

GCP ARM 2009
Brainstorming I—Mon 21st Sep (1330–1500)

GCP data curation, standards and quality

Convener: Elizabeth Arnaud

Minutes: Elizabeth van Strien

Panelists

- Theo Van Hintum: *Passport and SSR datasets in the GCP Central Registry: issues related to management of quality.*
- Claire Billot: *Validation of GCP SSR genotyping data: revisiting the process*
- Ruairadh Sackville-Hamilton: *Quality phenotyping data*
- Rosemary Shrestha: *Biocuration: Its Importance in Data Quality and Role in Standardizing Data*

Session's objectives

- Look at the quality checks performed this year or needed,
- To agree on necessary elements/processes to add or improve for quality data
- To get researchers' involvement in the process.
- To prepare the
 - The Ontology session will also be a part of the quality data in follow up to this session.
 - Data clinic session when the helpdesk team will show uploading and formatting of the data sets, including the phenotyping wizard.

Questions sent to panelists

1. What are the bottlenecks encountered by Biocurators with GCP data and what solutions can be applied?
2. How to build and sustain a curation community involving PIs for the GCP data sets and ensure peer reviewing?
3. What criteria can be applied to define the 'quality data' for breeders? for the GRSS?

All questions were not answered during the session as we encountered some technical problems.

A. Introduction:

'Data management is one of the essential areas of responsible conduct of research '

as outlined by the Office of Research Integrity <http://ori.dhhs.gov/>

Reference documents:

- OECD Principles and Guidelines for Access to Research Data from Public Funding
- Guidelines for Responsible Data Management in Scientific Research

Reference was made to the message sent by JM Ribaut

B. Data types:

SSR typical errors:

- Genotyping error – duplicated samples
- Incompatibility in frequencies
- Distortion in allele frequencies – allele calling problems?
- Differences in allele numbers per locus– misreading or contamination?
- Differences in allele numbers per accession – mixing, outcrossing?
- Differences between datasets: mislabelling or contamination?
- Does the structure represent all of the diversity?

Phenotyping data errors

- Ensuring high quality phenotyping data is about more than defining standards
- Need better elucidation of use case - User driven
- Data creators and managers
- **Take data quality more seriously**
- Software developers
- **Take more seriously the need to deny data managers the choice to get it wrong**

Typology of errors that can be applied to all data:

- Protocol error
- Results misreading, mislabelling, misinterpretation
- Incompatibility, Distortion
 - Wrong data
 - Right data wrong standard - this can be encapsulated to map proper standard (web services and ontology are possible solutions)
 - Right data right standard wrong place

Some errors cannot be avoided so computers are there to solve this-

C. What means quality data for GCP?

- Data should be fully interpretable
- Data to be attached to the reference sets of GCP: Passport and SSR, and..
- Presently, needs lots of curation action to be at this stage: decoding the accessions and markers, reference to literature is necessary
- Difficult to relate samples IDs to accession IDs
- Publish results / product / variety
 - Fit for purpose = data enable you to draw the conclusions relevant to your hypothesis

while

- Publish data
 - Fit for purpose = others able to use your data for other purposes
- who is responsible for quality ?
 - the PI generating the data?
 - GCP Central Registry?

D. Discussion summary and outcomes

Q FvE: is the PI going to use the information provided by the data quality check? A: Unfortunately, expectations on this are low.

Q: .not all errors are equally severe. A system of ranking would be useful. The data quality check result should be able to block downloading data files from the CR.

Feedback from users and PIs AFTER the data quality check is much required.

GCP data must strive to be globally meaningful. The CR needs a quality label, for data of which the GCP is happy to label as “GCP data”.

“Raw data” need to pass a **quality control process**, and can be subsequently released as meaningful “GCP data”. The PI is responsible for the raw data. This responsibility includes a small investment in effort and time for a data quality check.

Quality checks should be mandatory. Data sets and publications are required to meet set standards.

Q JCG raises the conversion of data lists and data matrices. Experience shows that the list is never ok or updated.

In this respect the organization of the process of the data quality check is really important. The communication with the PIs needs a tight organization. After uploading a data set, a **quick check** on data quality and subsequent **rapid feedback** are highly required.

In summary a protocol for the data quality check procedure must encompass the following points:

- the GCP needs to produce reports on data quality
- communication with PIs need to be in an organized fashion
- response of PIs on the DQ report must be guaranteed.
- a time schedule is required

Attention for data quality should not only focus on “old” data, but especially on the large amount of new data that are expected to be produced.

- Data are to be published so they ought to offer a certain level of quality but a debate went on the fact that quality of published data must enable other user to reproduce the experiment or not?
- OECD guidelines stipulates: **Data Storage** - This concerns the amount of data that should be stored -- enough so that project results can be reconstructed.

Note: Not much questions went on the standards to apply, particularly metadata and controlled vocabulary like Ontology so we hope to get more from the ontology session on Thursday.

E. Recommendations:

R1: Central registry structure and process

- organization Central Registry needs to change
 - meta information
 - version control
- Interaction between PIs and curator –iteration, feedback
- Uploading of corrected files by the curator
- Access to the raw data – versioning
- Simple mechanism to check data when uploading in CR



R2: A GCP Quality stamp

1. Data quality control applied at data sets level by PIs – curate row data
2. When project ends, let time for additional quality checking to assess if the data set can be labeled GCP sets – be attached to reference sets
3. Requires a GCP curation team and set of criteria to give the final approval

Still same questions:

Is there a long term vision of the GCP management for Central Registry, data quality tools and ontology?

Annex