Data model experiences from various EU projects MAGETAB, FuGE, XGAP, Pheno-OM

Metadata Capture Symposium
November 10, Utrecht

Morris Swertz and biodata modelers around the world*

BBMRI-NL, EU-GEN2PHEN, EU-CASIMIR, EU-EURATRANS, LifeLines, EU-SYSGENET, EU-PANACEA, NBIC and other consortia

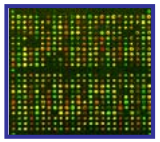
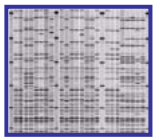
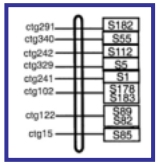
Genomics Coordination Center Groningen

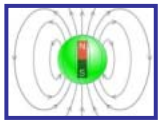
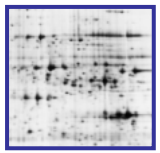
Genomics Coordination Center – UMC/University Groningen





etc.

etc.

GEN2PHEN

university of groningen

FIMM

LIFELINES

SYSGENET

CASIMIR

nbicassist

WAGENINGEN UR
For quality of life

EMBL-EBI

umcg

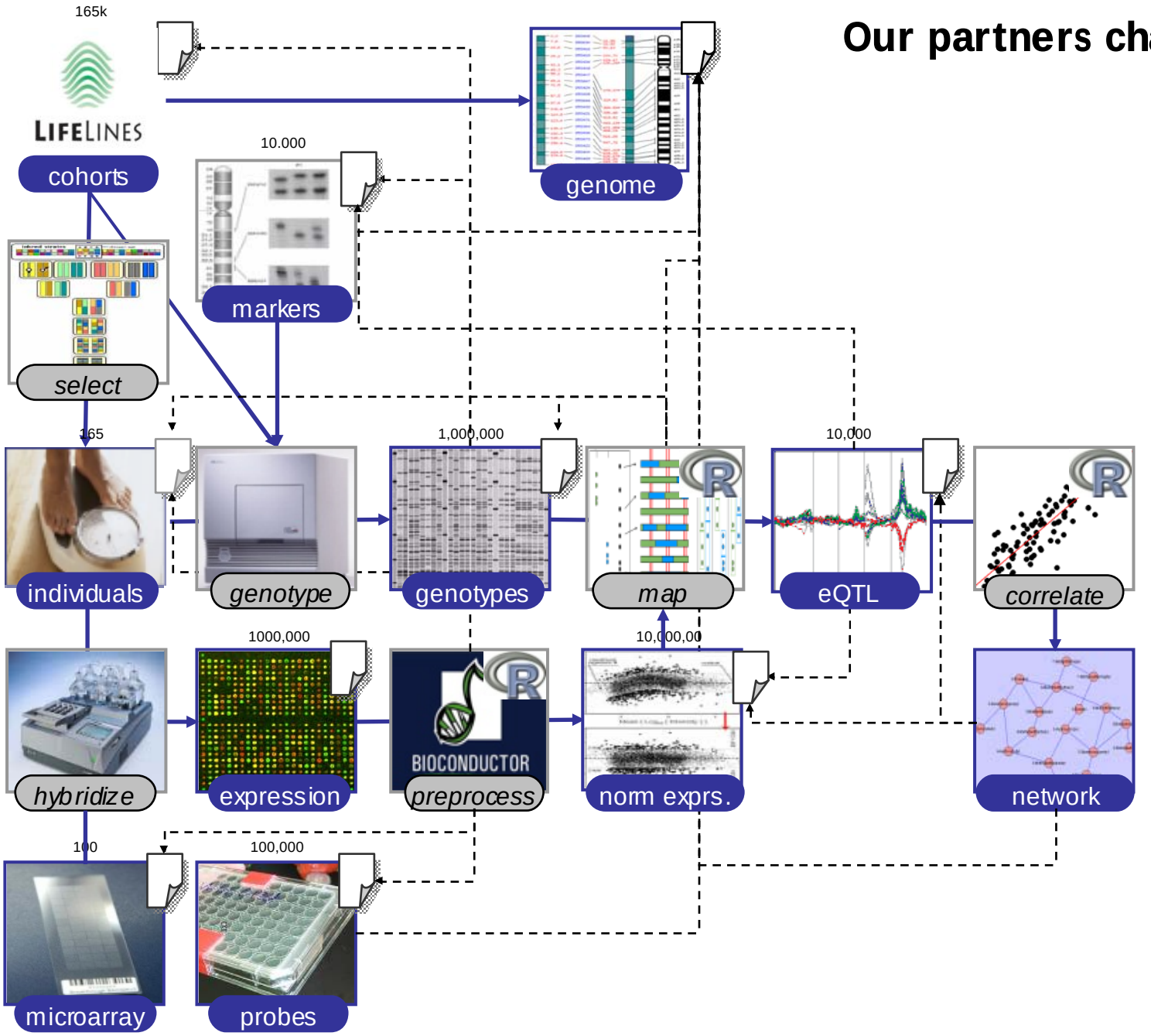
IOP

BBMRI • NL

University of Leicester

PANACEA

Our partners challenges



Some models

- MAGE-TAB for microarrays
  - Oldest model, born from experience microarray community
- Pheno-OM for phenotypes
  - Recent model, but present now as it shows core concepts
  - Born from XGAP and PaGE
- XGAP for all phenotypes/genotypes
  - Extended version of Pheno-OM and FuGE
- FuGE for functional genomics experiments
  - Examplar for its protocol – protocolApplication structure
- Others like ISA-TAB (aka MAGE-TAB++), CHADO, HL7, endless list

=> I will show convergence of these models

MAGE-TAB (format)

Born out of failed MAGE-OM (100+ pages spec)

Originally for microarrays

Contents:

- IDF Investigation description file
- SDRF Sample-Data-Relationship file
- Data files: raw format or matrix files
- Optional: array design file(s)

Widely used in many production systems!

http://www.mged.org/mage-tab/
http://www.ebi.ac.uk/microarray/doc/help/MAGE-TAB.html
Rayner et al (2006) BMC Bioinformatics 7, 489

IDF

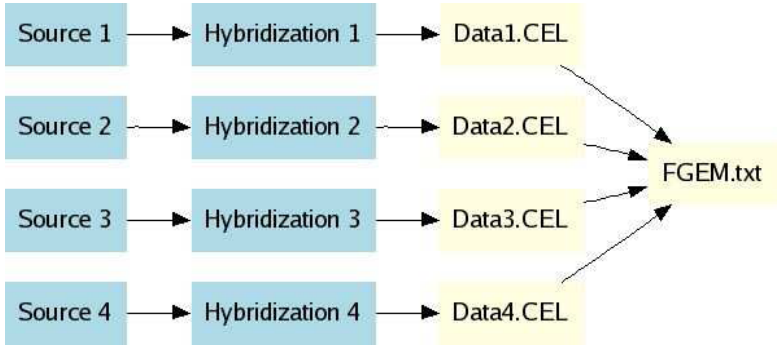
Investigation Title	University of Heidelberg H sapiens TK6		
Experimental Design	genetic_modification_design	time_series_design	
Experimental Factor Name	Genetic Modification	Incubation Time	
Experimental Factor Type	genetic_modification	time	
Experimental Factor Term Source REF	MGED Ontology	MGED Ontology	
Person Last Name	Maier	Fleckenstein	Li
Person First Name	Patrick	Katharina	Li
Person Email	patrick.maier@radonk.ma.uni-heidelberg.de		
Person Phone	+496213833773		
Person Address	Theodor-Kutzer-Ufer 1-3		
Person Affiliation	Department of Radiation Oncology, University of Heidelberg		
Person Roles	submitter; investigator	investigator	investigator
Person Roles Term Source REF	MGED Ontology	MGED Ontology	GED Ontology
Quality Control Type	biological_replicate		
Quality Control Term Source REF	MGED Ontology		
Replicate Type	biological_replicate		
Replicate Term Source REF	MGED Ontology		
Date of Experiment	2005-02-28		
Public Release Date	2006-01-03		
PubMed ID	12345678		
Publication Author List	Patrick Maier; Katharina Fleckenstein; Li Li; Stephanie Laufs; Jens Zeller; Stefan Fruehauf; Carsten Herskind; Frederik Wenz		
Publication Status	submitted		
Experiment Description	Gene expression of TK6 cells transduced with an oncoretrovirus expressing MDR1 (TK6MDR1) was compared to untransduced TK6 cells and to TK6 cell transduced with an oncoretrovirus expressing the Neomycin resistance gene (TK6neo). Two biological replicates of each were generated and the expression profiles were determined using Affymetrix Human Genome U133 Plus2.0 GeneChip microarrays. Comparisons between the sample groups allow the identification of genes with expression dependent on the MDR1 overexpression.		
Protocol Name	GROWTHPRTCL10653		TRANPRTCL10656
Protocol Type	grow	action	bioassay_data_transformation
Protocol Description	TK6 cells were grown in susp RPMI 1640 medium supplem horse serum (Invitrogen, Kar The cells were routinely ma ed at 37 C and 5% CO2.	6 cells were r (Qiagen). Total RNA was extracted from the ce using an RNeasy kit	Mixed Model Normalization with SAS Micro Array Solutions (version 1.3).
Protocol Parameters	media; time	Extracted Product; A	
Protocol Term Source REF	MGED Ontology	MGED Ontology	
SDRF File	e-mexp-428_tab.txt		
Term Source Name	Cell Type Ontology	MGED Ontology	NCI Metathesaurus
Term Source File	http://obo.sourceforge.net/cgi-bin/detail.cgi?cell	http://mged.sourceforge.net/ontologies/MGEDontology.php	http://ncimeta.nci.nih.gov/indexMetaphrase.html
Term Source Version		1.3.0.1	

Repeating columns

Separated Values

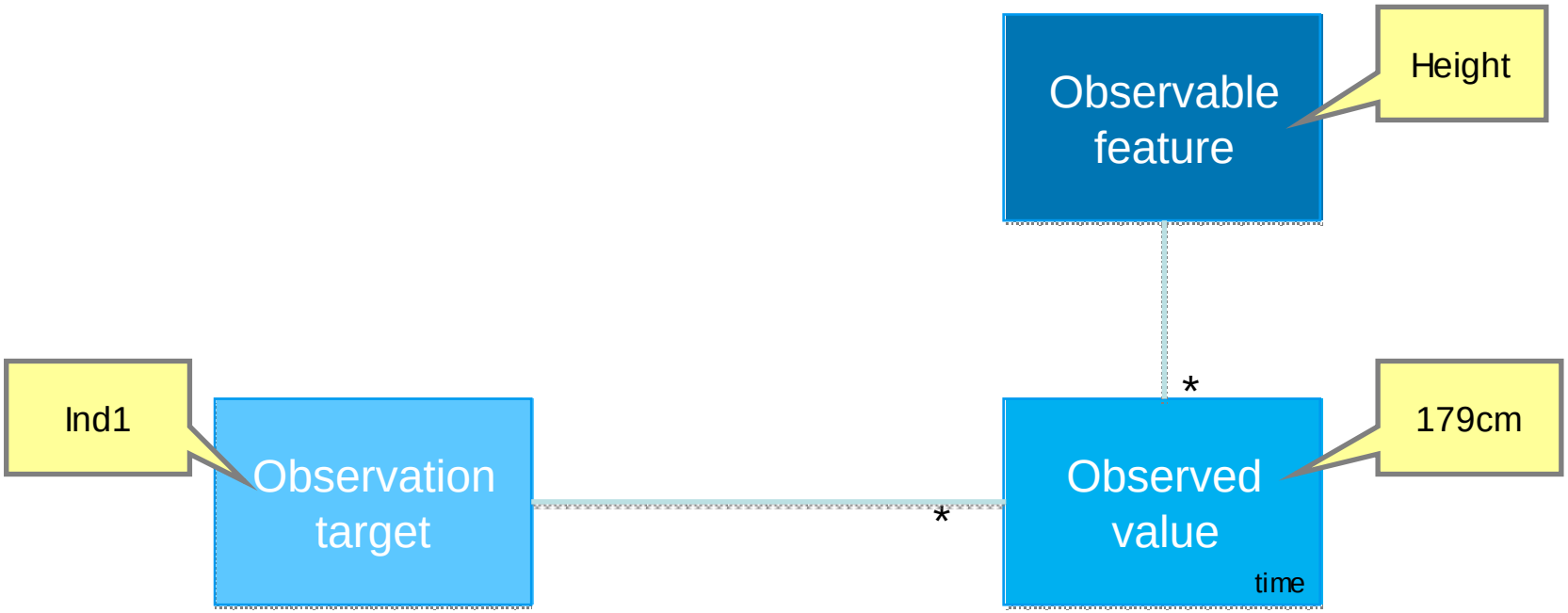
Ontology Sources

SDRF

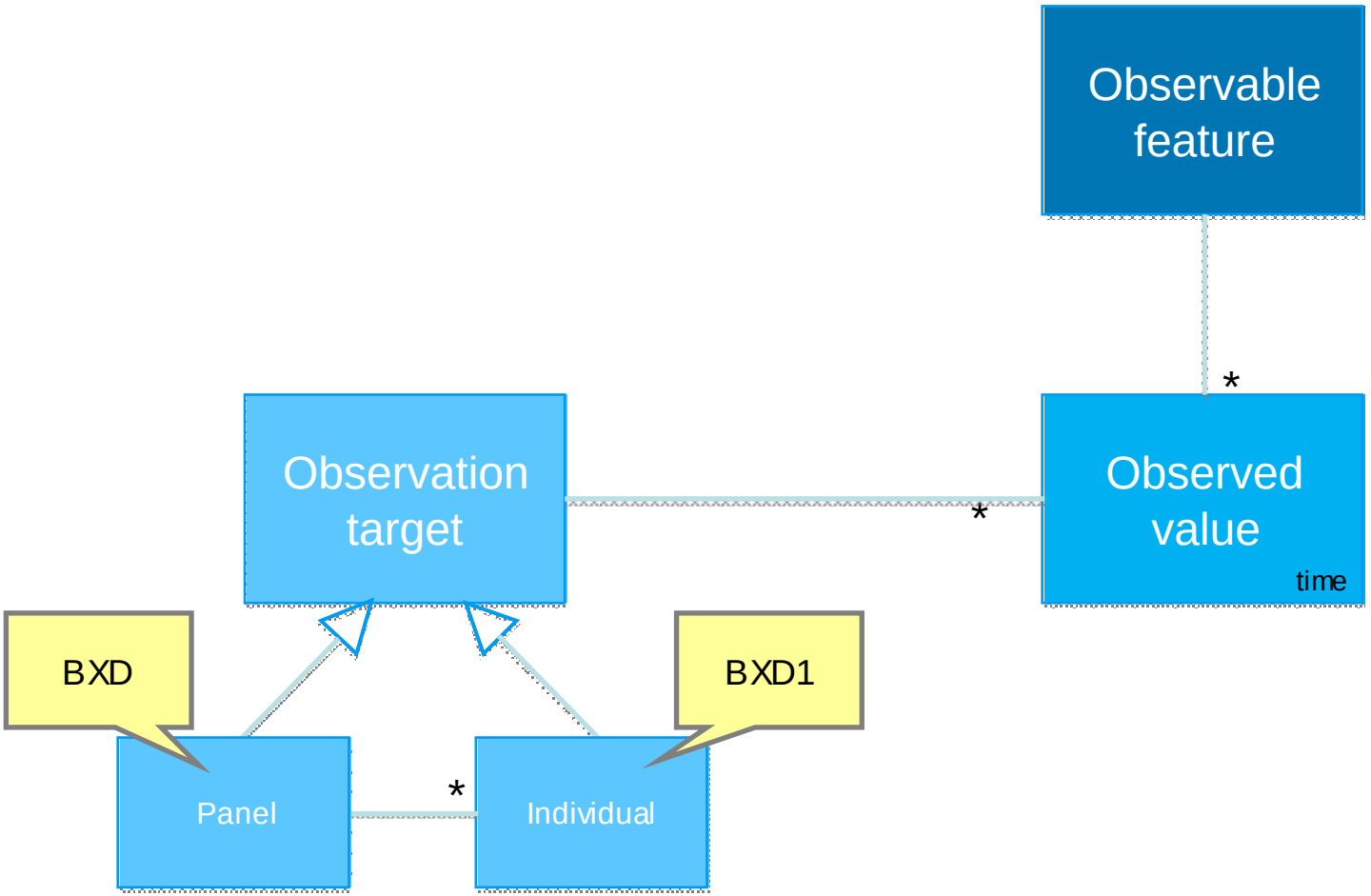


Source Name	Protocol REF	Hybridization Name	Array Data File	Derived Array Data Matrix File
Source 1	P-XMPL-10	Hybridization 1	Data1.CEL	FGEM.txt
Source 2	P-XMPL-10	Hybridization 2	Data2.CEL	FGEM.txt
Source 3	P-XMPL-10	Hybridization 3	Data3.CEL	FGEM.txt
Source 4	P-XMPL-10	Hybridization 4	Data4.CEL	FGEM.txt

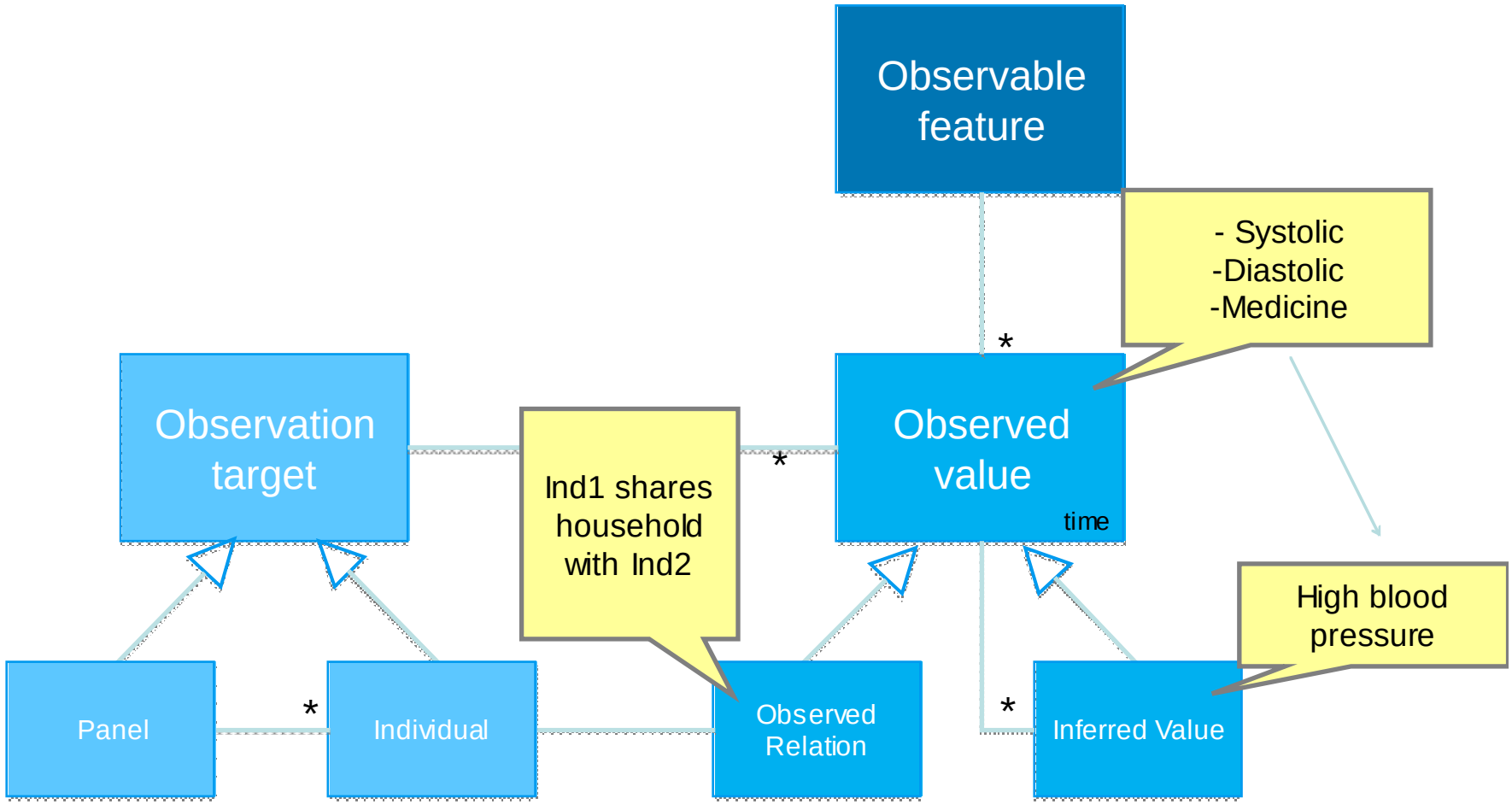
Pheno-OM ... for phenotypes



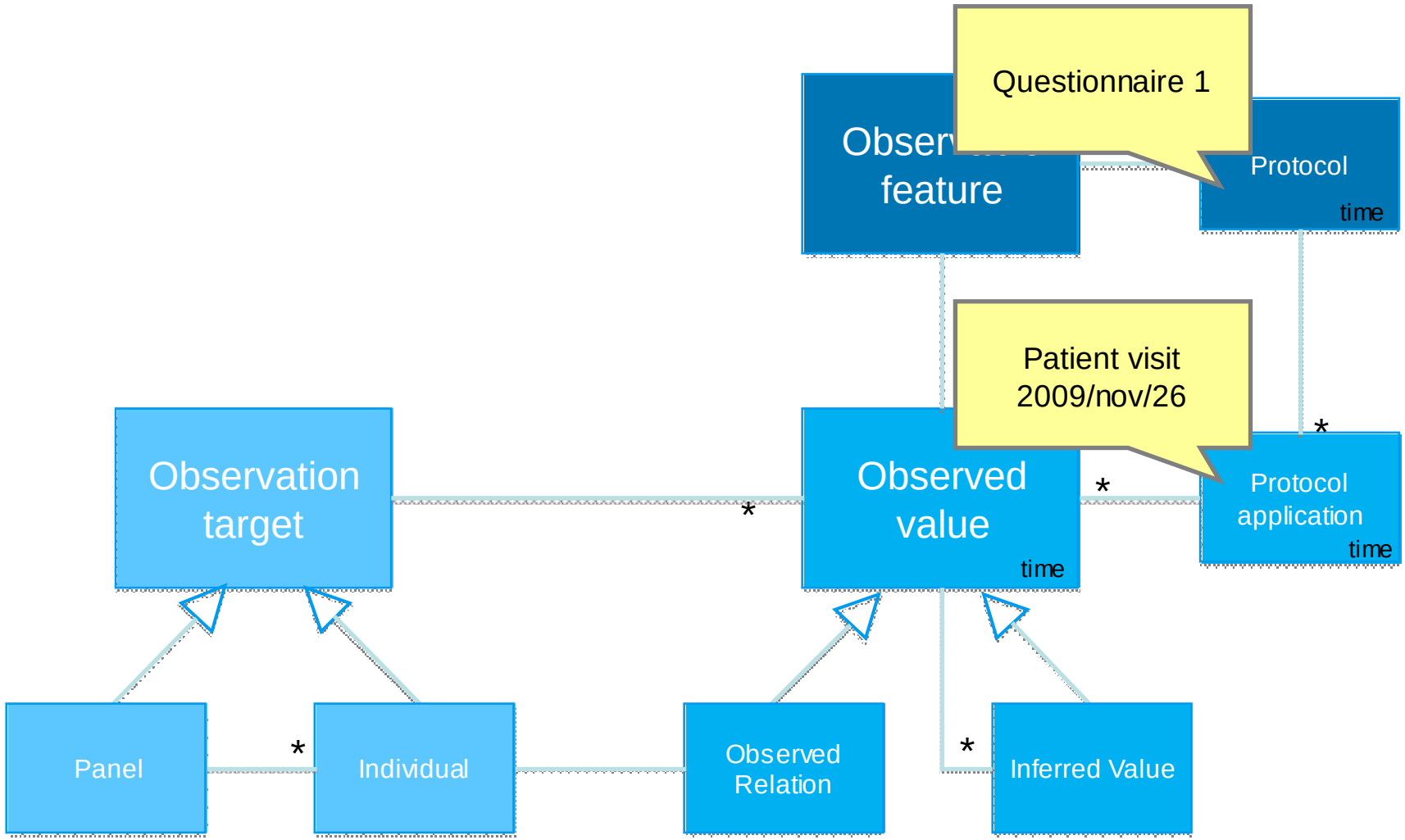
Individual and panel level data



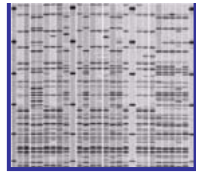
Values, relations, inferences



Protocols/SOPs used ...



XGAP for genotypes (superset of pheno-OM)



Data in matrices

Genotype data

Subjects: PANELS

TMrAaRiKtEsRS

DATA ELEMENTS

TRAIT × *SUBJECT*

Swertz et al (2010) Genome Biology 11(3).

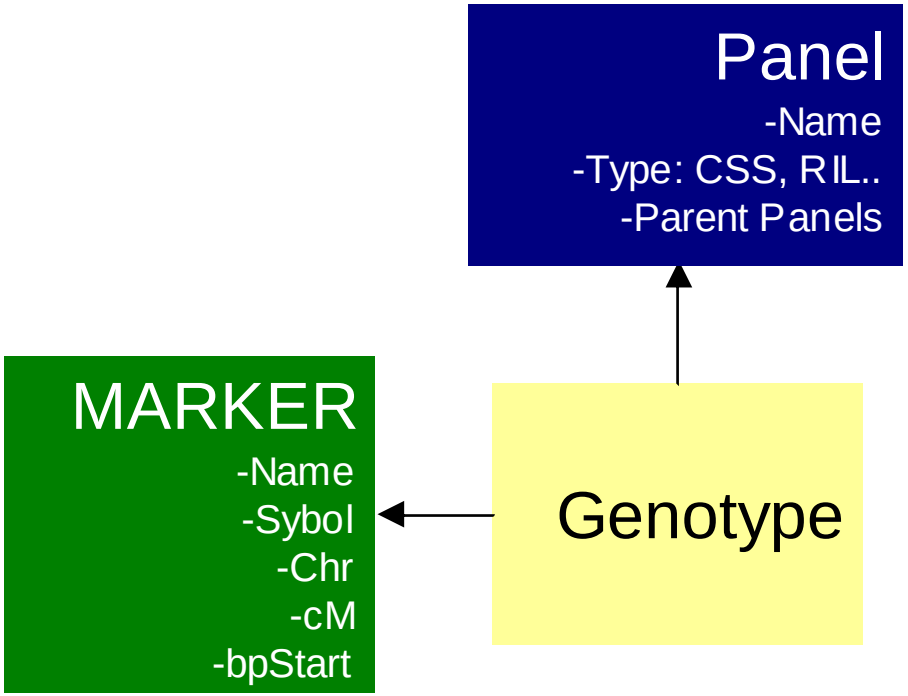
Minimal model

- Annotations in tables, e.g. Marker

MARKER
-Name
--Symbol
-Chr
-cM
-bpStart

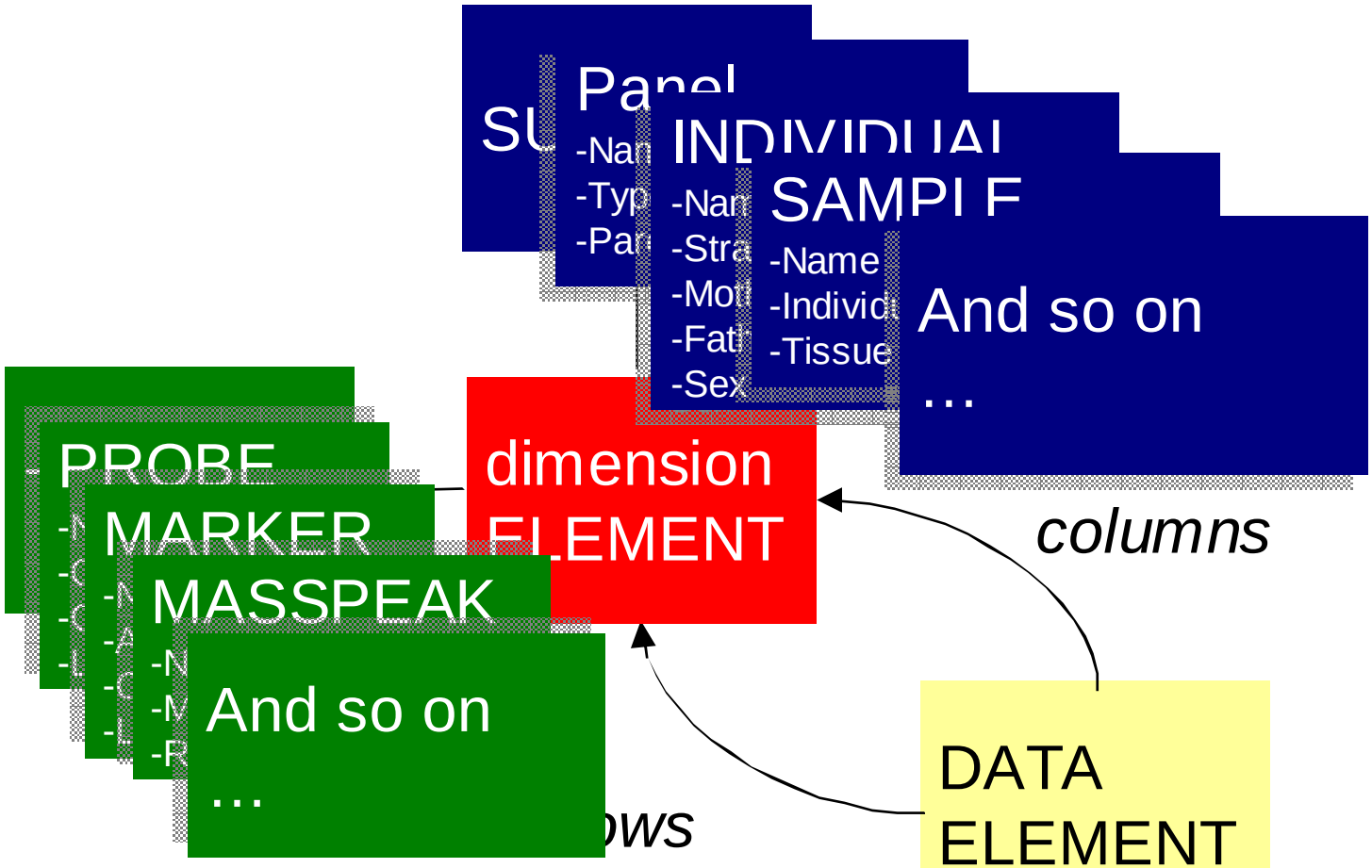
Name	Symbol	Chr	cM	bpStart	mb
C1M1	I_1_pkP1050	1	-18.2603	168807	0.168807
C1M2	I_2_pkP1101	1	-17.2825	992188	0.992188
C1M3	I_3_pkP1103	1	-11.959	1884415	1.884415
C1M4	I_4_pkP1052	1	-6.1004	2818973	2.818973
C1M5	I_5_egPE107	1	-3.5488	3502476	3.502476
C1M6	I_6_egPF101	1	-1.4887	4338254	4.338254
C1M7	I_7_pkP1054	1	-0.6162	4845515	4.845515
C1M8	I_8_egPH102	1	0.4597	5893622	5.893622
C1M9	I_9_pkP1057	1	0.9366	6359867	6.359867
C1M10	I_10_pkP1116	1	2.1576	7589863	7.589863
C1M11	I_11_egPK103	1	2.4087	7894081	7.894081
C1M12	I_12_pkP1059	1	2.9456	8654360	8.65436
C1M13	I_13_pkP1122	1	3.7959	9569914	9.569914
C1M14	I_14_egPN104	1	4.7801	10259909	10.259909
C1M15	I_15_egPO105	1	6.0193	11085295	11.085295
C1M16	I_16_pkP1068	1	7.5226	11760182	11.760182

Model: try 1



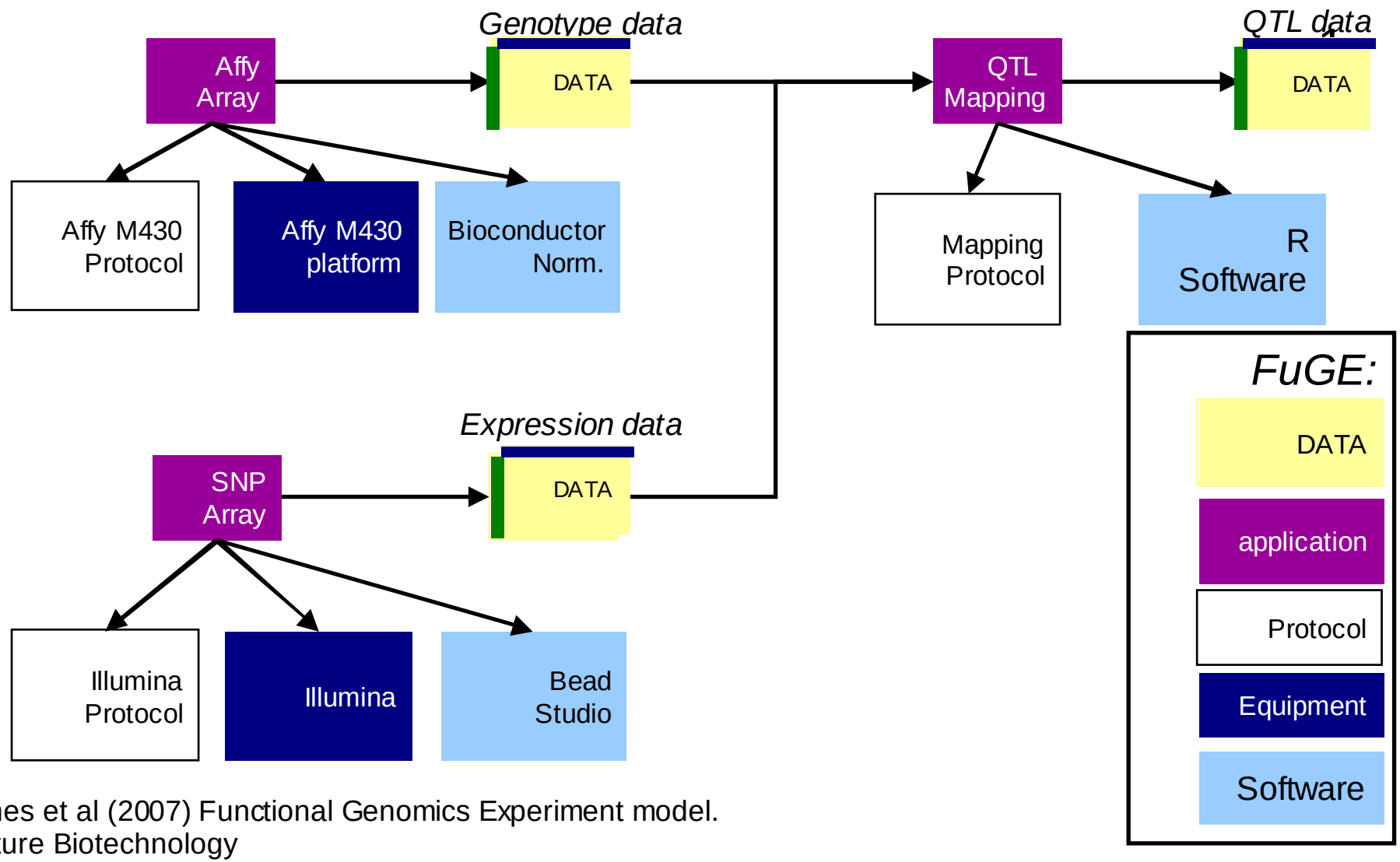
Swertz et al (2010) Genome Biology 11(3).

XGAP (extension based variation mechanism)



Swertz et al (2010) Genome Biology 11(3).

XGAP extends on FuGE



Jones et al (2007) Functional Genomics Experiment model. Nature Biotechnology

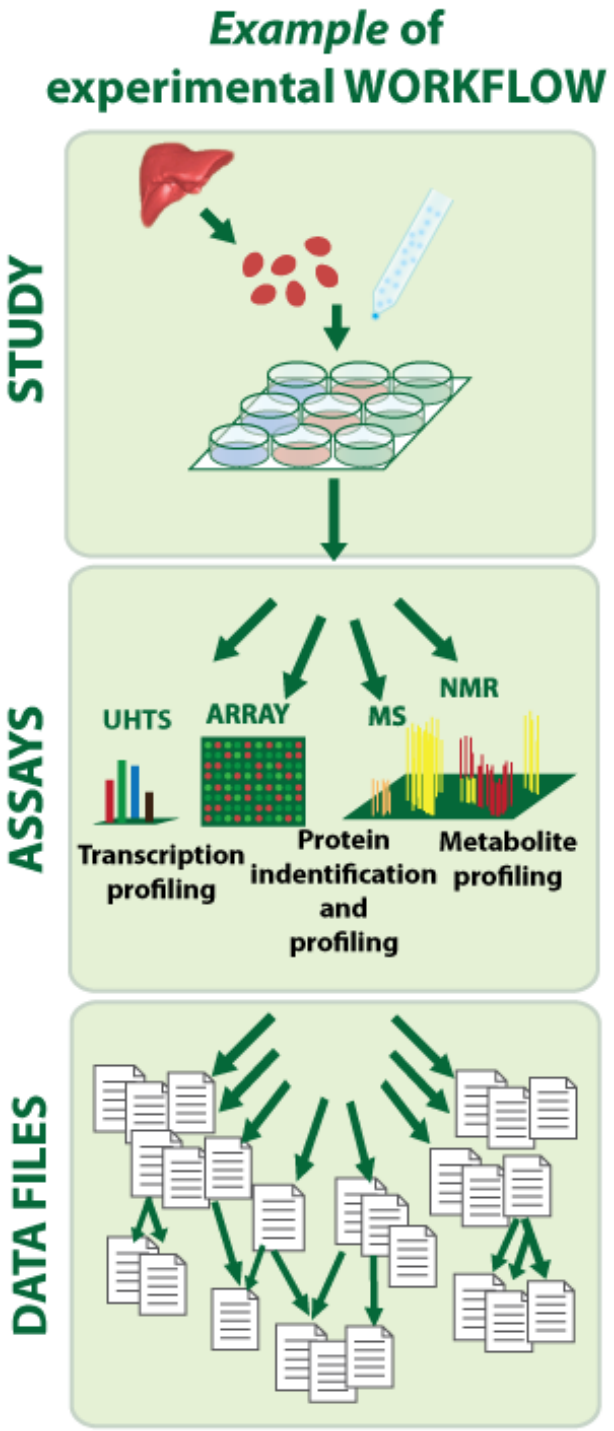
ISA-TAB(generic model)

Differs from MAGE-TAB

- Nested investigations (as studies)
- To have templates assays
- More aligned to FuGE
- But some find it still difficult
  - Reflects largely the complexity of the experiments not that of the format.

ISA =

- Investigation
- Study (Investigation component)
- Assay (a component of Study)
- Data files
  - Used at the Harvard Stem Cell Discovery Lab, NERC and Sysmo-DB



http://isatab.sf.net

Current work: merging the models

Pheno-OM	FuGE + XGAP	MAGE-TAB	PaGE
Investigation	Investigation	Investigation	Investigation
ObservableFeature	Trait	~Characteristic	Observable_feature
ObservationTarget	Subject	Source	Abstract_observation_target
Individual	Individual	NA	Individual
Panel	Panel	NA	Panel
ObservedValue	DataElement	CharacteristicValue	ObservedValue
ObservedRelationship	NA	NA	NA
InferredValue	via ProtocolApplication	NA	ObservedValue (self assoc)
OntologySource	OntologySource	OntologySource	Ontology_source
OntologyTerm	OntologyIndividual	OntologyTerm	Ontology_term
Protocol	Protocol	Protocol	Observation_method
ProtocolApplication	ProtocolApplication	ProtocolApplication	Observation_method
Code	NA	~Characteristic	NA

Discussion for afternoon

- How to properly map this to RDF
- Map data values to RDF/OWL
  - Most models ontology enabled
- Map data structures for RDF/OWL
  - So annotate data model
- Some experiences with MOLGENIS D2RQ
  - Report of GEN2PHEN internship student