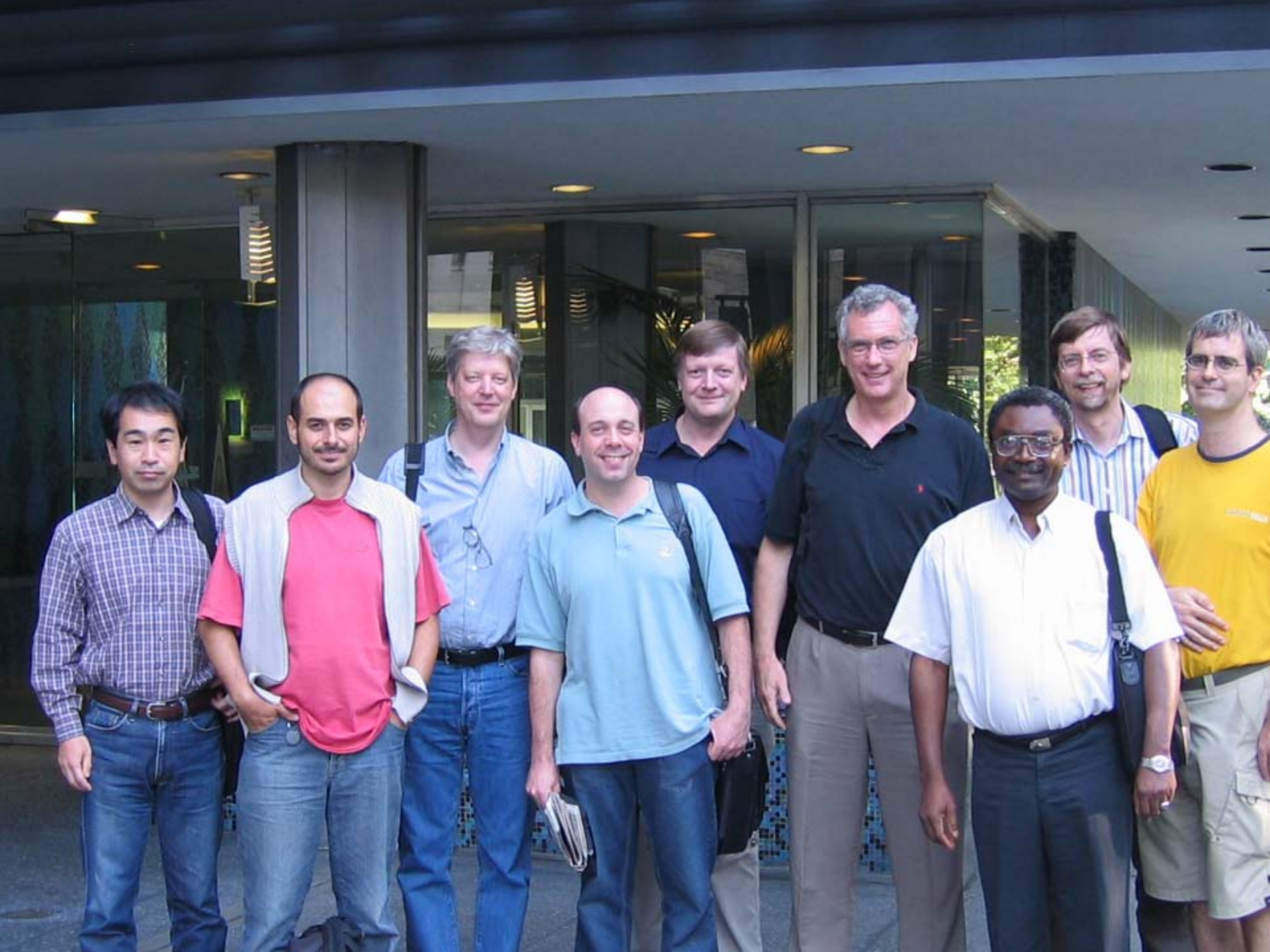




Improvement of quality of existing GCP databases

Next Steps



Data Quality

The Conventional Wisdom:



- Data quality control and quality assurance are very important
- You should just see how bad my colleagues' data are!
- Something should really be done about it!

A GCP Data Quality Strategy



- Five GCP data curation centers conducted base-line surveys
- Results of these surveys were discussed at the Vancouver meeting with the following main conclusions:
 - Data quality should be everyone's concern
 - The problem is with your own data
 - The problem starts as soon as you collect it



GCP Data Quality Vision

Achieve recognition as a repository of public research data of certified quality

- **Quality** means ‘fit for the intended purpose’
- **Certified** means that the quality control is documented and assured

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



- Mandatory application of QA/QC best practices for GCP funded research
- Promote the application of such practices to existing data compiled into the project
- Develop a blueprint and policy supporting the generation and maintenance of high quality data

Components of Quality Assurance



- **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 Depends on Data Type



- **Identification and ownership** of materials and data
- **Passport data:** secondary data (mostly, from existing genebank databases)
- **Characterization** (environment and phenotype): secondary data (from existing genebank databases)
- **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
Derived genetic data: e.g. genetic (QTL) maps
- **Sequence and molecular expression data:** gene expression, proteomics, metabolomics, etc.



Next Steps

- Get a clear understanding that QC and QA are everyone's problem
- Develop data-type specific QC protocols
- Train data collectors and curators in applying the QC protocols
- Implement QA components in GCP data repositories