
Microsoft® Research
Faculty Summit
10
YEAR ANNIVERSARY

Microsoft Research

Improving Meta-Analysis based GWAS Through Data Quality Management

Raul Ruggia
Professor
InCo, Univ. de la República
Uruguay

Hugo Naya
Head of BioInformatics
Institut Pasteur Montevideo
Uruguay

Outline

Microsoft Research

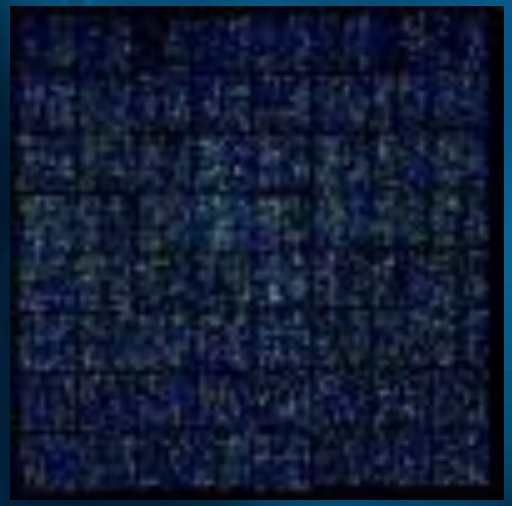
- GWAS and Meta Analysis
- The approach
- GWAS data quality model
- GWAS data quality management tools
- Conclusions
- Perspectives

GWAS and Meta-Analysis

Microsoft Research

- Genome Wide Association Studies
  - Objective: associate genotypes to phenotypes
  - Technology allow to genotype thousands of markers (SNPs)
  - Screening (genotyped) of case (ill) and control populations
    - Genotype markers associated with the disease
    - Typical sample sizes are 1000 cases and controls
- Meta Analysis
- Statistical approaches that address a set of related research hypotheses combining the results of several existing studies

Courtesy Affymetrix



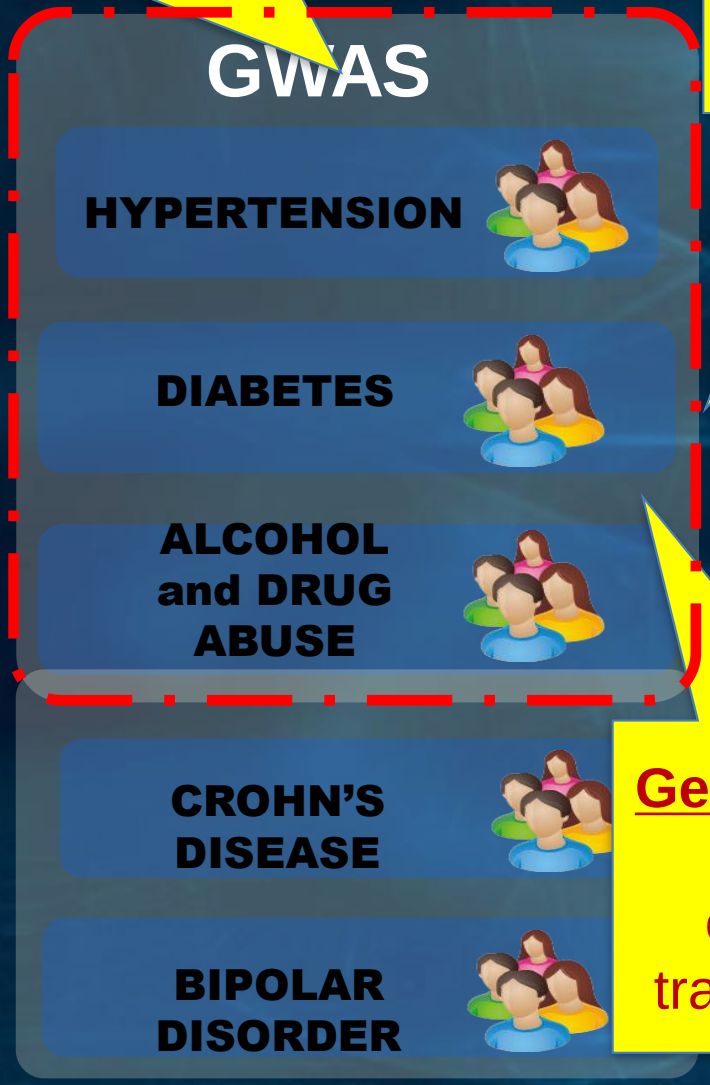
Combining GWAS Studies for Meta-analysis

Microsoft Research

subset: GWAS with CAD data recorded

Challenge: data from different GWAS possess different specifications and quality levels

Goals:
-Enable to combine different studies
-Apply Data Quality Management Tools
-Provide a data quality assessment environment to "GWAS Meta-Analysts"



Approach:
META-ANALYSIS

NEW GWAS STUDY

ex: Coronary Artery Disease (CAD)

General Problem: Costs and complexity of traditional GWAS

Research problems:

- Specify main GWAS data quality properties.
- Evaluate GWAS satisfaction & select the best combination of GWASs.
- Support interactive analysis of multiple properties on multiple GWAS.
- Compute data quality of GWAS combination.

Meta-analysis: Advantages and Challenges

Microsoft Research

- Advantages:
  - Increases sampling size and statistical power
  - Imputation of markers between different platforms (HapMap):
    - ➔ increase in markers density
  - Reduce costs of new studies
  - NEW!! Information could be reused to addresses different targets or diseases (depending on phenotype info available)

- Challenges:
  - Source heterogeneity
    - differences in terms referring to similar concepts
    - quality heterogeneity (genotype and phenotype)
  - Addressing similarities and differences between studies ➔ canonical study
  - Population structure (internal and combined)

GWAS Meta-analyst Perspective

Microsoft Research

- Given a new target (e.g., a disease to study):
  - How to find the best subset of studies to combine?
    - Balance between maximizing the size of combined data against quality of the result (from data heterogeneity)
  - Which are the main quality properties to look for in GWASs?
  - Which GWAS combines favorably with others?
  - How to address the quality of the resulting dataset?
- This requires:
  - Specify a Data Quality Framework for GWAS
  - Manage multidimensional quality information in a highly interactive manner

Approach of the Project

Microsoft Research

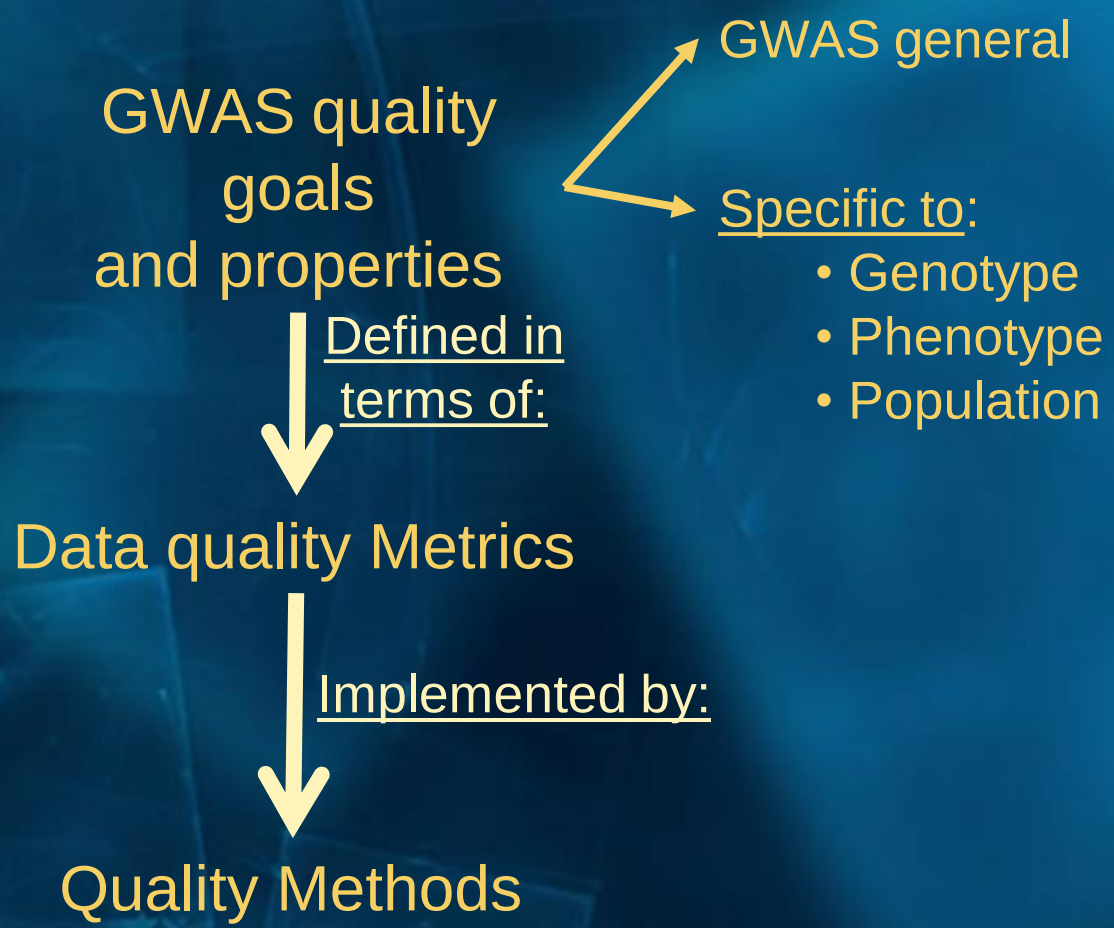
- Specify a GWAS Data Quality model:
  - Representing the meaningful data quality properties for Meta-Analysis-based GWAS
  - Implementable ➔ include data quality measurement methods
  - Top-down approach: from high level goals to low level methods
- Implement GWAS data quality management Tool
  - Enable interactive exploration on data quality of the GWASs
  - Evaluate data quality of the existing GWASs
  - Define user oriented quality properties and their computation methods

GWAS Data Quality Model

Microsoft Research

- High level quality goals:
  - G1: Combinability
    - How combinable a GWAS is
  - G2: Study data quality
    - Accuracy, Precision, Completeness, Specificity, etc.
  - G3: Reputation
    - Depending on cross-validations, study follow-up, history of studies
  - G4: Accessibility
    - Availability, performance, privacy and security regulations
  - G5: Statistical power
    - How the study enables better statistical reasoning

- Structure of the quality model:



Combinability of Genotype Data

Microsoft Research

SNP based plaftform overlap

	Affymetrix 100K	Affymetrix 500K	Illumina 100K	Illumina 550K	Perlegen 600K
Affymetrix 100K	100,0	23,5	4,2	32,0	22,4
Affymetrix 500K	5,5	100,0	3,2	16,9	26,8
Illumina 100K	4,5	14,7	100,0	22,5	23,5
Illumina 550K	6,7	15,2	4,4	100,0	35,1
Perlegen 600K	4,3	22,4	4,3	32,6	100,00

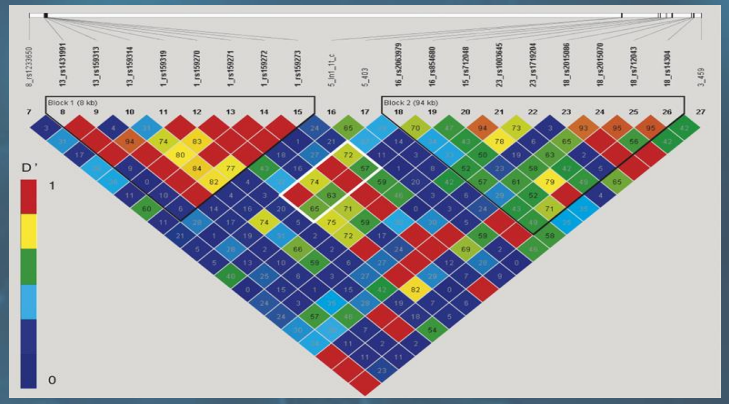
COMMERCIAL GENOTYPING PLATFORMS

AFFYMETRIX illumina PERLEGEN

SNP set intersection based on refSNP numbers (rs) from dbSNP Project

Correlated SNPs groups

	SNP_1	SNP_2	...	SNP_n
$Platform_1$	1	0.90		1
$Platform_M$	0.85	0.86		0.87

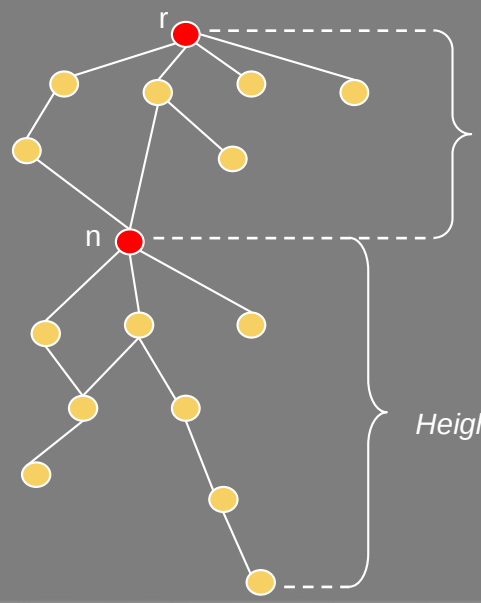


SNP coverage measure based on Linkage Disequilibrium (LD) [HapMap Proj.]

$\text{Max} \{ r^2(SNP_n, SNP(platform_M)) \}$

Phenotype: Mapping Terms to SNOMED

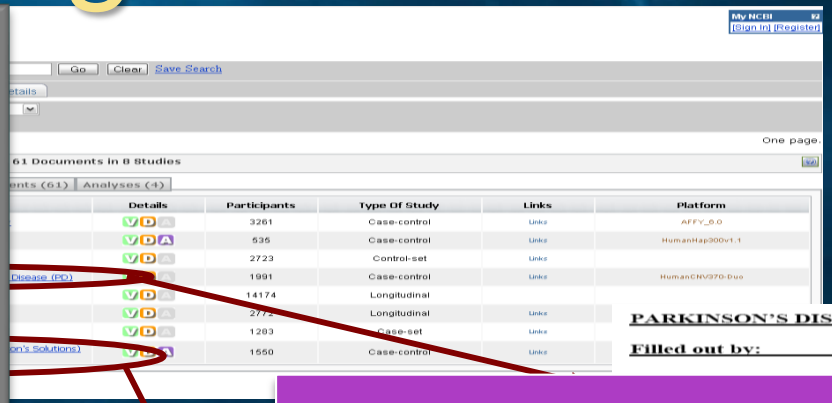
Microsoft Research

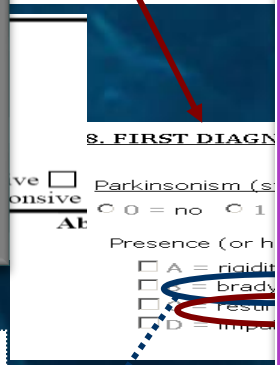


LengthMax(r, n) = 3

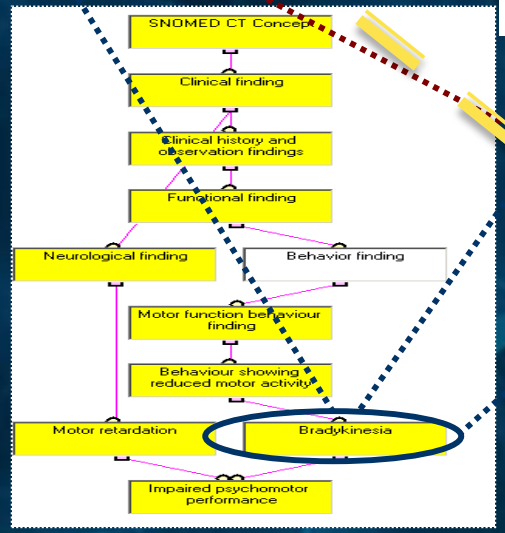
SpecificityDegree(n) = $\frac{4}{7}$

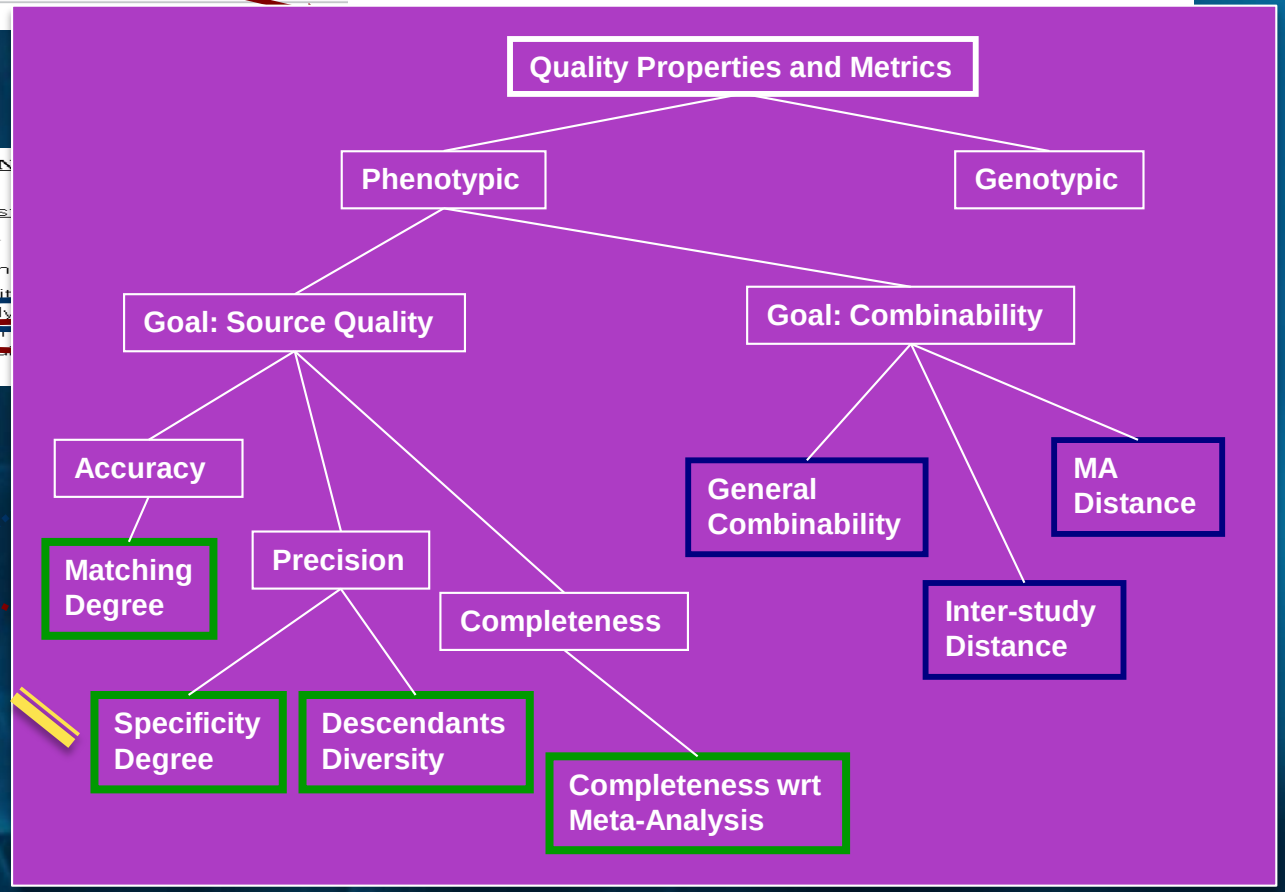
HeightMax(n) = 4





CliniClue® SNOMED browser





Quality Model → Management Tool

Microsoft Research

GWAS OBJECTS

GENOMIC DATA

PHENOTYPIC DATA

POPULATION DATA

Identify GWASs with adequate quality level in the variables of interest.

Qualified according to

QUALITY FRAMEWORK

- G1: Combinability
- G2: Study data quality
- G3: Reputation of the GWAS
- G4: Accessibility of GWAS data
- G5: Statistical power

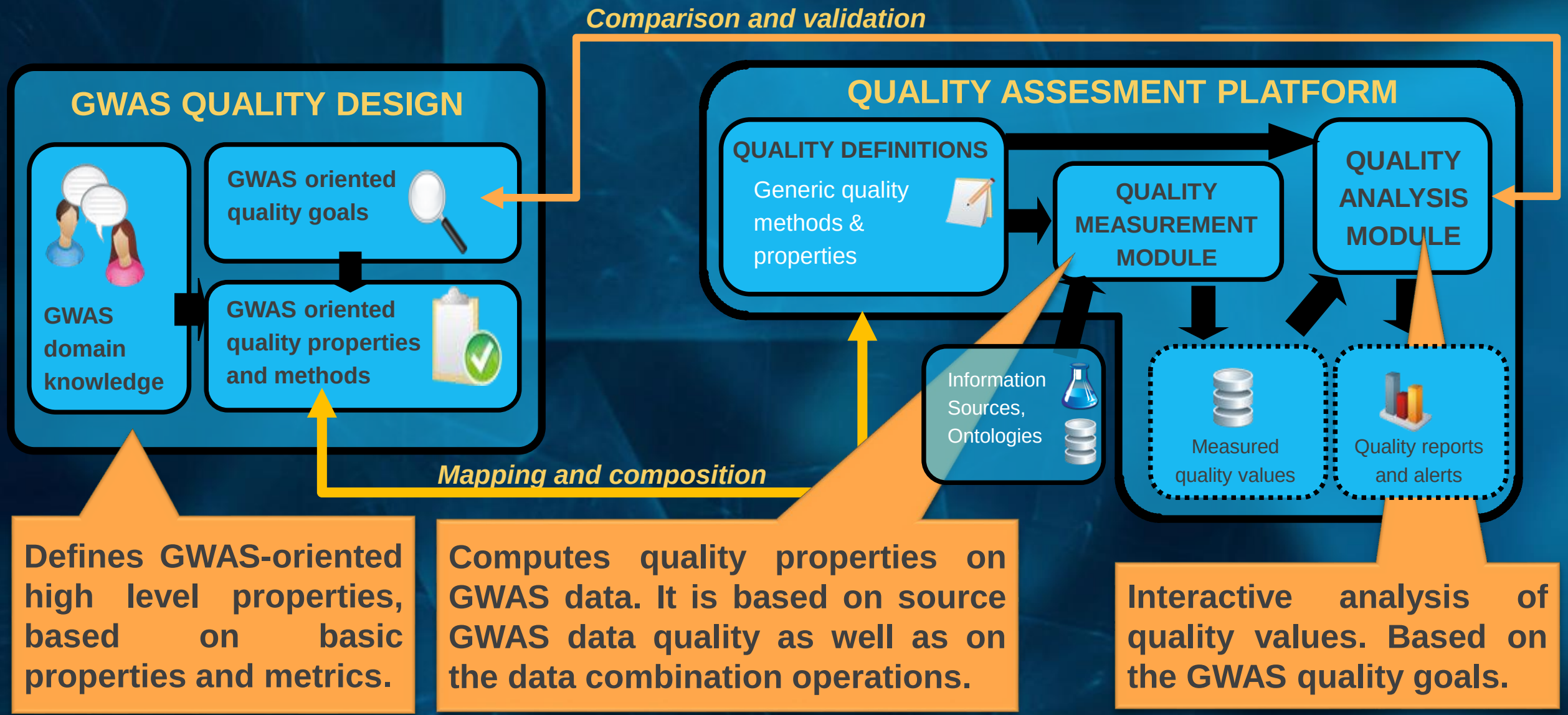
GWAS Data Quality Management tool

1. Specify GWAS data quality properties
2. Measure quality values on each GWAS
3. Analysis of quality requirement satisfaction on multiple GWAS
4. Analyze of quality results on GWAS combination

Adjust quality property definitions

Data Quality Management Tool

Microsoft Research



Defines GWAS-oriented high level properties, based on basic properties and metrics.

Computes quality properties on GWAS data. It is based on source GWAS data quality as well as on the data combination operations.

Interactive analysis of quality values. Based on the GWAS quality goals.

Technical Approach

Microsoft Research

- Quality analysis ➔ OLAP + Customizations
  - Using SQL Server 2008 and Analysis Services
  - Specific Roll-Up functions. Intensive use of MDX language
  - DSLs: GWAS specific languages for quality properties and for analysis conditions
  - Starting with Excel as OLAP client, towards a specific interface
- Using SNOMED as medical ontology
  - The basis to define phenotypic quality properties
  - Assisted mapping tool: GWAS ➔ SNOMED
- Large data processing ➔ Cloud computing
  - Genomic data: aprox. 50 GB (hundreds of GB in public repos)

Conclusions

Microsoft Research

- Meta-analysis is a powerful tool for performing GWAS
  - However, studies' combinability should be guided in a clear/objective way
  - The developed Data Quality Framework for GWAS enable to systematically apply such approach
- Specification of GWAS Data Quality Framework:
  - Constitutes a main achievement of the project; however, still remains much to be done
  - Privacy regulations and information restrictions are main issues for info availability
- The implemented tool:
  - Provides analytical functions to explore the data quality characteristics of GWAS
  - Intends to guide the analyst in the process of reusing studies
- The prototype is based on several available tools:
  - Shows the potential of "general purpose" software to be used in the scientific area

Perspectives

Microsoft Research

- GWAS:
  - Extending quality goals and metrics (on genotype and phenotype info)
  - Specifying “population environment” quality goals and metrics
  - Developing a module for automatic search of risk factors
  - Addressing privacy and data restrictions through a *GWAS Data Quality Service* ?
- Biological data quality:
  - Massive methods (High Throughput) are nowadays pervasive in biology and data quality approaches could led light in data sharing
  - Extending our approach to microarrays or Next Generation Sequencing is straightforward
- Medical/Healthcare application:
  - The concepts and quality metrics developed could be interesting in Medical/Healthcare recording
  - Assisted Medical Diagnosis is another very interesting application

Microsoft Research

Thank you very much

- Contact:
  - http://www.fing.edu.uy/inco/grupos/sibio/gwas/

  - Hugo Naya, naya@pasteur.edu.uy
  - Raul Ruggia, ruggia@fing.edu.uy

Microsoft Research

Microsoft®

Your potential. Our passion.™