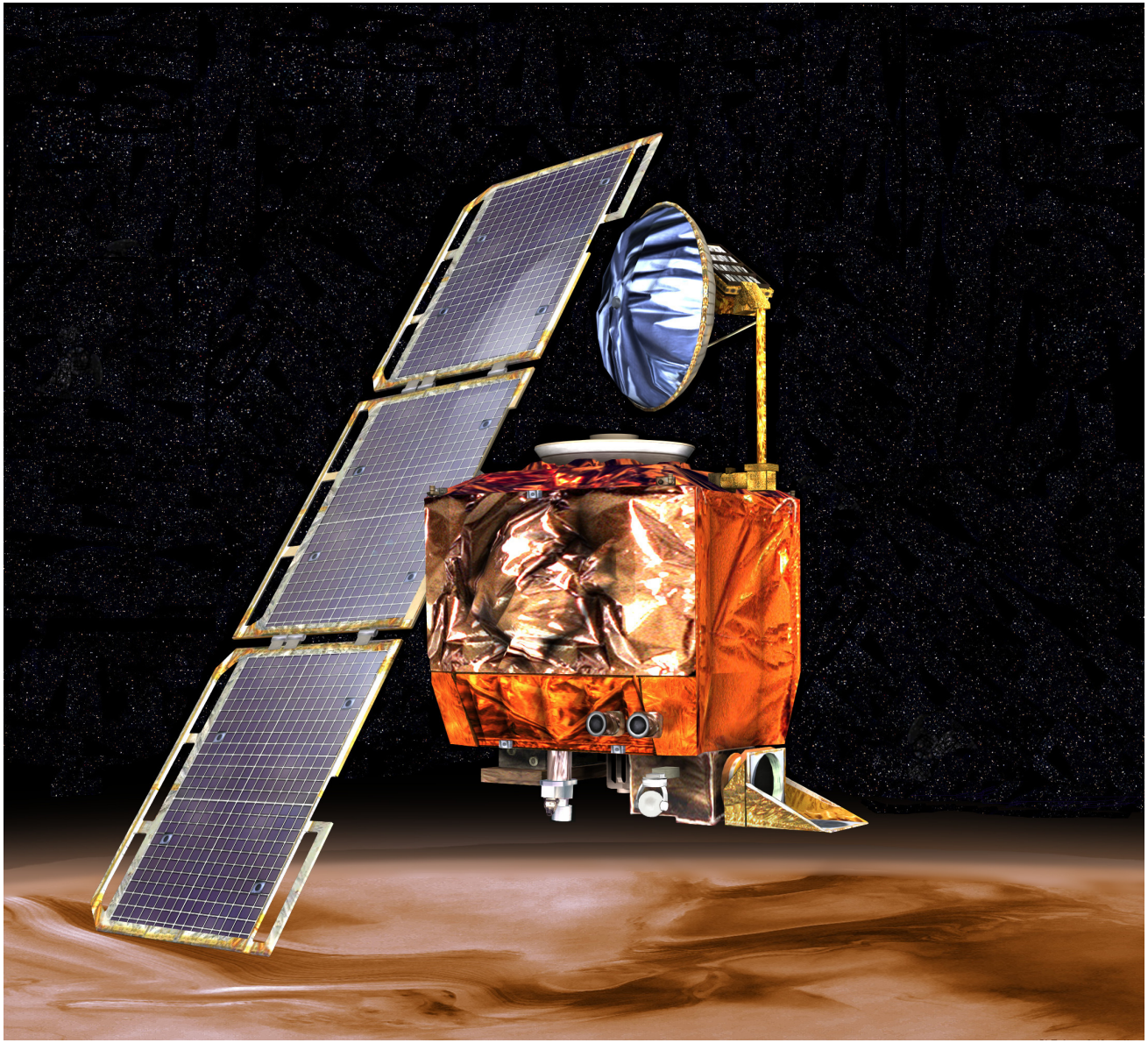
SISG 2018 Module 12

Adrienne Stilp

July 19, 2018

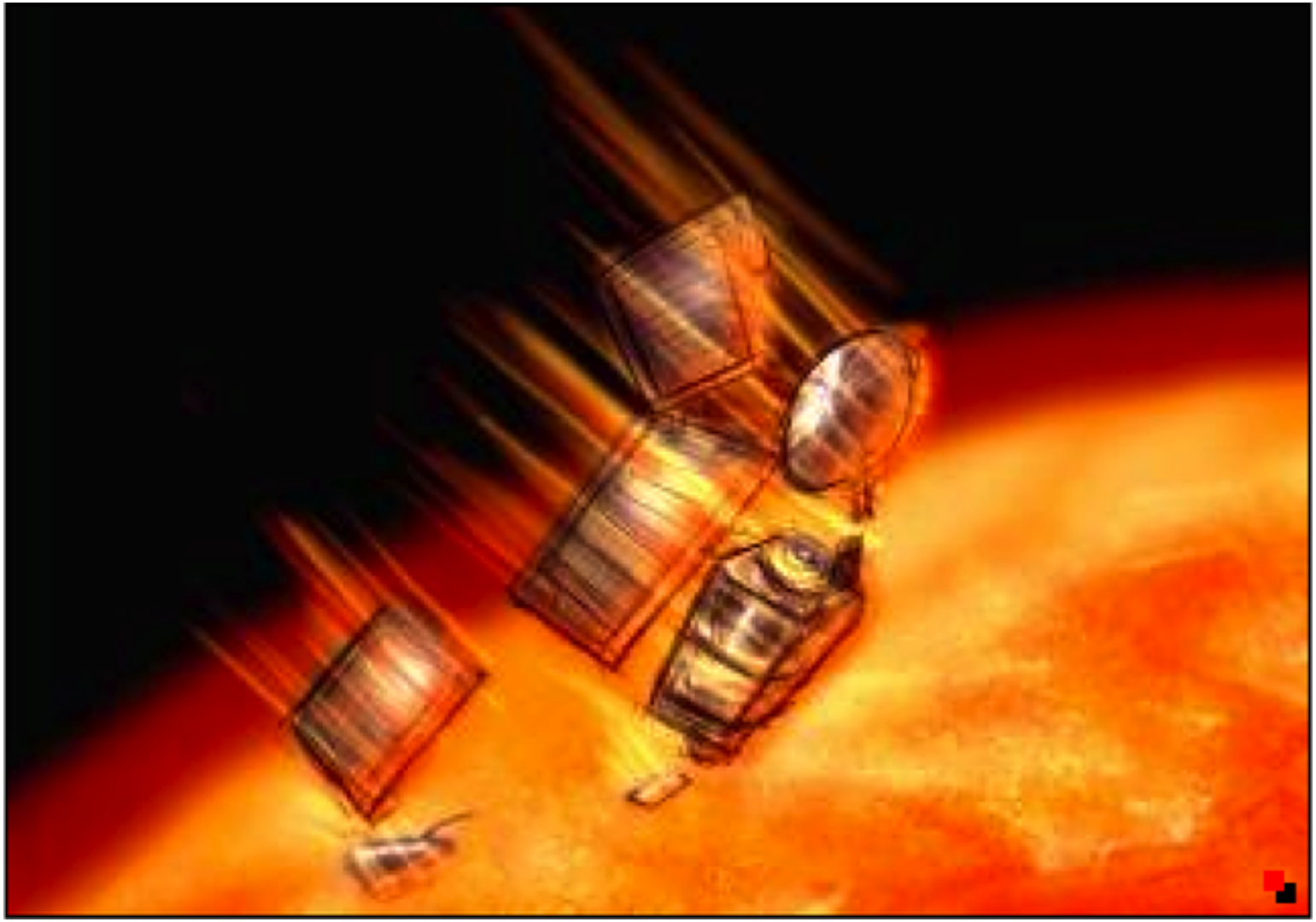
Phenotype harmonization is the process by which source phenotypes from different studies are transformed so that they can be analyzed together.

The Mars Climate Orbiter



NASA/JPL/Corby Waste

Disaster!



https://www.unleesh.com/single-post/2015/11/18/Three-times-not-being-on-the-same-page-ended-in-disaster

But really, why?

To find genetic associations, we need:

1. Genotypes
   - big but **homogeneous**
   - similar across studies -> automated processing
2. Phenotypes
   - small but **heterogeneous**
   - every study collects data differently -> manual effort required

- Too much noise can cause a loss of power and mask true associations.

What needs to be done in phenotype harmonization?

1. Define the target phenotype
2. Decide which studies can be included
3. Process source data by study
   - Perform QC
   - Determine harmonization algorithm
   - Once per study
4. Estimate quality of harmonized dataset
   - More QC
   - May need to repeat previous steps
5. Document and disseminate harmonized phenotypes

QC of study phenotypes

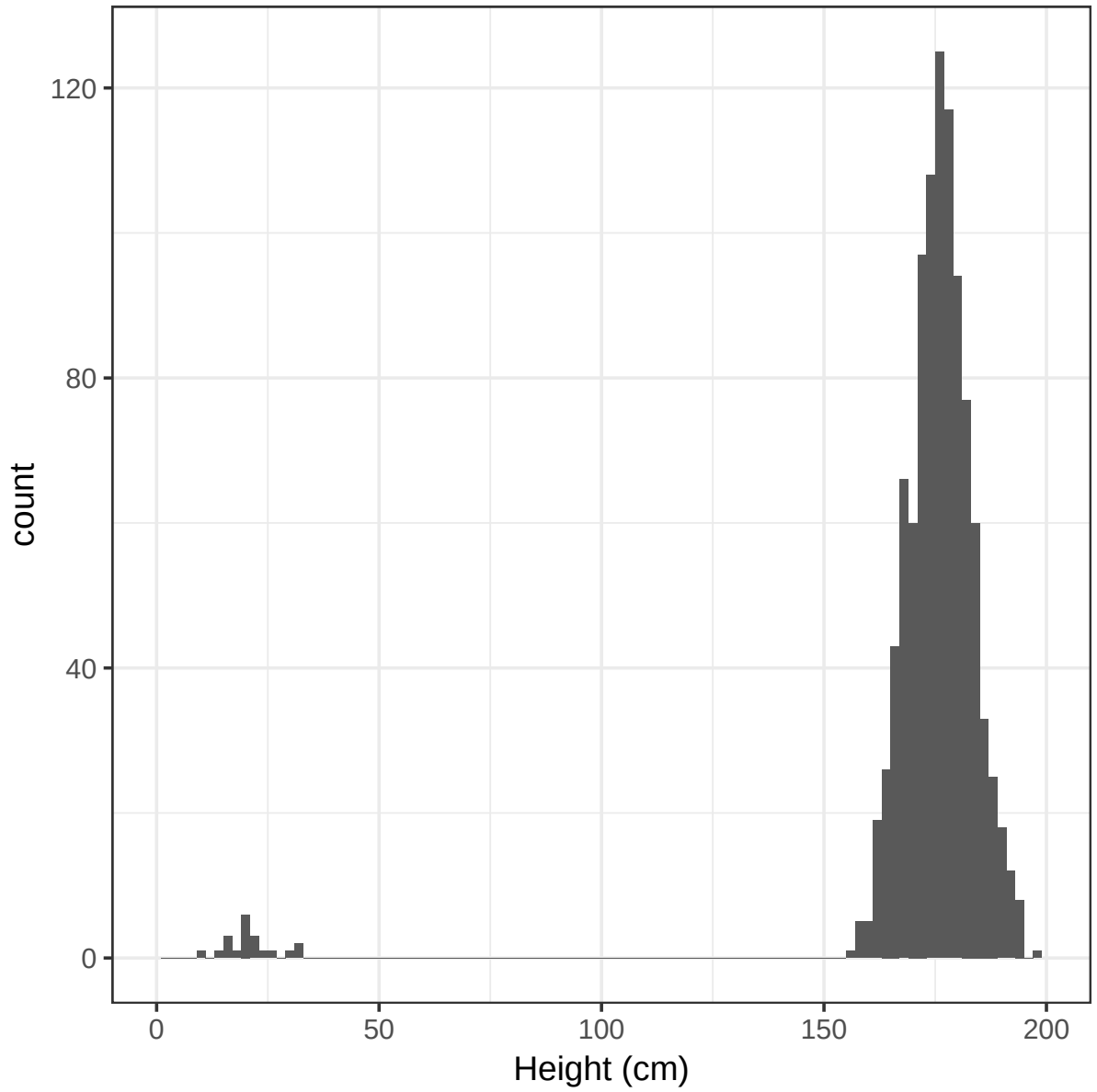
Potential QC issues:

- Biologically invalid values
- Extreme phenotypes
- Missing data
- Internal inconsistencies

And a lot of others you can't predict!

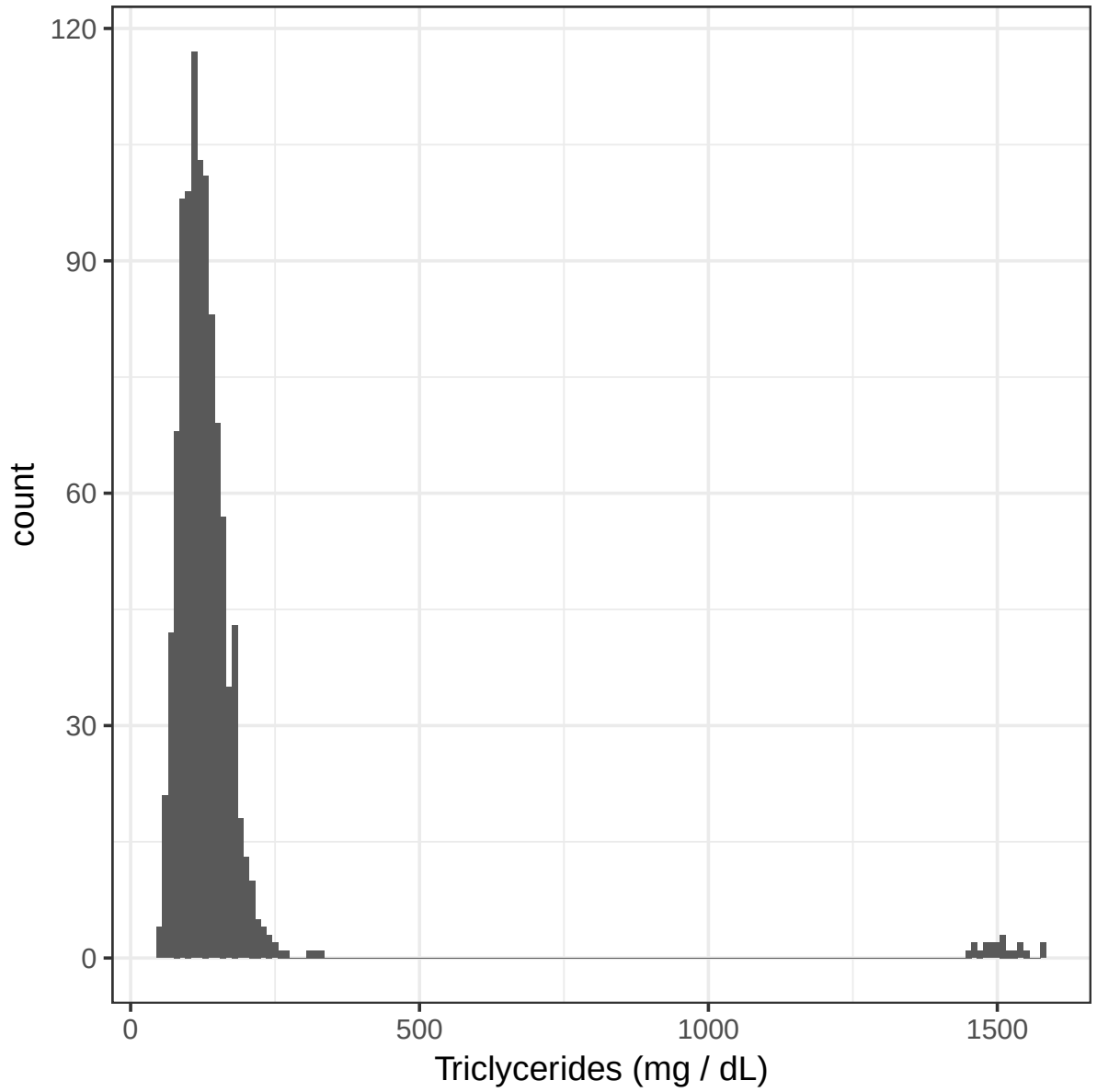
Biologically invalid values?

Example: implausibly small height measurements



True extreme phenotypes?

Example: Extreme triglycerides levels



Missing data?

Example: missing data in some components for diabetes

	subject_id	diabetes_self_report	diabetes_meds
1	a	1	1
2	b	0	.
3	c	1	1
4	d	1	.
5	e	0	0
6	f	1	1
7	g	0	0
8	h	1	1
9	i	1	1
10	j	1	1

Internal inconsistencies?

Example: self-reported vs. MD-diagnosed diabetes

	subject_id	self_report	md_diagnosis	
1	a	0	0	
2	b	0	0	
3	c	0	0	
4	d	1	1	
5	e	0	0	
6	f	1	0	# discrepant
7	g	0	0	
8	h	1	1	
9	i	0	0	
10	j	0	1	# discrepant

How do you fix problems?

- Which measurement (if any) is correct?
- Should you exclude subjects with discrepant data?
- Should outliers be excluded?
  - Measurement issue?
  - Real values indicative of rare variants with high effects (e.g., LOF)?

No blanket answer for all phenotypes!

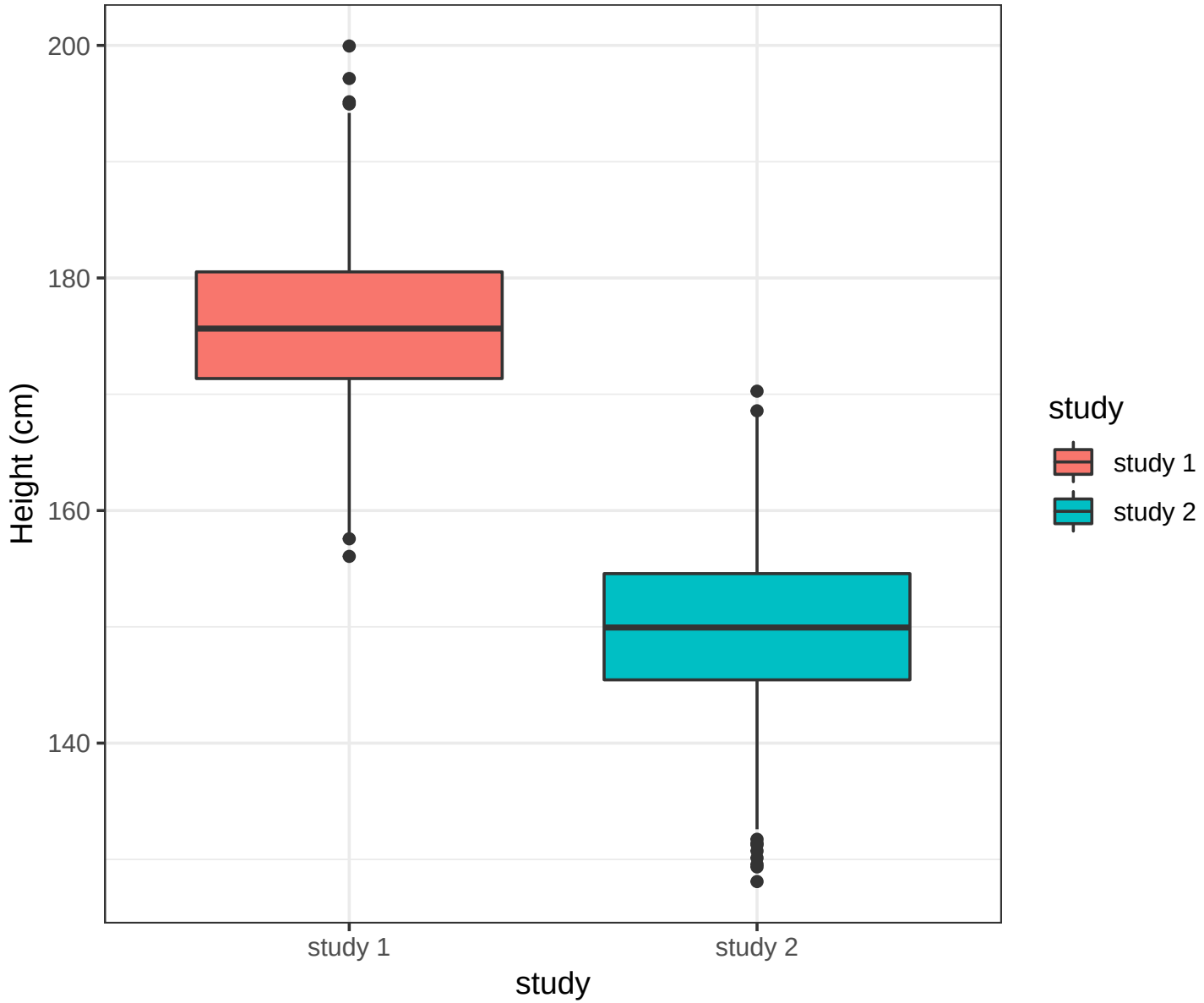
- Involve both study members and domain experts
- Clearly specify how to handle these QC issues

QC of harmonized data

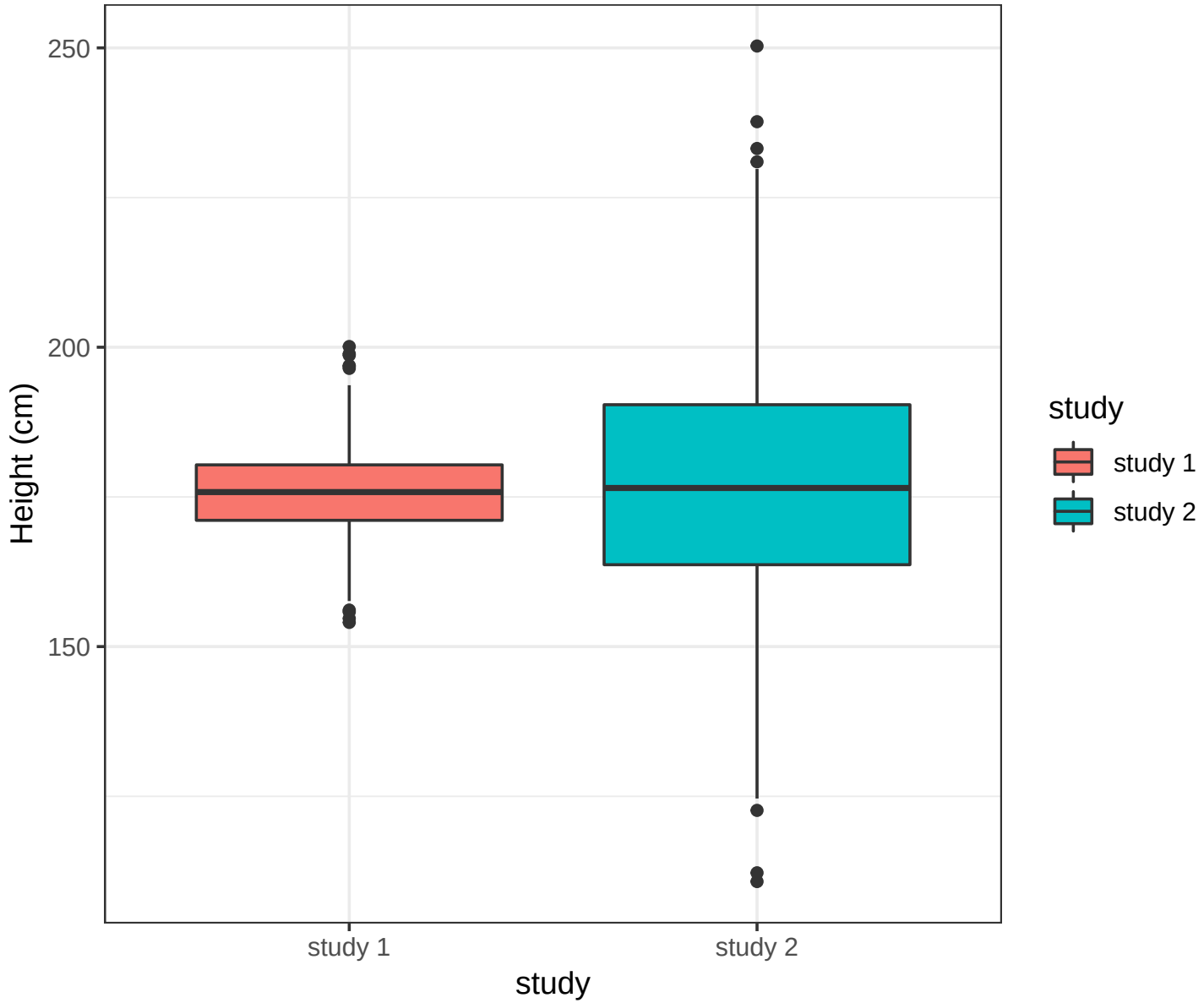
Are some studies very different than others?

- Quantitative data:
  - mean
  - standard deviation
  - general distribution
- Categorical
  - frequency
- May need to look at batch effects from other variables, e.g.:
  - Assay or device used?
  - Questionnaire version?
- For WGS with related subjects, fit a mixed model:
  - Fixed effects: age, sex, study
  - Random effects: genetic relatedness matrix

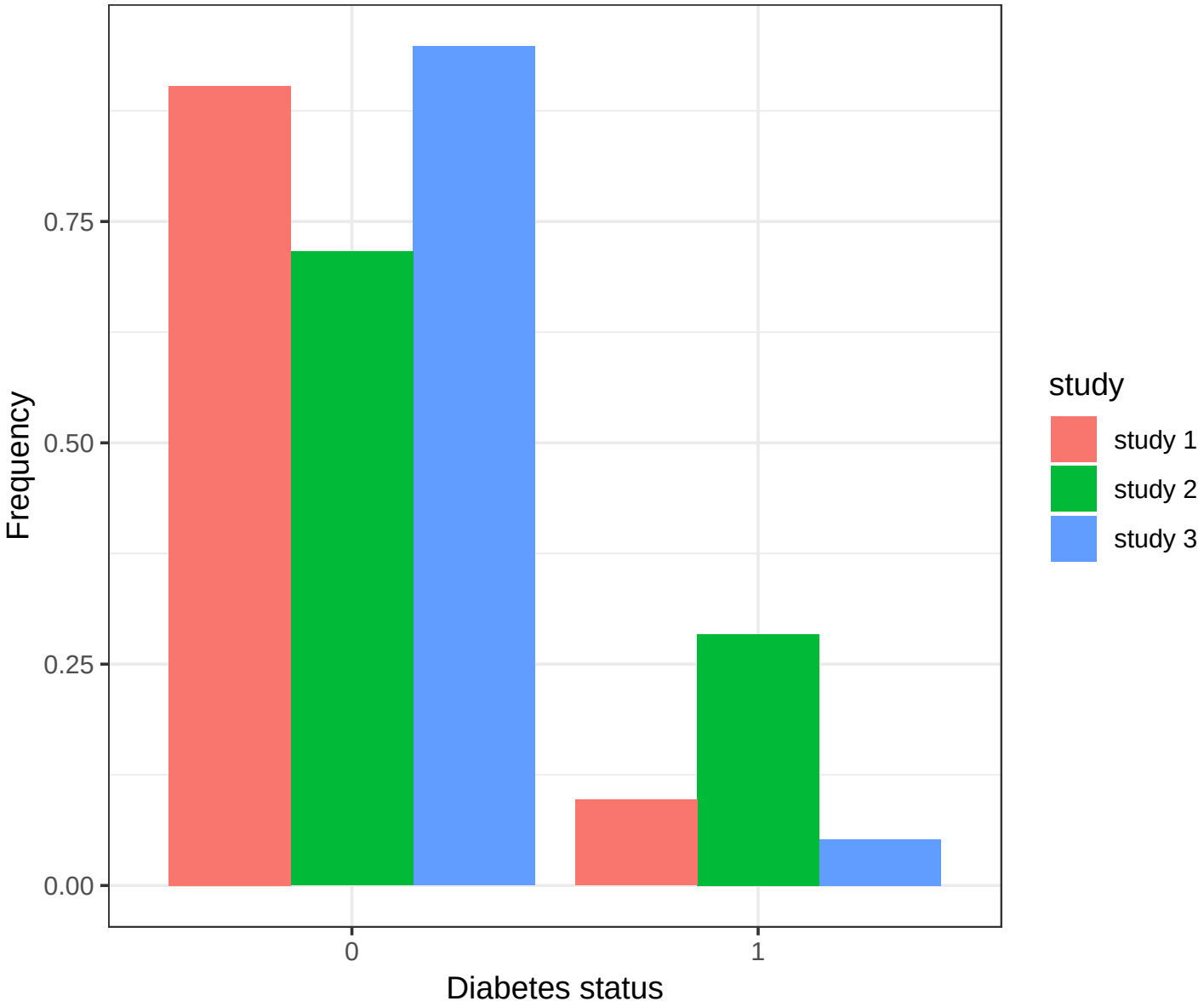
Different means?



Different standard deviations?



Different frequencies?



What do you do if you find a difference?

- Is there a valid reason for the difference?
  - Expected differences due to study design?
    - e.g., higher prevalance of disease in a study targeting cases
  - Different distributions due to ancestry?
- Is there an error in the harmonization algorithm?
- Do this study's data need to be treated differently?
- Is the study too different to be included?
- Do you need to adjust for the difference in analysis?

Again, no blanket answer to these questions!

- Need to involve both study members and domain experts

Documentation

Your phenotype should be reproducible.

- Accurate reporting in papers
- Able to add new studies in the future

What do you need?

- Definition of the harmonized phenotype
- Which component phenotypes were used
  - source file?
  - version?
- What algorithms were used
  - ideally, the exact code you used
- How QC issues were addressed

DCC harmonization for TOPMed

- Acquire study data from dbGaP
  - Provides a bookkeeping trail for documentation
  - Available to the general scientific community
- Store data in a relational database
  - Both study phenotypes and harmonized phenotypes
  - Includes everything needed to recreate a harmonized phenotype
    - Metadata
    - Component phenotypes and versions
    - Algorithms
  - Allows automated production of datasets and documentation

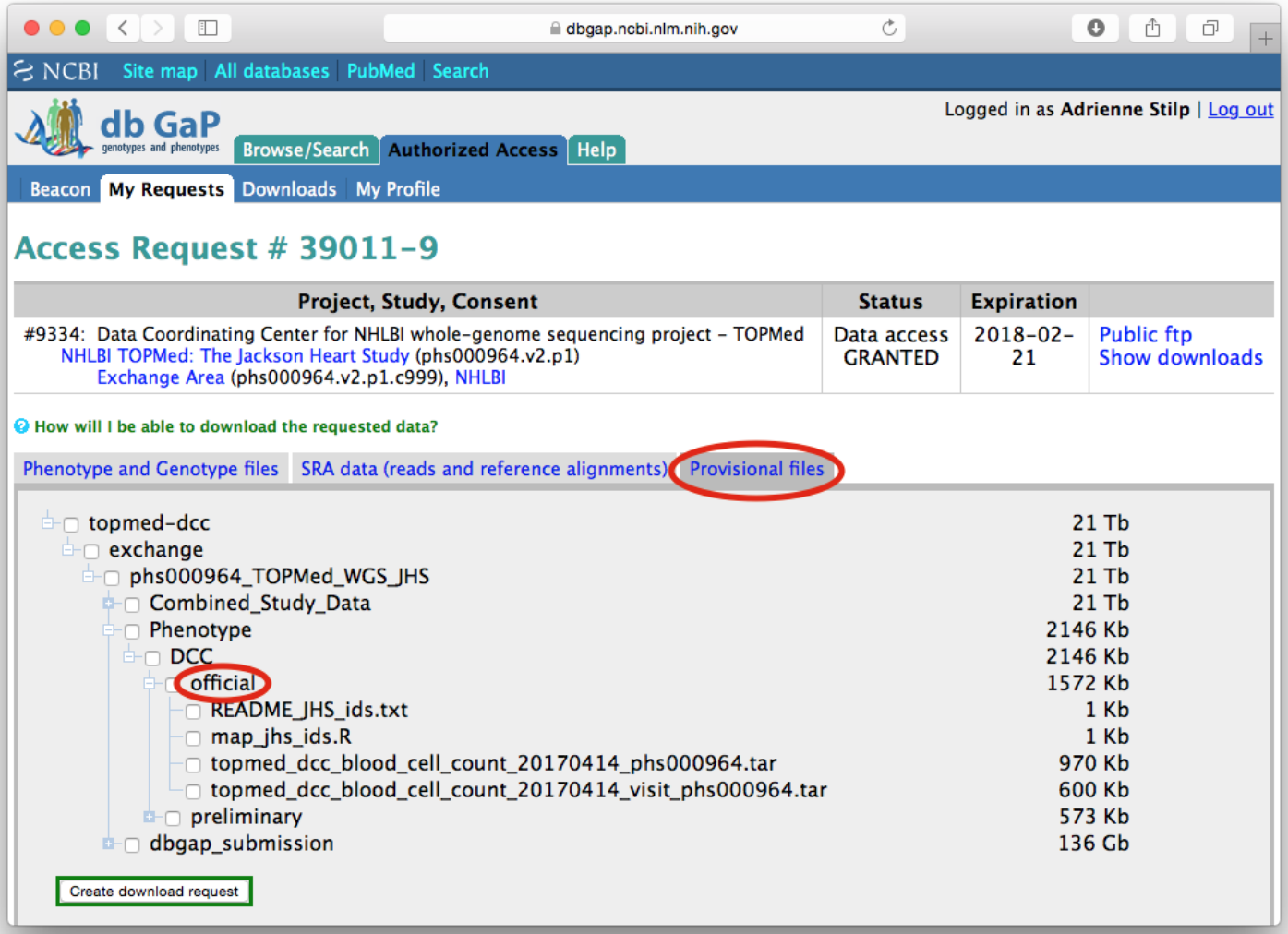
What phenotypes is the DCC harmonizing?

1. Key NHLBI phenotypes
   - Blood cell counts
   - VTE
   - Atherosclerosis-related phenotypes
   - Lipids
   - Blood pressure
   - ...
2. Common covariates
   - Height
   - Weight
   - BMI
   - Smoking status
   - Race/ethnicity

The DCC is in the process of preparing harmonized phenotype files for upload to dbGaP.

DCC-harmonized phenotypes in the exchange areas

If you are a TOPMed investigator, look in the exchange area for phenotypes harmonized by the DCC:



Guidelines for Phenotype Harmonization

- Always use subject ids in phenotype files
- Decide who will do the harmonization
  - You or the studies?
- Provide clear instructions to the harmonizers
  - Description of target phenotype
  - Clear algorithm definition
  - How to handle missing data and QC issues
- Perform sanity checks on the files you receive
- Document, document, document!

If you are interested in a specific phenotype area, join the appropriate TOPMed working group!

Helpful references

- Bennett, SN et al. Phenotype harmonization and cross-study collaboration in GWAS consortia: the GENEVA experience. Genet Epidemiol. 2011 Apr; 35(3): 159-73
- Doiron, D et al. Data harmonization and federated analysis of population-based studies: the BioSHaRE project. Emerg Themes Epidemiol. 2013 Nov 21; 10(1): 12
- Fortier, I et al. Maelstrom Research guidelines for rigorous retrospective data harmonization. Int J Epidemiol. 2017 Feb 1; 46(1): 103-106