

Genotyping Data Quality Workshop

Genotyping Data Quality Workshop, 19-23 February 2007, IRRI

Contents

Monday, February 19

AM 1 Welcome, Introduction, Overview, Use of the GCPWiki for workshop documentation

Session leader: T. Metz

[Initial Workshop agenda \(pdf\)](#)

Review of the agenda

Demonstration and discussion of the GCPWiki functionality for the workshop documentation

AM 2 Rice and Sorghum Datasets

[Rice dataset - Ken McNally \(ppt\)](#)

[Sorghum dataset - Claire Billot \(ppt\)](#)

Important points: missing data, controls and ladders (regression -bp size vs migration time- of the commercial size ladder is not always congruent with regression of the SSR marker), repeat of experiments

PM 1 Maize & Potato datasets

[Maize dataset - Céline Mir \(ppt\)](#)

[Potato dataset - Jorge Nuñez \(pdf\)](#)

PM 2 Lentil dataset

Session leader: A. Hamwiah

[Lentil dataset - A. Hamwiah \(pdf\)](#)

Tuesday, February 20

AM 1 Templates, CIMMYT & ICRISAT Dataset

The presentation on templates was given remotely by Guy Davenport via AccessGrid and VNC connection. Present at CIMMYT were also Marilyn Warburton, Susanne Dreisigacker, and Philippe Monneveux.

[GCP Templates - Guy Davenport \(ppt\)](#)

[Wheat dataset - Maria Zaharieva \(pdf\)](#)

[Chickpea, groundnut and pigeonpea datasets - Rajeev Varshney \(pdf\)](#)

AM 2 Development of common documentation scheme

[Documentation Presentation - Maria Zaharieva \(pdf\)](#)

The presentation of Maria Zaharieva reviewed existing documentation schemes that are used in the GCP, e.g. Multicrop Passport Descriptors, Descriptors for Genetic Markers, and their application and in the GCP SSR templates and the GCP Central repository.

The major conclusion of the discussion was that the experimental documentation (protocols. procedures) was generally sufficient, but that more emphasis needs, explanation, and institutional support needs to be given to the correct use of identifiers.

SampleID: Is a unique identifier within the experiment

GermplasmID: Leads to the (biological) material, and existing or future data associated with that material

Accession Number: Leads to passport data and other related (e.g. pedigree) data at the accession level

The use of GermplasmIDs is an issue that needs to be solved at the institutional level, as it spans across different domain, e.g. genetic resources, breeding, genomic research, etc.

The concepts of germplasm creation methods and neighbourhoods as implemented in ICIS were discussed. See: [1]

PM 1 Completing dataset documentation

Session leader: T. Metz

PM 2 Minimizing Error Occurrence

Chairman: K McNally, Notes: C. Billot

Minimizing error occurrence = Good laboratory practice

Discussion involved several different areas of concern:

Human resource management

Experimental design and implementation

Manipulation of data including tracking and scoring

Good computing practices

Wednesday, February 21

AM 1 Methods and tools for error detection

Error detection Presentation - Claire Billot (ppt)

The session deals mainly with controls/standards and preparing the data merging.

Controls

The example of sorghum.

After having put into evidence the discrepancy between allelic scoring based on standard size ladder and allelic sizing based on the stutter scales, we decided to use allelic references as controls. Three controls were defined. They are composed of 10 DNA samples mixed in 2 pools of 3 and 1 pool of 4 individuals. The 10 individuals were chosen from 48 Sorghum samples presenting a fair picture of the overall genetic diversity, in order to represent a large range of allelic diversity, both in terms of allele

number and allele sizes. Each control is amplified for each marker, and allelic sizes (controlled through sequencing of the alleles) are used as control sizes. All information concerning composition of the controls, allele sizes, and pictures can be found [here](#).

Gestion of SampleID, GermplasmID, AccessionID, raw data obtained from Licor, controls on the redundancy of the data according to SampleID and GermplasmID, merging of dataset according to allelic size standards, defining of a dataset and exporting it under different formats are performed through an access db.

Different species SSR allelic references are available in the frame of the GCP. Homogeneisation of presentation and information available for each kit will follow.

Common advices. As good laboratory practices, we would advise the user to use the ladder obtained by the amplification of the different alleles as a size ladder, to amplify the controls at the same time as the samples (positive controls), to put them at the same places whatever the DNA plate, to provide them in the dataset when merging data is concerned, and to document the use of these controls in the template submitted to GCP repository.

Merging of the data is possible only once the same allelic controls are present in the different dataset. See presentation on data merging.

Internal data quality estimation of the dataset can be estimated, provided that repeats of part of the experiment are present in the dataset released to GCP depository.

References

AM 2 Measurement of Genotyping Errors and QA

Session leader: Graham McLaren

Presentation on Genotyping Errors: Controlling and Measuring them

Discussion concentrated on the need and requirements for replicated observations. Three types of repeated measurements were discussed:

Standards - Samples with precisely known results used to calibrate the measurement process.

Controls - Samples with known results used to check that the process is working correctly.

Replicates - Repeated samples with unknown results used to measure the variability of the sampling and measurement process

Examples of these for the genotyping process are the commercial standards with precise molecular weight bands added to runs on the Licor genotyper to calibrate the separation. Samples of mixed DNA with known alleles are also added to each plate and these are controls for the DNA amplification step since they can be used to detect poor samples but they are also standards for the separation step since they are used to adjust the allele sizes for test samples. A sample with a known genotype was also added to each plate as a control. Finally some test samples were repeated and genotyped independently and these replications can be used to measure the variability in genotyping.

There are a number of within run measures of quality produced by various machines and processes. These should be made available in datasets where possible because they indicate the precision of the observation method. However they do not say much about the quality of the dataset because their scope is too narrow. Sources of error occur at all stages of the process from sample selection to allele calling and some overall measure of repeatability is required to document the quality of a dataset. Unfortunately this can only come from replicated samples. As mentioned in the presentation, partial replication is available by genotyping relatives, but no overall statistic

To measure the variability of genotyping within marker (locus) / we should calculate the average

pairwise Rogers' distance between all replicates. For replicate sample pair i and j the distance is:

[Barley SNP dataset & software - David Marshall \(pdf\)](#)

Friday, February 23

AM 1 Merging genotyping data

[Merging genotyping data - Céline Mir](#)

AM 2 Workshop documentation

Working and writing groups with (leadership) on:

Template recommendations (Maria Zacharieva)

Good laboratory practice (Ken McNally)

Detecting errors in genotyping data (Claire Billot)

Error assessment and quality measures (Graham McLaren)

PM 1 Workshop documentation

Working and writing groups with (leadership) on:

Template recommendations (Maria Zacharieva)

Good laboratory practice (Ken McNally)

Detecting errors in genotyping data (Claire Billot)

Error assessment and quality measures (Graham McLaren)

PM 2 Workshop documentation & closing

Workshop documentation review

Follow up???

Closing

Workshop Recommendations

The main conclusion of the workshop is that SSR genotyping leads to distribution of raw allele sizes (ie decimal numbers estimating allele sizes in relation to commercial ladders), whether accurate or not.

Errors found have different sources

Identity errors

Control errors: best practice for different situations

Measure errors: keep replicate data

Most of them can be minimized by Good Laboratory Practises and use of standards. However, an internal data quality index should be given to each data set. The use of Allelobin programme, as shown in case of ICRISAT, will be helpful. In order to assess for that, new templates have been defined that lead to changes in the central GCP depository db.

For existing data:

If you have loaded the data in the GCP repository, please refresh your data to take into consideration the new templates

If you haven't yet, use the new template

For future work, we advise the following points:

Germplasm resources should be documented as adequately as possible: if the plant corresponding to a DNA sample cannot be reliably identified, it is as if the data do not exist.

To maximize scoring accuracy, the use of sequenced allelic references should be a pre-requisite of any new genotyping project

The design of the SP1 genotyping project on data quality evaluation should take into considerations the conclusions of this workshop. Replication of some genotyping should be undertaken on each large dataset.

A LIMS is a key resource in minimizing errors for large scale genotyping exercise (and should be connected to seed collection management). Genotyping centres are encouraged to use the LIMS for future genotyping projects. Some participants felt that proper use of a LIMS should be a pre-requisite for participating in any future large-scale genotyping project.

GCP Data Submission Templates

Recommendations of the Genotyping Data Quality Workshop

working group: Maria Zaharieva, CIMMYT; Claire Billot, CIRAD; Aladdin Hamwiah, ICARDA; Graham McLaren, IRRI; Rajeev Varshney, ICRISAT

Documentation is one of the most critical issues in the study and use of genetic resources. It is essential for data to be recorded accurately and with the desired degree of precision. Guidelines for data submissions as well as examples in Excel sheet format are available on GCP site: <http://www.generationcp.org/bioinformatics.php?da=0650728>, in GCP Passport version 1.0 and SSR genotyping (fingerprinting) version 1.0.

However it has been noted by CRIL participants that two new improved versions had been created for genotyping (version 1.1.) and passport data (1.2) templates although not yet available in GCP site. Consequently the group decided to work on these new versions to avoid that some recommendations become obsolete soon.

Their prompt availability on GCP site with clear definitions and detailed examples taking into account the remarks/suggestions/recommendations of this workshop and users needs is an important requisite for a high quality documentation of the GCP SP1 datasets.

General remarks, suggestions and recommendations

Data must be easily documented in a suitable format, in order to be easily analyzed and used by

researchers with different backgrounds and experience.

1. In the templates guidelines, **clear definitions**, based on precise genetic and biological concepts, are needed.

- a good correspondence should exist between field description and field naming

- all definitions should appear in template cells of the excel files examples

- it should be better to use real data (e.g. data already posted in the registry) as examples, to explain and illustrate how to fill the different fields

2. **Data quality** should be recognized as a priority and this should be reflected in the GCP templates. An additional "Quality Assessment" section containing data for standards, controls, replicates, error estimators should be added. Clear guidelines for quality control assessment (examples, techniques, references...) should be established and the corresponding field naming defined.

3. **Passport data**: this information is essential and should be accurately collected and documented. Unfortunately, in many cases, passport data posted in the registry does not fit well with the required information, and consequently needs adjustments.

Having better defined passport descriptors could perhaps improve the quality of the collected passport data information. It is expected that in the future passport data would be automatically obtained from the genebank databases via database connection, but at the moment it is not the case and even when this link exists it will be not possible to have a direct access to passport data for accessions that are not originated from a genebank.

4. One **expert** has been designed by the GCP to revise the registered data. However, there is a need for a working group to take the recommendations in this workshop and coordinate or monitor activities relevant to them. Less clear is how such a group might be formed and maintained.

In order to improve datasets quality and their further utilization, assistants should be hired or mandated by the GCP to ensure the activities of revising, completing and adjusting the templates with the content of datasets and this, in strong collaboration with the persons who generated and provided the data (presently, these activities are assumed by investigators who have little time to devote to the revision of datasets presentation). Additional funding should consequently be found for these activities.

5. Increased **training** of scientists and technical assistants in the application of data documentation procedures, standardization of data definitions and collection methods, and education and assistance of data users would all greatly improve data quality and utilization.

Specific remarks, suggestions and recommendations

A. Definitions

Discussions on the meaning of "**Sample ID**", "**Germplasm ID**" and "**Accession number** (or identifier?)" showed that the definitions of these three identifiers were not easily understandable and need to be clarified. The definitions we can find in the GCP Templates Guidelines are as follows:

Sample ID is defined in **SSR Genotyping Guidelines** as "*a unique identifier of a DNA sample, which can be a sample in a well on a gel or a LIMS entry, or even a unique ID created specifically for this dataset. The Sample ID is specific to a laboratory but is not a universal identifier*"

Germplasm ID is defined in **Passport Data Guidelines** as "*a unique alphanumeric value which identifies the germplasm. This global identifier links data across domains. The format proposed is concatenation of "**HoldingInstitute:CollectionName:LocalUniqueID**" (e.g. NGA333:Genebank:252), where:*

- the "**Holding institute**" is the Institute holding the germplasm

- "**Collection name**" is "an unique identifier which identifies the collection to which the germplasm belongs within a holding institute" (e.g. CIMMYTWHEAT)

- "**Local unique ID**" is a number which "serves as the current unique identifier for the germplasm sample; if it concerns an accession within a genebank collection this refers to the accession number which is assigned when a sample is entered into the genebank collection."

It still needs to be clarified what should be the local ID if the accession does not belong to genebank collection and if only accession name is available. In some cases the "Local ID" has been confounded with "Sample ID", locally assigned for DNA sample.

Germplasm ID is also defined in **SSR Genotyping Guidelines** as "a unique identifier for the germplasm (e.g. seed sample) from which the DNA sample was extracted. Germplasm ID are often unique within a specific database. For this reason it should be prefixed by the data name or abbreviation. For example, an entry with Germplasm ID 2341 in IWIS, would be IWIS:2341. A new Germplasm ID is assigned each time an accession is regenerated or for some other reason a new seed or germplasm sample is taken".

It was noted by the participants that some institutions managing genetic resources collections have already adopted the Germplasm ID (GID)* as unique germplasm identifier automatically generated by ICIS for their material, thus the Germplasm ID is already known (as for example for IRRI rice collection) and this number should be used for GCP dataset presentation.

(*GID definition: "a unique identifier for a germplasm record within an ICIS database. GIDs are integer values that are automatically generated by the system when new germplasm records are created. Their primary use is for rigorous internal identification of germplasm within the database, not for external publication.").

This should be reflected clearly in the templates guidelines in order to avoid confusions. Moreover the definitions and examples in Passport and Genotyping data for Germplasm ID should be the same.

Accession is defined in **SSR Genotyping Guidelines** as "The number or name of the accession" and "as a collection of samples with different Germplasm IDs."

There is no field in Passport data named Accession or Accession number or Accession name. Only in the definition of "Local ID" is noted that "if it concerns an accession within a genebank collection this refers to the accession number which is assigned when a sample is entered into the genebank collection".

In a genebank, accessions are the key genetic entities and are identified through their accession number. This number is used in publications. Users who want to access to the germplasm material of interest from a genebank will refer to this number, and it is important to be added in passport data template (see comments for Passport Data).

Comments on the three categories of identifiers:

As defined, there could be more than one Germplasm ID per accession, and more than one sample ID per Germplasm ID. Each Germplasm ID is linked to one accession, but one accession can be linked to more than one Germplasm ID. If one accession (identify by its local genebank accession number) has gone through different generations of management it is documented under different Germplasm ID (in IRRI example), and receive automatically a GID. But it is not the case for all genebanks, for institutions collections, or breeder's collections. In these cases, the creation of Germplasm ID for the GCP data registration should follow the rules described in Passport data guidelines.

Each sample ID is linked to one Germplasm ID, but one Germplasm ID can be linked to more than one sample ID. If there are multiple extractions from the same germplasm material (same Germplasm ID) each DNA sample would have a unique Sample ID. This definition does not link the sample ID to the really existing seed material (a subset or one seed from a packet of seeds from what DNA was extracted). In some cases, as in ICRISAT, the progeny of these materials was obtained and stored in

a separate packet with the same accession number as the original packet from which the seed subset was obtained. But in the majority of studies, DNA was extracted, used in the study, but no more available seeds exist. So the Sample ID is just a virtual number, related to the local experience, but not linked to the physically existing seed material. It was postulated by a number of participants that the information obtained at the sample ID level is important for our knowledge about heterogeneity or heterozygosity of a Germplasm ID (but cannot be easily separate, taking into account the biological status of the germplasm, e.g. breeder's line, landrace, wild...). In SP1 Data analysis Workshop held in CIHEAM, Zaragoza, Spain June 21-25, 2004, the issue of heterogeneous accessions was considered and it was suggested that "*when this is possible (not too late) and efficient (suitable multiplication rate), it is advisable to extract DNA from a single plant per accession and to self it, and use the seeds as a foundation*". But for a number of laboratories this was not possible and was not done.

Sample ID does not only correspond to different extractions from a Germplasm ID seed packet. In the genotyping templates guidelines it is also defined as "*DNA sample, which can be a sample in a well on a gel or a LIMS entry, or even a unique ID created specifically for this dataset*". This means that we can have different sample ID even from the same DNA extraction but used in different experiences (different conditions, different gels...), and there is no field corresponding to the description of the sample ID status (different extractions from the same GID, or the place in different gels etc.). The multiple genotyping data information obtained at Germplasm ID and Accession (number) levels could be confusing for data users. In the majority of datasets in GCP registry (e.g., CIMMYT wheat dataset), a Sample ID is related to only one Germplasm ID and to only one Accession number. The DNA was extracted only once from one accession (number) and used for the whole composite set genotyping.

Conclusions

Definitions and semantics should be improved taking into account:

the reflections, remarks and recommendations of the participants of this workshop, principally data collectors

users (GCP partners, collaborators, national programs...) understanding and needs

B. Genotyping data: SSR genotyping (fingerprinting), version 1.1

Sections (spreadsheets): Source, Experiment, Conditions, Data list, Data Matrix, Markers, Maps, Accessions, Institutes

1. Source: Institute, Principle investigator, Email contact, Species, Ploidy, Dataset name, Version, Creation Date, Remark), all mandatory. *There were no remarks for this section.*

2. Experiment: Operational Taxonomic Unit, Purpose of the study, Missing Data (Mandatory) Remark (optional). *There were no remarks for this section.*

3. Conditions: Sampling strategy, Control genotypes, Size Standard, DNA extraction, DNA amplification and detection (Mandatory), Genotyping Software, Quality Measure, Reference (optional)

It was suggested to delete "Quality Measure" from Conditions section and was recommended to create a new section entitled "Quality Assessment" and containing description of the quality assessment, standards, controls, error estimators. Clear guidelines for quality assessment should be established.

4. Data list and 5. Data Matrix (different formats for genotyping data presentation)

In the Guidelines of the new version (1.2) as well in the old one it is postulated that "*the data can either be in a list format or a matrix. The list format is preferred since it can be produced directly from both ABI and LI-COR genotyping software and contains more information.*"

A number of participants have submitted their datasets in "matrix" format. They found this format easier for submission, more appropriated and adapted for further analyses. However, it was also noted that in previous meetings it was decided that "Data will be present only as a list data" in order to

relate with the amount for each allele. This decision, if adopted, should be argued and reflected in the guidelines for data submission.

List Format (Data List)

Fields: Sample ID, Accession, Marker (Mandatory), Gel/Run, Dye, Allele, Size, Quality, Height, Amount (optional).

It was recommended to add a Mandatory column "Germplasm ID" as a link to passport data and to define "Accession" field as accession number only, or to rename it in "Accession number" in order to avoid confusions with the different writing of the accession name.

Matrix Format (Data Matrix)

Fields: Sample ID, Accession, and columns (labeled with the marker name) containing the alleles for each marker, with the number columns equal to the ploidy of the species (...2 columns per marker for diploids, 3 columns per marker for triploids etc. For bulked data there would be a variable number of alleles per marker in each sample". All columns are Mandatory.

As for Data List, it is recommended to add a Mandatory column "Germplasm ID" and to rename "Accession" in "Accession number".

6. Markers: Marker (Mandatory), Chromosome, Map, Position_cM, Motif, Forward_Primer, Reverse_Primer, Annealing_Tm, Min Allele, Max Allele, Accession GenBank, References

It was recommended that Motif, Forward_Primer, Reverse_Primer, Annealing_Tm, Min Allele, Max Allele, Accession GenBank, References should be Mandatory.

7. Accessions: Sample ID, Germplasm ID, Holding institute, Collection name, Local unique ID, Genus, Species, Country of origin, Accession name (all fields Mandatory).

There are changes in the fields in the new version 1.1 in comparison to the version 1.0, "Holding institute", "Collection name" and "Local unique ID" fields being added and Germplasm ID is now mandatory. However "Accession Number" was excluded in the new version but should be kept as Mandatory. This whole section is marked as optional in the two genotyping versions. It contains some selected passport data linked with Sample ID, however the fields naming differs from passport data (there is no field for Accession number and Accession name in passport data).

Some participants proposed to delete this section "Accessions" (if "Germplasm ID" is added in Data list and Data matrix sections as link to passport data). An alternative could be that this section contains only "Sample ID" and "Germplasm ID" as link to the passport data. However it was also noted that it would be practical for users to have in the same file some selected passport data corresponding to the sample ID (linked to genotyping data).

8. Institutes: FAO institute code, Name organization, Street, City/State, Zip Code, Country,

Institutional email, Institutional telephone, Fax, Url, Primary contact name.

There was no remark for this section.

C. Passport data: Passport, version 1.2

Sections (spreadsheets): Source, General Passport Data, Collecting location data, Additional Passport Data, Institutes.

1. Source: same fields as in Genotyping data; replace the Meta data provider of version 1.1

2. General Passport Data: Germplasm ID, Holding institute, Collection name, Local unique ID, Is genebank accession (Mandatory), Acquisition date (optional), Genus, Species (Mandatory), Rank, Intraspecific epithet, Cultivar, Cultivar group (optional), Full scientific name (Mandatory), Scientific name author, Taxonomic reference, Crop name, Sample status, Sample status remarks, Collecting

location ID, Collecting date, Collecting institute, Collecting number, Collecting source, Collecting source remarks, Ancestral information, Pedigree release year, pedigree developer, Donor institute, Donor ID, Alternative genebank ID, Special plant characteristics, prevailing stresses, Ethnic group, Local uses, Other names or ID, Germplasm sample Url, Date last modified (Mandatory), Remarks (optional).

3. **Collecting location data:** contains data related to the collecting site; only one Mandatory field "Collecting location ID"

4. **Additional Passport Data:** any additional descriptors defined by the data provider, two Mandatory fields: "Germplasm ID" and "Collecting location ID"

5. **Institutes:** same fields as in Genotyping data.

Comments:

Ontology: avoid use terms that differ to those used internationally (MCPD, EURISCO), e.g., "infraspecific epithet" instead of "subtaxa" or "sample status" instead of "biological status". For more clarity, the nomenclature should be improved in closer association with genetic resources managers and users.

"Full scientific name" generally includes the name of the author (e.g., *Triticum turgidum* L. subsp. *carthlicum* (Nevski) A. Love and D. Love, or *Triticum aestivum* L. subsp. *aestivum*), so that there is no need to have two different fields "Full scientific name" and "Scientific name author"

"Cultivar", "Rank" and "Infraspecific epithet" (=subtaxa), " should be mandatory

"Accession number" and "Accession name" should be added in General Passport data section. At least "Accession number" should be Mandatory. As defined in EURISCO descriptors "*this number serves as a unique identifier for accessions within a genebank collection, and is assigned when a sample is entered into the genebank collection*"

Minimizing error occurrence - Good Laboratory Practice

Chairman: K. McNally

Notes: C. Billot

Discussion involved several different areas of concern:

Human resource management

Experimental design and implementation

Manipulation of data including tracking and scoring

Good computing practices

Human resource management

Hire or choose the right person for a project who has appropriate and adequate technical and communication skills to accomplish the project. In other words, the management ensures the right person is at the right place.

A common working language is needed. Care should be taken when the working language is not the first language of the staff involved (technicians, students, visitors or trainees).

Provide training when needed.

Prior to project start, orient staff regarding procedures, expectations, and documentation.

Staff should adequately understand all protocols so that they are aware of how and why errors can

occur as a result of failed equipment or manipulation of the samples or the chemistries.

Ensure staff can interpret the results.

Ensure staff understand the effect(s) error would have on the quality of the data set and its ultimate use.

Hold regular meetings of staff involved in a project.

Experimental design and implementation

Standardized protocols should be used. These must be optimized, documented and reproducible.

Equipment should be calibrated so that measurements are accurate and precise.

Experiments must be documented with records such that results are verifiable and errors can be traced.

DNA extraction, from one seed or not depends on the species.

These standard operating procedures are embodied in ISO Certification. In some cases (SCRI), obtaining funding is dependant on certification. The certification process involves good guidelines and should not be frightened by it.

Specific recommendations:

Sample tracking must be in place, preferably with barcodes. If not, then labels should be printed using a computer-generated list, preferably from a database (LIMS). Where not possible, a standardized labeling scheme and unambiguous script should be used.

Better to pre-label tubes, bags before harvesting. If possible, plant in same arrangement as for plates.

Systematic checking of labels by dual reading to ensure proper placement and accuracy of labels should be followed.

When handling DNA extractions, quality assessment, quantification, PCR amplification, etc, samples should be handled in a systematic fashion such as a specified order that is never changed for a set of 96, i.e. a "plate." The plate then becomes the trackable object.

All experiments must include appropriate positive and negative controls and standards to verify results. With positive controls, failed procedures can be identified. Standards allow the identity of sample order to be followed from a plate onto other devices such as a sequencer. The fate of a particular sample is then traceable. Negative controls allow identification of assays where contamination may have occurred.

When handling PCR plates with sticky plastic sealers, cross-contamination can be avoided by removing the plastic from a frozen plate.

Plates should be centrifuged before and after PCR or use for sample transfer and loading.

Sources and lots of chemicals should be from the same supplier over the course of a project. * Records of suppliers, lots, expiration dates, etc. must be maintained. Yet, for many purposes different suppliers can be used as long as controls are in place that verify the accuracy and reliability of results.

Laboratory automation (e.g. full robotics for extraction, PCR, etc) is recommended especially when dealing with thousands of samples. This would allow integration of barcodes and sample tracking directly with a LIMS, partially removing the "human" factor.

At a minimum, use of multichannel pipettes (8 or 12 channels) must be used when handling samples.

Dispensing of sample volumes should be well within the accuracy of the instrument.

A pipettor with a maximum volume of 20 ul, has an accurate range from 2-20 ul. Therefore, the pipettor must not be used to dispense volumes below 2 ul, and better accuracy would occur if the volume was 5-20 ul.

Larger volumes are more accurately dispensed than small volumes.

96-pin manual replicators (~3000 USD) can be used for DNA transfer. These allow one drop of DNA to be transferred from a stock into a plate, and can be calibrated such that the amount transferred is adequate for a PCR amplification.

Approaches to reduce experimental costs, such as reusing tips with liquid handling machines for the same DNA plates or polypropylene PCR plates (CIMMYT, with negative and positive controls after washing with bleach), are not recommended practices since untrackable errors could be incurred.

Attempting to reduce costs might result in more repetition and ultimately higher costs. Cheap quick solutions are not always the best.

Experiments should follow prescribed order of addition of components with regard to their chemistry.

Certain "failed" results that seem perplexing could be the result of equipment failures such as gradients occurring during PCR or capillaries failing during genotyping. Operators should gain a sense that if samples from columns 11-12 always give poor results, then the PCR machine might be at fault. Or, if the runs on capillary two are always noisier, then the capillary might need replacing prior to the rest.

Proper experimental design should be implemented, such as randomization of sample order, inclusion of replicates, and blind controls. These will allow statistical assessment of result quality to be determined.

Length of gel runs or type of capillary matrix should give the appropriate separation needed for a locus.

Manipulation of data including tracking and scoring

Dual reads by independent scorers. Reconcile calls in disagreement by consensus.

Score all runs for a primer at one time, so that the same approach is used to call alleles and the entire range of a locus is taken into account.

Certain SSR loci may give poor results. The effort to attempt call these loci may introduce more error than worthwhile. Don't hesitate to eliminate bad loci!

For the quantification of the frequency of an allele, sometimes peak height is used instead of peak area. Tests to ensure that peak height is correlated with peak area should be done on a per locus basis. Peak height is more advantageous when dealing with peaks that overlap. Use of peak area in these cases would require deconvolution.

In instances where samples were mixed prior to loading, the allele range for all loci should be checked. Since sequencers use filters that have overlapping wavelength ranges, some peaks might be the result of dye "bleaching." A band appears that is not from the locus-dye being scored but from one of the other locus-dye combinations in the loaded sample. Scoring needs to account for this. In capillary systems, bleed through can occur between capillaries, especially when the amount of sample loaded is high relative to the others.

Good computing practices

Spreadsheets should be used with caution, use a database instead.

Staff involved in computing should be knowledgeable about the software they are using.

Error detection

Standarts enable the scoring of the data (peaks or bands). Two types of standarts can be used:

- commercial standarts, for which quality is assessed by the provider
- allelic references (ie mixture of alleles of known sizes, obtained through allelic sequences or estimated and considered to be stable in time).