

Vancouver Quality Assurance Workshop report

Contents

- [1 Towards a Data Quality Strategy for the Generation Challenge Program](#)
 - [1.1 GCP Data Quality Vision](#)
 - [1.2 GCP Data Quality Goals](#)
 - [1.3 GCP Data Quality Strategy](#)
 - [1.3.1 Quality Assurance Characteristics](#)
 - [1.3.2 General principles](#)
 - [1.3.3 Datatypes](#)
 - [1.4 Type specific QA and QC Strategies](#)
 - [1.4.1 Identification and Ownership](#)
 - [1.4.2 Passport data](#)
 - [1.4.3 Characterization and Evaluation Data](#)
 - [1.4.4 Genotyping data](#)
 - [1.4.5 Derived genetic data](#)
 - [1.5 Table of QA Characteristics for each Data Type](#)

Towards a Data Quality Strategy for the Generation Challenge Program

GCP Data Quality Vision

- To achieve recognition as a high quality public repository of research data.

GCP Data Quality Goals

- To document data resources to a sufficient standard to allow integration and support use for germplasm improvement and comparative biology.
- To promote quality assurance for all data sets and establish guidelines for quality control for GCP funded data resources.
- To train scientists in the application of data quality procedures

GCP Data Quality Strategy

Data quality consists of both data quality assurance (QA) and data quality control (QC). The GCP should make the application of QA/QC best practices mandatory for GCP funded research and should promote the application of such practices for existing data compiled into the project. These best practices should form a blueprint and specify policy supporting the generation and maintenance of high quality data.

Quality Assurance Characteristics

QA is complete documentation of the perceived quality of a given datum. This may consist of estimates of:

- Accuracy: whether or not the datum is close to its "real" value (statistically?)
- Precision: the level of uncertainty in the "real" value of the datum
- Consistency: logical consistency (temporal sequence, causality, etc.) between data elements
- Lineage: source of the information
- Completeness: completeness in availability
- Fitness for purpose: are the assumptions congruent with purpose

QC is the assessment of QA levels at defined input/output control points in a data flow from data source to usage point, with possible feedback remedial process efforts applied to the workflow to increase data quality.

General principles

- Ensure unambiguous and stable identification of and associations between all related material and data entities
- Driving forces for QC are:
 - Aggregation: accumulating and comparing data from many diverse sources
 - Publication: have all the data available to external parties
 - Visibility: ease with which a human party can interpret and assess the data (i.e. data visualization)
 - Attribution: identification of the user that d
- Pulling forces for QC are:
 - Provide tools & training for QA application to GCP data sets by end users:
 - To assist proper codification and (QA) documentation of data (ie. with structured formats and ontology)
 - To assist in QA activities by data generators on data, at source
 - To easily analyze data after capture and codification
 - Promote application of QA protocols and supporting technologies (i.e. controlled vocabularies and ontologies)
- Assess quality at the lowest level at which are opportunities exist:
 - Data collection (LIMS) - reporting mechanism that feedback critical outliers in quality (i.e. highlight deviations from expectations or error signals)
 - Publication to repositories (i.e. GCP data templates with data validation)
 - Acceptance protocols by repository ("gatekeeper" function)
 - Comparisons between data sources.

Datatypes

GCP data resources can be classified into the following data types which may be subject to specific quality measures in addition to the general principles.

- Identification and ownership of materials (i.e. germplasm) and data
- Passport data: secondary data (mostly, from existing genebank databases); need tools to assess quality of this data
- Characterization (environment independent phenotype) data: secondary data (mostly, from existing genebank databases); need tools to assess quality of this data
- Genotyping data: more local control (i.e. LIMS) with more opportunity for QC as well as QA
- Evaluation (environment dependent phenotype) data: field data that is environment specific (ie. G x E)
- Derived genetic data: e.g. genetic (QTL) maps
- Sequence and molecular expression data: gene expression, proteomics, metabolomics, etc.

Type specific QA and QC Strategies

Identification and Ownership

GCP datasets should document the following QA pertinent meta-data:

- Global identification/designation (like an ISBN for a book, etc.)
- Citation
- Description
- Time periods
- Status
- Scope of domain
- Keywords
- Access constraints
- Use constraints
- Point of contact
- Dataset credit

- Security information
- Native data set environment: raw data format (version), software version, etc.
- Cross references

Passport data

- Capacity building: Do not concentrate solely on passport data associated with GCP accessions, but work with other organizations to develop guidelines for improvement of passport data in general. SP5 should concentrate on training on how to implement best practices which are now being developed in several networks - SINGER, EURISCO etc. The GCP has an opportunity to take this role.
- Value: The GCP represents a unique opportunity to assess the value and quality of passport data for use in crop improvement. By looking at impact of the GCP we can assess the importance of passport data in these results and assess the missed opportunities resulting from lack or poor quality of passport data.
- Accuracy: What would a central repository look like? Will there be duplication of data from primary resources. If so these need to be synchronized between the repository and the primary data resource. This is crucial for the credibility of the GCP repository. The methodology and frequency of synchronization needs to be documented.
- Identification and Ownership of Germplasm: Cross-checking received germplasm passport identifiers (and underlying passport descriptions) against originating database and public passport databases: SINGER, GRIN, EURISCO, FAO, etc.
- Other Passport Descriptors: Verify consistency (matching) between passport descriptions across received and public records.

Characterization and Evaluation Data

Characterization is the measurement of traits assumed to be stable over environments while evaluation measures an environment dependent phenotype.

- Attribution: The importance of clear attribution in the quality of characterization and evaluation data

- Unambiguous identification of germplasm: Germplasm needs to be identified down to seed sample level to be compatible with and allow integration of characterization and evaluation data with genotype data
- Distributed data resources: Encourage local curation of characterization and evaluation data in a network of integrated databases through development and extension of CVO technologies for plant and phenotype concepts.
- Citation: Monitor use of data resources as a measure of quality and acceptability of GCP data.

Genotyping data

- Data submission: Though consultation with data producers finalise the data submission template taking into consideration the QA factors described in the table below.
- Timeline:
 - Circulate this document to all GCP members for review at the Annual Research Meeting in September/October 2005.
 - Final data submission template in October 2005
 - Final data submission deadlines
 - 2004 projects - December 2004
 - 2005 projects - June 2006
 - 2006+ project by March the following year.
- Data repository: Implement the data repository with the following functionality:
 - Single data repository, with back up and mirroring.
 - Data upload from template (By November 2005)
 - Validation tools
 - Check internal consistency of data (By November 2005)
 - Establish a list of validation rules (e.g. check external links, such as check germplasm ids against singer) as agreed in consultation with data providers at ARM (By October 2005)

- Implement agreed validation rules to check external links (By November 2005)
- Visualisation tools. For example, use of datamart (By November 2005)
- Temporary restricted access for period of grace (e.g. 6 months) to promote the early submission of data into the repository
- Summary statistics of data held in the repository
- Data querying tools using basic meta data such as
 - Provider
 - Crop
 - Data type (e.g marker, accession, location etc)
- Download of data in to various standard formats required for analysis and visualization software.
 - Simple output in CSV format (By November 2005)
- Full functionality (upload, download, querying) provided via both web interface and web services.
- Regular release schedule. At least twice a year: March and September.

Derived genetic data

For example genetic (QTL) maps), Genomic sequence and Molecular expression data

- Accuracy precision and completeness:
 - Encourage recording of GIDs for germplasm resources for all experiments and treatment conditions and experimental designs using well established 'template protocols' (like MIAME)
 - Enforce application of CVOs in data encoding of all GCP funded datasets. This can be facilitated by providing CVO curation and application tools with training as well as analytical reporting tools that survey the completeness of data documentation (application of the CVOs). (The default for labeling should be 'no labeling' so that deficiencies are easily discovered.)

- Commissioning and certification of LIMS installation at all GCP data collection sites.
- Archive original raw data and documentation for future external validation (eg. Sequence trace files from EST projects, including PHRED scores; gel images from genotyping).
- Ensure accurate cross linkages to related gene and sequence data. For public sequence information always record the sequence version associated with the specific set of annotation.
- For molecular experiments record quantification of biological samples. (Eg DNA dilutions etc)
- Use valid experimental design with appropriate biological and technical replication to ensure that results are meaningful.
- For DNA cloning based experiments (sequence analysis) document cloning vectors, bacterial hosts, restriction enzymes, treatment types etc. This is important for QA during sequence analysis.
- Capacity building:
 - Empower data generators and curators with tools, protocols and training for the application of data templates (eg MIAME), semantic encoding of data at source (CVO), data entry into LIMS, experimental design and statistical analysis.
 - Provide full context specific definitions and examples of all data components.
- Consistency:
 - Compare maps from different studies of the same mapping population for consistent marker positions. (In cases of new markers being added and sub-setting of mapping lines).
 - For consistency across sequence annotations compare outputs from different annotation algorithms and different published research results.
- Fitness for purpose:
 - Fully document experimental designs.
 - Fully document algorithms applied and analytical processes including program versions, algorithm details, assumptions, limitations (impact of missing or erroneous data points) and theoretical model options selected.

Table of QA Characteristics for each Data Type

	Accuracy	Precision	Consistency	Data Origin	Completeness	Fitness for purpose
Identification and Ownership of Data Sets	<p>Is this material/data item exactly what you say it is?</p> <p>Did you deliver what was expected (contractually)?</p> <p>Validate delivery/receipt of data sets</p>		<p>Compare identification of received data against original database and public sources</p>	<p>Validate data provider</p> <p>Location of original data files documented with formats, versions, etc</p>	<p>Ensure complete meta-data documentation generated/received</p>	
Passport data	<p>Report data accuracy checks of implemented in original dataset</p> <p>Verify input data against source</p> <p>Periodic verification and propagation</p> <p>Auditing accuracy with reference to other data</p>	<p>Misspellings corrected;</p> <p>Promote use of controlled vocabularies, standards, e.g. MCPD</p>	<p>Report data consistency checks of original dataset</p>		<p>Count/% of missing values per descriptor, and overall average</p>	<p>Compare the diversity of the GCP core collections against complete collections</p>

	sources					
Characterization (environment independent) and Evaluation (environment dependent) phenotype data	<p>Report data accuracy checks of implemented in original dataset</p> <p>Verify input data against source</p> <p>Periodic verification and propagation</p> <p>Auditing accuracy with reference to other data sources</p>	Promote use of controlled vocabularies, standards, e.g. PO & PATO ontologies	<p>Do all sources of characterization data from local and public sources agree with each other for a given accession.</p> <p>Does normalization of differentially scaled phenotype values resolve to the same value (including discrete labelings)</p>	<p>Are the protocols for characterization data accessible and compatible (consistent)</p> <p>Availability (location) and format of raw data is documented</p> <p>Documentation of location, environmental conditions (including temporal) and experimental treatments underlying the measurement of the phenotype</p>	Count/% of missing values per descriptor, and overall average	Compare the diversity of the GCP core collections against complete collections
Genotyping data	<p>Ensuring adoption of best practices, e.g. using replicate samples of control genotypes, replicate gels</p> <p>Use of control</p>	Publication of estimates of precision for control genotypes within and between labs.	<p>Consistency of coding and data formats across and within labs.</p> <p>Consistent, transparent, and documented approach to handling ambiguous</p>	Location of original gel images and associated data files documented with formats, versions, etc.	Clear reporting of missing data due to experiment failure, experiment not done and if possible null alleles.	Population structure analysis

	<p>markers and genotypes across labs.</p> <p>Analysis of the full sample set within each lab.</p> <p>Publication of detailed protocols including steps taken to assure data accuracy.</p> <p>Ensure that sample identification is maintained during the process and that these sample ids relate to defined germplasm.</p> <p>Ensure consistency of marker naming.</p>		data. E.g.			
Genetic maps, sequence and molecular expression data	<p>Report GID's of parents and progeny of mapping populations;</p> <p>Document name and version of</p>	<p>Promote use of controlled vocabularies, standards, e.g. GO and SO, ontologies (mandatory</p>	<p>Verification of sequence and marker identities with public databases</p>	<p>Are the protocols for characterization data accessible and compatible (consistent)</p> <p>Availability</p>	<p>Promote use of MIAME (MAGE-OM/ML) documentation standards (mandatory for GCP</p>	

	<p>mapping algorithms/software uses</p> <p>Correctly correlate sequence versions with annotation;</p>	<p>for GCP funded activities)</p>		<p>(location) and format of raw data is documented</p> <p>Documentation of experimental treatments underlying the production of (germplasm) sample reagents and derivatives</p>	<p>funded activities)</p>	
--	---	-----------------------------------	--	---	---------------------------	--
