

TARGET Conference 2013
Probing Big Data for answers
3 – 5 April, 2013
Groningen, Netherlands

Flexible datamodels for life sciences

**K. Joeri van der Velde, Martijn Dijkstra, Morris Swertz,
members of the Genomics Coordination Center**



Europese Unie
Europees Fonds voor Regionale Ontwikkeling



SNN

SENSOR
UNIVERSE



umcg



university of
groningen

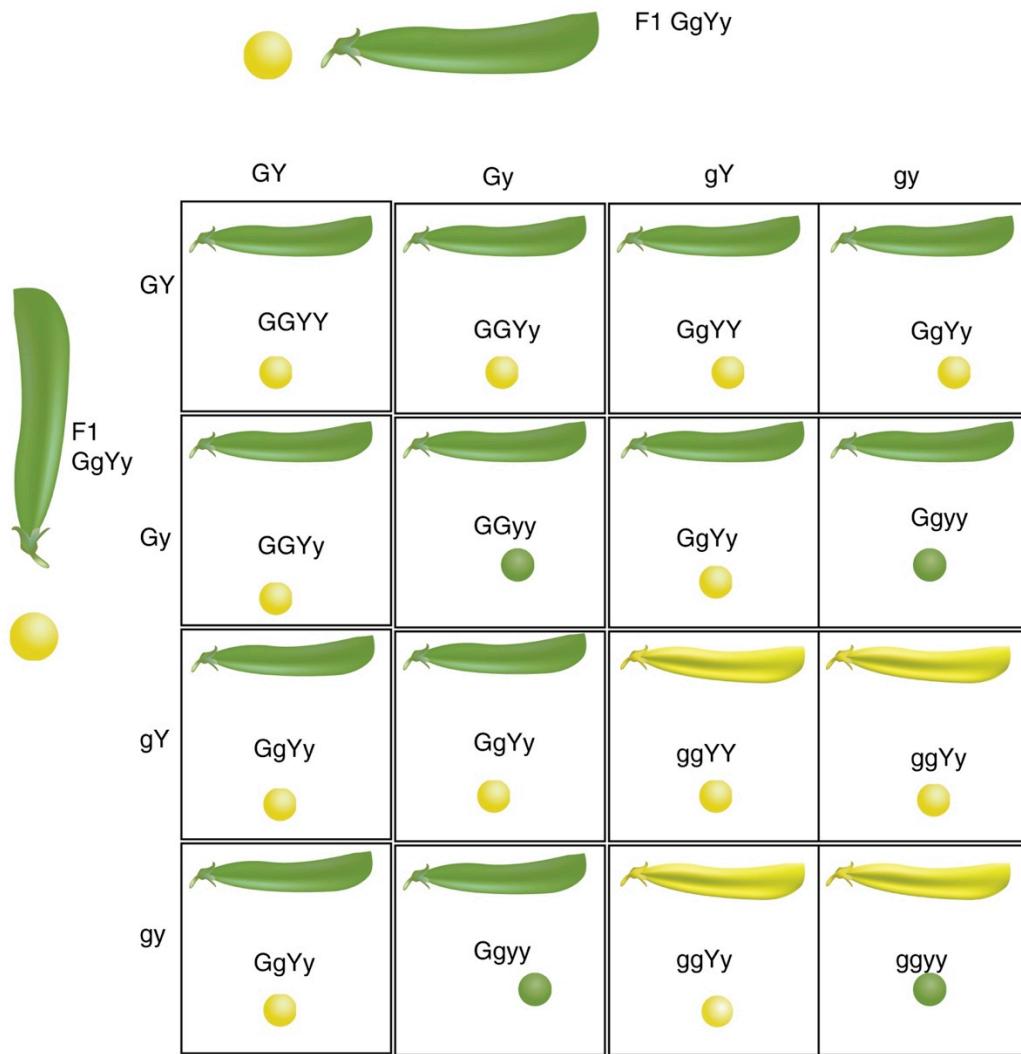
genomics coordination
center

Outline

- Introduction
 - Motivation
 - Dealing with variation in research needs
 - Design-time vs runtime configuration
- Results
 - XGAP model for homogeneous data ('molecules')
 - Observ-OM model for heterogenous data ('phenotypes')
- Showcase
 - EB Registry
 - WormQTL
- Current work

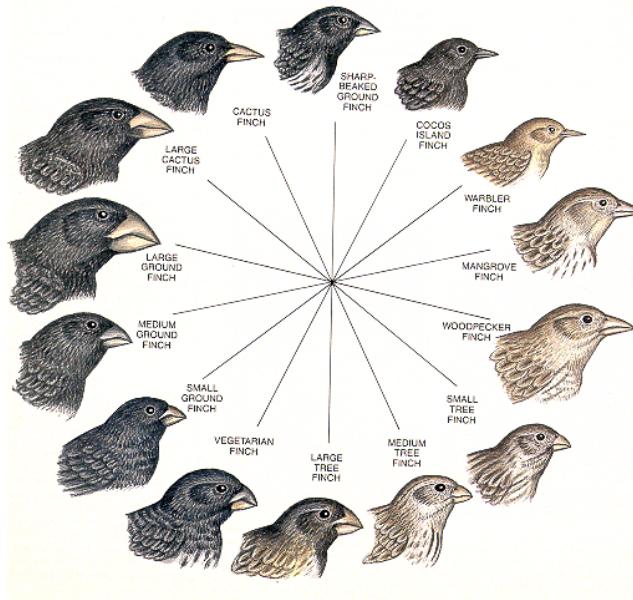
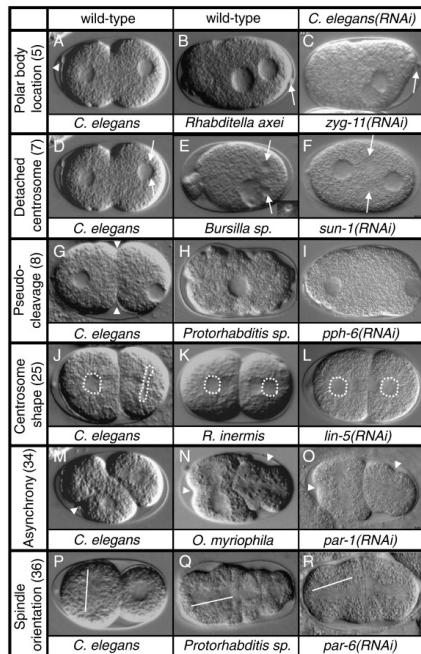
Introduction Motivation

Biological variation first explained by genetics

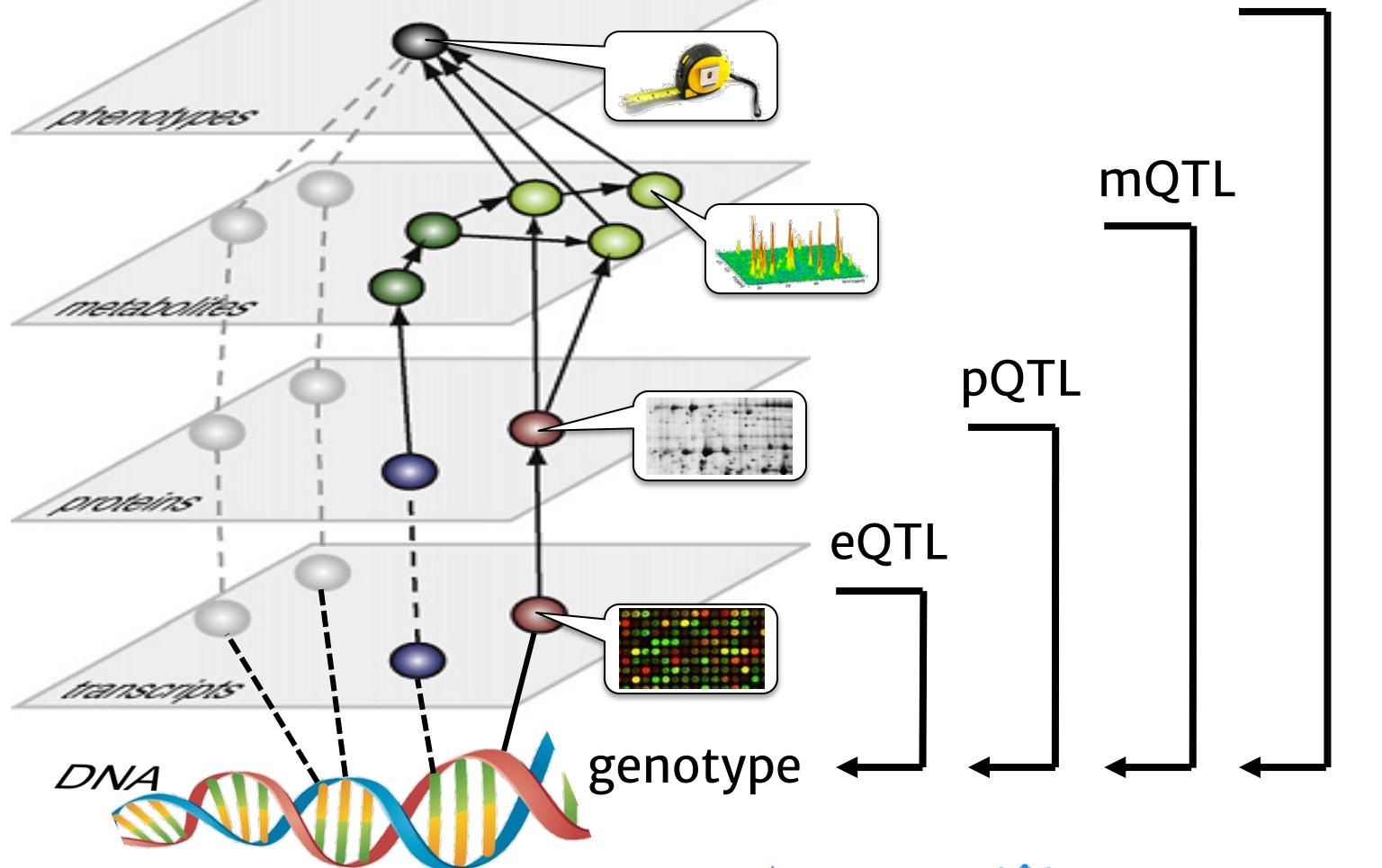


*Mendel got lucky..
single-locus traits*

Motivation: Understanding complex variation



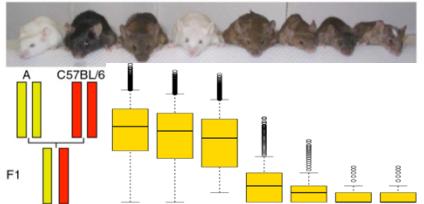
How? Understanding geno-to-pheno



Introduction

Dealing with variation in research needs

What needs to be done?



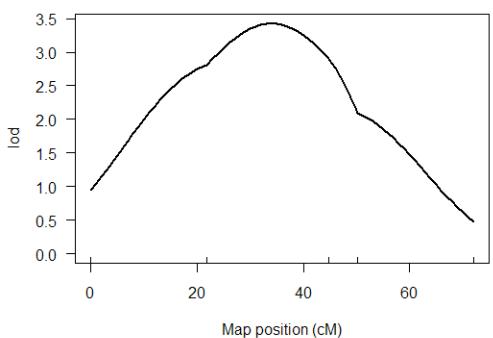
Project that generates data

‘My’ data

H1	0.3363646	0.4403331	0.3944101	0.9123233
H2	0.3363646	0.4403331	0.4400896	0.1241076
H3	0.3363646	0.4403331	0.4400896	0.1241076
H4	0.3363646	0.4403331	0.4400896	0.1241076
H5	0.3363646	0.4403331	0.4400896	0.1241076
H6	0.3363646	0.4403331	0.4400896	0.1241076
H7	0.3363646	0.4403331	0.4400896	0.1241076
H8	0.3363646	0.4403331	0.4400896	0.1241076
H9	0.4403331	0.4403331	0.4400896	0.1241076
H10	0.4403331	0.4403331	0.4400896	0.1241076
H11	0.4403331	0.4403331	0.4400896	0.1241076
H12	0.4403331	0.4403331	0.4400896	0.1241076
H13	0.4403331	0.4403331	0.4400896	0.1241076
H14	0.4403331	0.4403331	0.4400896	0.1241076
H15	0.4403331	0.4403331	0.4400896	0.1241076
H16	0.4403331	0.4403331	0.4400896	0.1241076
H17	0.4403331	0.4403331	0.4400896	0.1241076
H18	0.4403331	0.4403331	0.4400896	0.1241076
H19	0.4403331	0.4403331	0.4400896	0.1241076
H20	0.4403331	0.4403331	0.4400896	0.1241076
H21	0.4403331	0.4403331	0.4400896	0.1241076
H22	0.4403331	0.4403331	0.4400896	0.1241076
H23	0.1092371	0.02043251	0.0150711	0.01500009
H24	0.1092371	0.02043251	0.0150711	0.01500009
H25	0.1092371	0.02043251	0.0150711	0.01500009
H26	0.1092371	0.02043251	0.0150711	0.01500009
H27	0.1092371	0.02043251	0.0150711	0.01500009
H28	0.1092371	0.02043251	0.0150711	0.01500009
mean	0.1092371	0.02043251	0.0150711	0.01500009

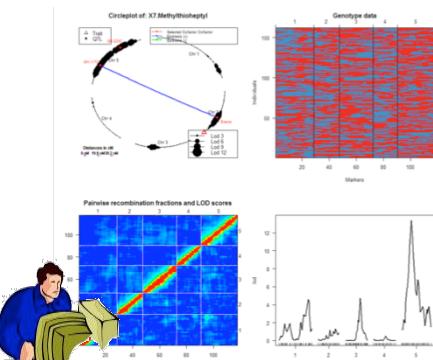


Direct interpretation



A little extra effort

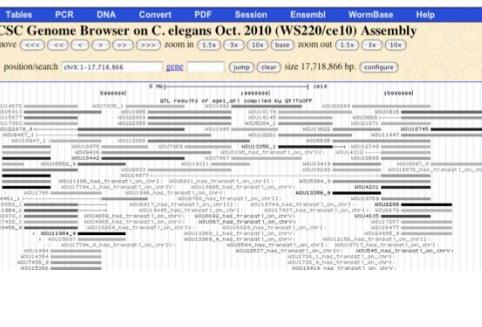
Rich interpretations,
geno-to-pheno



‘Team’ data



*C0A3 0.3751181790822	0.32256793827174	0.2746154
*C0A2 0.13287350660943	0.18064732468933	0.03367282
*C0A1 0.0082241283755568	0.0082241283755568	0.0082241283755568
*C0A0 0.0082241283755568	0.0082241283755568	0.0082241283755568
*C0A4d 0.1823216344987278	0.1823216344987278	0.1823216344987278
*C0A4d 1.00469140878811	1.00469140878811	1.00469140878811
*C0A4d 0.1823216344987278	0.1823216344987278	0.1823216344987278
*C0A3 0.79333184640894	0.591	0.591
*C0A3 0.635581547690894	0.591	0.591
*C0A4 0.388412604278893	0.4727292634161143	0.49797172
*C0A5 0.13287350660943	0.18064732468933	0.03367282
*C0A6 0.0082241283755568	0.0082241283755568	0.0082241283755568
*C0A7 0.0082241283755568	0.0082241283755568	0.0082241283755568
*C0A8 0.0082241283755568	0.0082241283755568	0.0082241283755568
*C0A9 0.0082241283755568	0.0082241283755568	0.0082241283755568
*C0A10 0.0082241283755568	0.0082241283755568	0.0082241283755568
*C0A11 0.0082241283755568	0.0082241283755568	0.0082241283755568
*C0A12 0.0082241283755568	0.0082241283755568	0.0082241283755568
*C0A13 0.0082241283755568	0.0082241283755568	0.0082241283755568
*C0A14 0.0082241283755568	0.0082241283755568	0.0082241283755568
*C0A15 0.0082241283755568	0.0082241283755568	0.0082241283755568
*C0A16 0.0082241283755568	0.0082241283755568	0.0082241283755568
*C0A17 0.0082241283755568	0.0082241283755568	0.0082241283755568
*C0A18 0.0082241283755568	0.0082241283755568	0.0082241283755568
*C0A19 0.0082241283755568	0.0082241283755568	0.0082241283755568
*C0A20 0.0082241283755568	0.0082241283755568	0.0082241283755568
*C0A21 0.0082241283755568	0.0082241283755568	0.0082241283755568
*C0A22 0.0082241283755568	0.0082241283755568	0.0082241283755568
*C0A23 0.0082241283755568	0.0082241283755568	0.0082241283755568
*C0A24 0.0082241283755568	0.0082241283755568	0.0082241283755568
*C0A25 0.0082241283755568	0.0082241283755568	0.0082241283755568
*C0A26 0.0082241283755568	0.0082241283755568	0.0082241283755568
*C0A27 0.0082241283755568	0.0082241283755568	0.0082241283755568
*C0A28 0.0082241283755568	0.0082241283755568	0.0082241283755568
mean	0.0082241283755568	0.0082241283755568



Introduction

Design-time vs runtime configuration

Autogenerate the software

Model in DSL



NextGenSeq



Gen

What models to use?

Mutation c

Can we have a model that rules them all?



Model organisms

Use generated software

Solexa Sequencer LIMS

| About | Object model | Report API | HTTP API | Web Services API

Flowcell Preps

flowcells

Add new record

id	flowcell
researcher_id	
flowcell_date	
reagent	
reagent2	
primer	
comments	
sequence_primer_id	

database of COL7A1 mutations

SearchDatabase UploadData Contact References Background Help Login

Find a specific mutation/variation

Variation: or Select mutations

Nucleotide No:

Amino Acid No:

Find mutations/variations

Exon/Intron: or Select exon/intron

Select mutation type: Select protein domain:

Search term:

Animal Observatory

Animals Projects Animals Events

Currently showing eventlog of animal ID 5.

Event ID around 2000-2510 in tree null at location null. Details: Healthy

See full description at 2007-12-17. Details: Not set

Last update at 2007-12-17. Last event

Selected animal ID: 5

Species: Rodent

AnimalID: 5

Date: 2007-12-17

Status: Alive

Sex: Male

Type: F1

Background: RICR:CD1-Tg(16G-CAG)C57BL

Mutation: Trisomy 21D

Genotype: +/+

Uterus: Healthy

Location: Horn

ProjectGroup: DefaultProject

Events

details

apply event to selected animals:

biometry | born | breeding | death | details | events | sample |

details

confirm event example:

MMR: December 17, 2007

confirm

<http://www.molgenis.org>

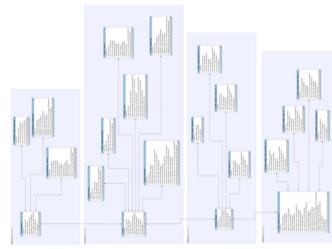
Swertz & Jansen (2007) *Nature Reviews Genetics* 8, 235-243

Swertz et al (2004) *Bioinformatics* 20(13), 2075-83

Examples of models.. How to choose?

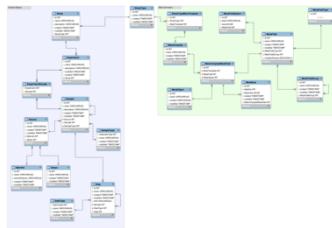
Stable
'design-time'

ISA-TAB
Investigation, study, sample



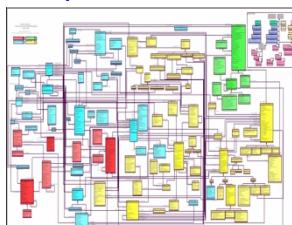
Dynamic
'runtime'

GSCF
Studies, flexible properties

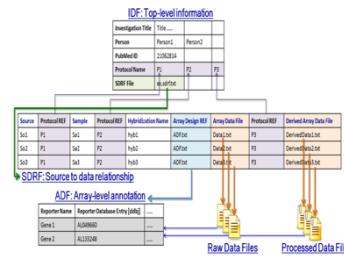


Exhaustive

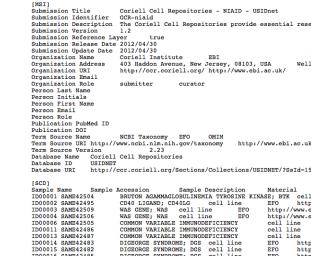
HL7
excerpt



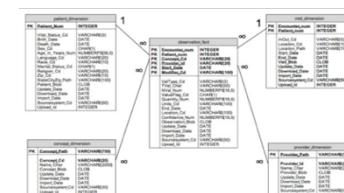
MAGE-TAB
Microarray provenance



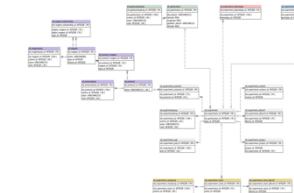
SAMPLE-TAB
Biobank sample tracking



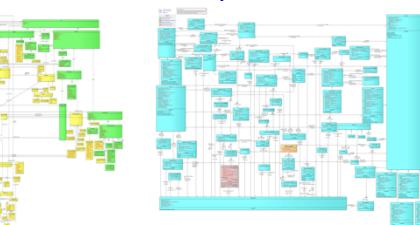
i2b2
Observations on patients
during doctor visit



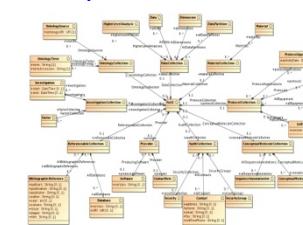
Chado NDM
Boilerplate for biological experiments and protocols



CDISC BRIDG
excerpt



FuGE
excerpt

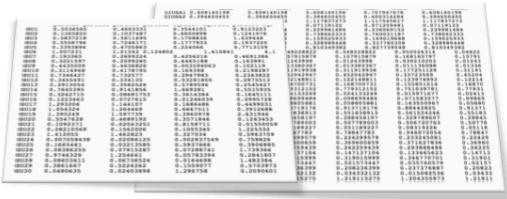


What are we dealing with?

Genomic features,
individuals, ontologies ..



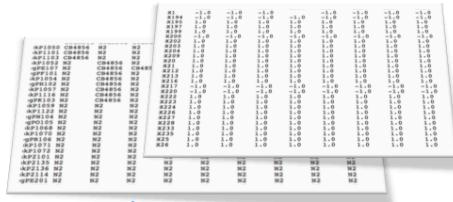
Biomolecular measurements,
association results ...



Metadata for phenotypes,
datasets, samples, panels ...



Genotypes,
conditions ...



Stable?

Annotations of concepts
used in data sets,
mostly static content

Dynamic?

Experimental data sets,
usually flexible and
volatile content

Apparently we need something **stable** AND **dynamic**
..without becoming exhaustive

Example: eQTL data

Probe (annotation)

name	mismatch	description
WSU1	true	NA / SpotReport / blast_match_NA / n
WSU2	false	C25A1.8 / cea2.c.00914 / blast_match_
WSU3	false	F21F3.6 / cea2.c.02677 / blast_match_
WSU4	false	F25H2.9 / cea2.c.02801 / blast_match_
WSU5	false	F56H1.4 / cea2.c.04344 / blast_match_

Stable?

Marker (annotation)

name	chromosome	bpstart	cm	description
pkP1050	I	169018	-18.26	PCR_non_cu
pkP1101	I	992189	-17.28	PCR_non_cu
pkP1103	I	1881116	-11.96	PCR_non_cu
pkP1052	I	2818974	-6.1	PCR_non_cu
egPE107	I	3502476	-3.55	PCR_non_cu

Stable?

	pkP1050	pkP1101	pkP1103	pkP1052	egPE107
WSU1	0.5036565	0.4603331	0.3544101	0.9123223	0.4157701
WSU2	0.1365825	0.1037487	0.6600898	0.1241076	0.1672705
WSU3	0.5837218	0.5611695	0.1708836	1.439448	1.94431
WSU4	0.5558796	0.7246171	0.1777933	0.1937225	0.4413371
WSU5	0.3393896	0.4705863	0.224066	0.7713159	0.01334126

Dynamic?

eQTL profiles (data set)

Stable = good for code generation

Annotations: Column-oriented data

name	chromosome	bpstart	cm	description
pkP1050	1	169018	-18.26	PCR_non_cu
pkP1101	1	992189	-17.28	PCR_non_cu
pkP1103	1	1881116	-11.96	PCR_non_cu
pkP1052	1	2818974	-6.1	PCR_non_cu
egPE107	1	3502476	-3.55	PCR_non_cu

← Attributes

3. import

↓ 1. model

```
<entity name="Locus" abstract="true">
  <description> position. Typical examples of such traits are genes, probes and markers. Common structure for entities that have a genomic</description>
  <field name="Chromosome" label="Chromosome" type="xref">
    xref_entity="Chromosome" xref_field="id" xref_label="name" nullable="true"
    description="Reference to the chromosome this position belongs to." />
  <field name="cm" label="cMPosition" type="decimal" nullable="true">
    description="genetic map position in centi morgan (cM)." />
  <field name="bpStart" label="Start (5')" type="long" nullable="true">
    description="numeric basepair postion (5') on the chromosome" />
  <field name="bpEnd" label="End" type="long" nullable="true">
    description="numeric basepair postion (3') on the chromosome" />
  <field name="Seq" type="text" nullable="true">
    description="The FASTA text representation of the sequence." />
  <field name="Symbol" type="varchar" nullable="true">
    description="todo" />
</entity>
<entity name="Chromosome" extends="ObservableFeature">
  <field name="orderNr" type="int" />
  <field name="isAutosomal" type="bool" description="Is 'yes' when number of chr" />
  <field name="bpLength" type="int" nullable="true" description="Length of the chromosome" />
  <field name="Species" label="Species" type="xref" xref_entity="Species">
    xref_field="id" xref_label="name" nullable="true"
    description="Reference to the species this chromosome belongs to." />
</entity>
```

2. generate



Strains (panels)							Chromosomes	Markers	Genes	Measurements	DerivedTraits	Probes	Samples	
							File	Edit	View	1 - 10 of 68,452				
Micro-array probes blasted against WS220 and linked to the genes, used for gene expression phenotypes and eQTLs.														
1.	105050	WSU1	NA / blast_match_NA / non_cumu_bp_start_0											
2.	105051	WSU2	C25A1.8 / cea2.c.00914 / blast_match_50 / non_cumu_bp_start_10184580											
3.	105052	WSU3	F21F3.6 / cea2.c.02677 / blast_match_60 / non_cumu_bp_start_4912043											
4.	105053	WSU4	F25H2.9 / cea2.c.02801 / blast_match_60 / non_cumu_bp_start_10567120											
5.	105054	WSU5	F56H1.4 / cea2.c.04344 / blast_match_60 / non_cumu_bp_start_5741975											
6.	105055	WSU6	H06O01.1 / cea2.c.04508 / blast_match_60 / non_cumu_bp_start_2015820											



Challenge: Modeling 'dynamic' data sets

Can we use 'entity-attribute-value' modeling?

	BF221L	FD85C	C6L9	T7M24	BF151L
RIL1	A	A	A	B	B
RIL2	B	A	B	B	A
RIL3	B	B	A	B	A
RIL4	A	B	B	A	B
RIL5	A	B	A	A	B

← *Not stable columns!
Different per data set...*

↓ 1. model

*Values all comparable:
Columns all of the same type*

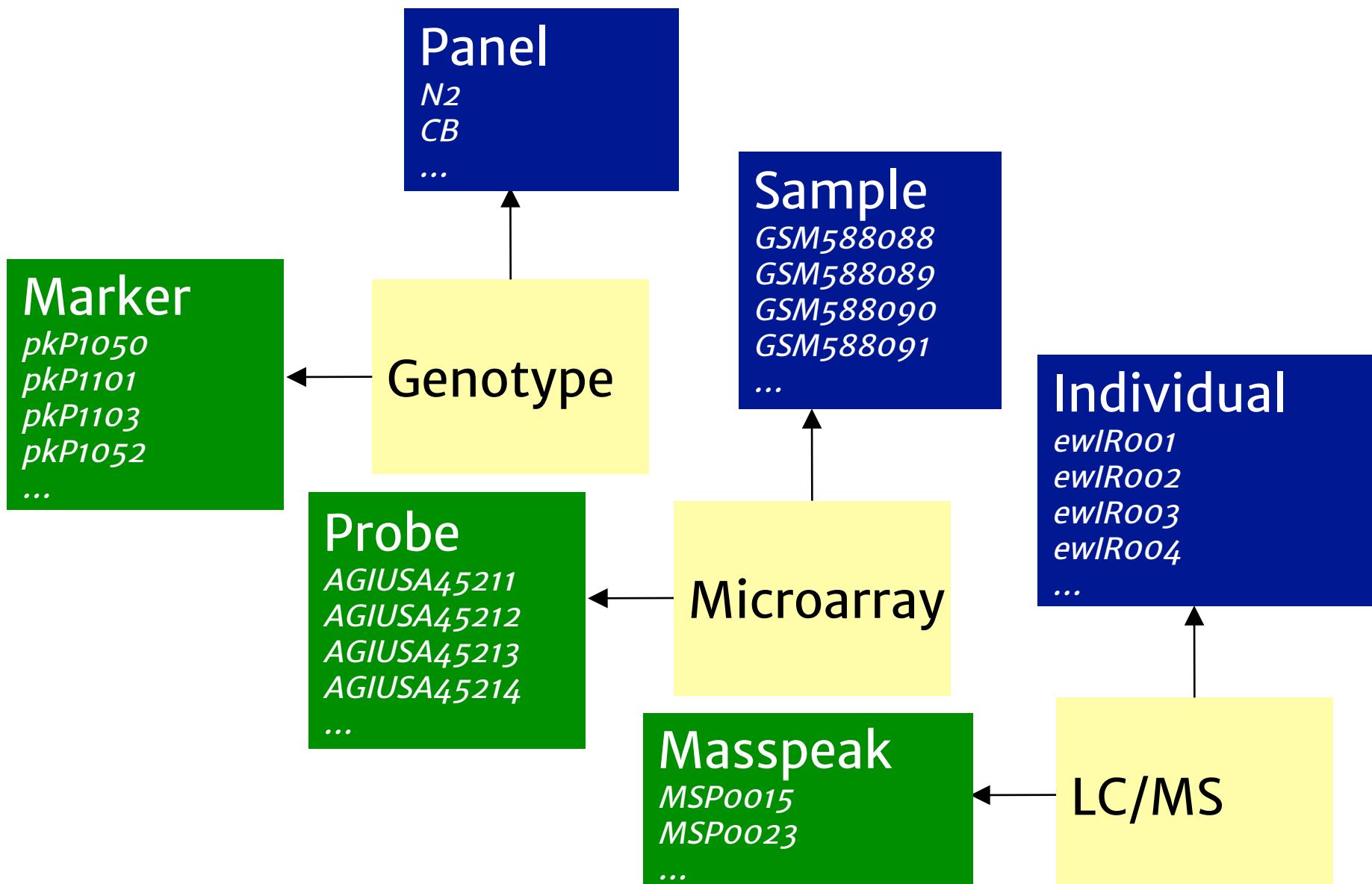
```
<entity name="Locus" abstract="true">
  <description> position. Typical examples of such traits are genes,
  probes and markers. Common structure for entities that have a
  genomic</description>
  <field name="Chromosome" label="Chromosome" type="xref"
    xref_entity="Chromosome" xref_field="id" xref_label="name" nullable="true"
    description="Reference to the chromosome this
    position belongs to." />
  <field name="cm" label="cMPosition" type="decimal" nullable="true"
    description="genetic map position in centi morgan (cM)." />
  <field name="bpStart" label="Start (5')" type="long" nullable="true"
    description="numeric basepair postion (5') on the chromosome" />
  <field name="bpEnd" label="End" type="long" nullable="true"
    description="numeric basepair postion (3') on the chromosome" />
  <field name="Seq" type="text" nullable="true"
    description="The FASTA text representation of the sequence." />
  <field name="Symbol" type="varchar" nullable="true"
    description="todo" />
</entity>
<entity name="Chromosome" extends="ObservableFeature">
  <field name="orderNr" type="int" />
  <field name="isAutosomal" type="bool" description="Is 'yes' when number of chr
  <field name="bpLength" type="int" nullable="true" description="Lenght of the
  <field name="Species" label="Species" type="xref" xref_entity="Species"
    xref_field="id" xref_label="name" nullable="true"
    description="Reference to the species this
    chromosome belongs to." />
</entity>
```



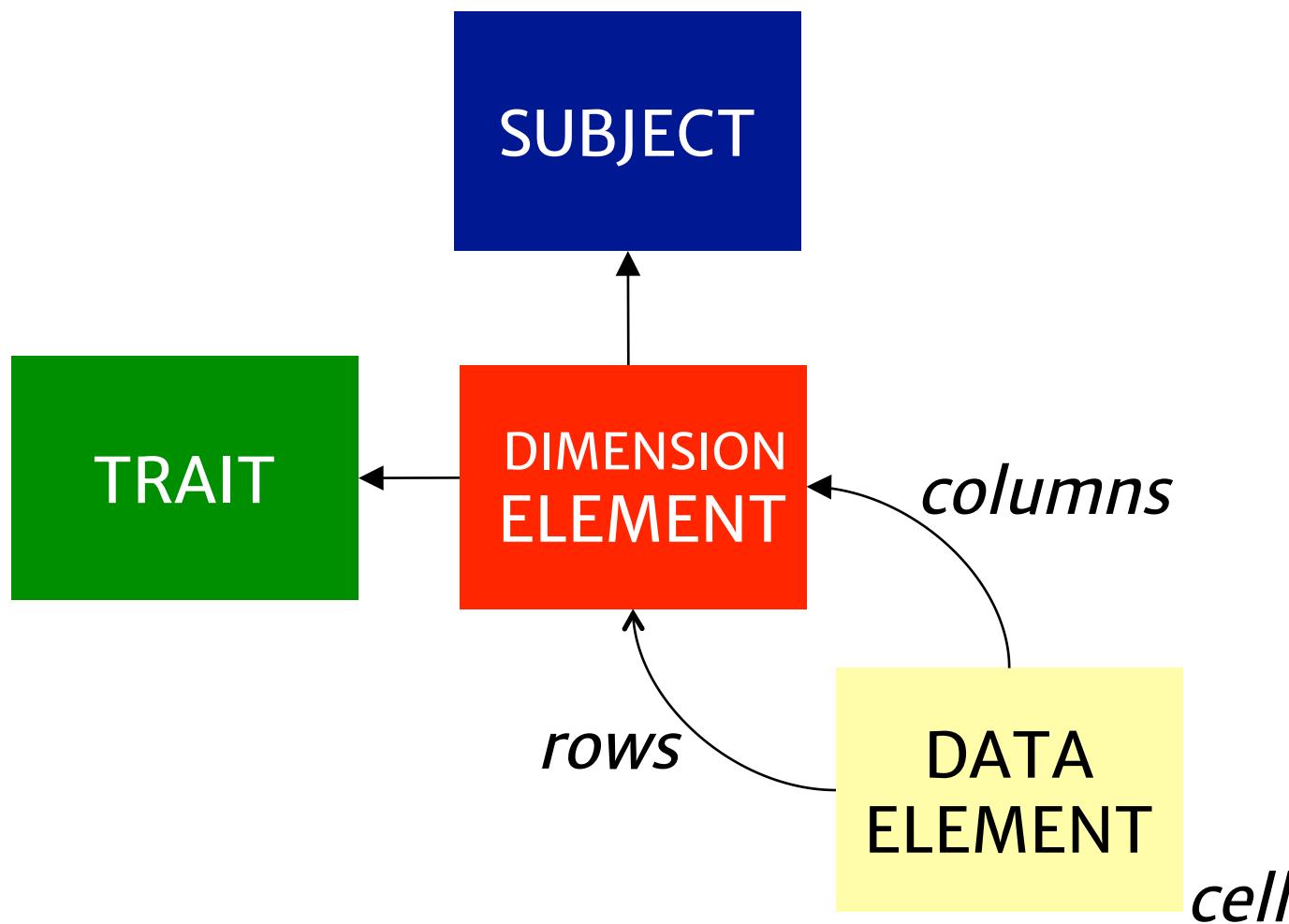
Results

XGAP model

Challenge: Data sets can be variable combinations

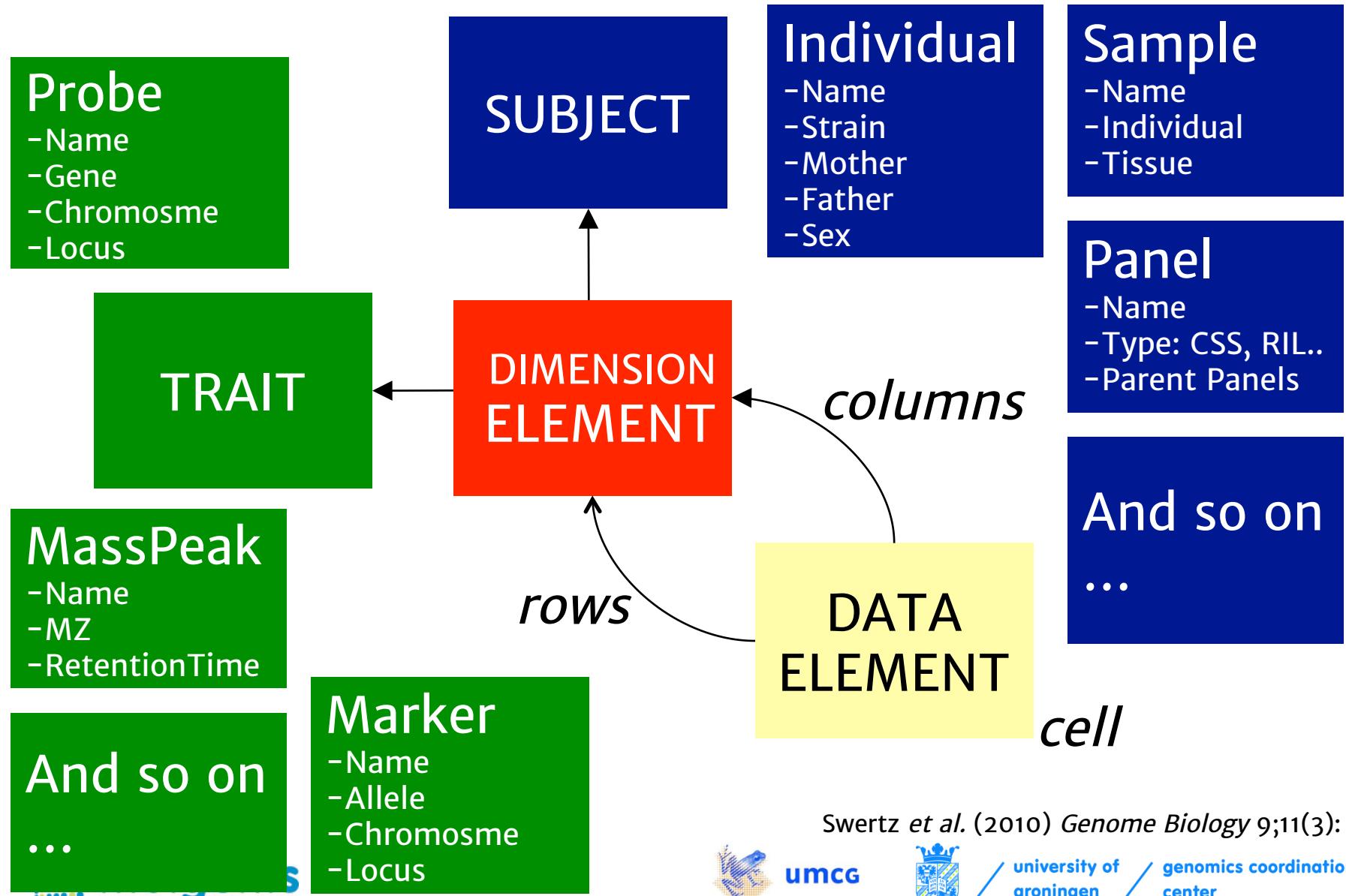


XGAP model: <any trait> X <any subject>



Swertz *et al.* (2010) *Genome Biology* 9;11(3): R27.

Extensible core model



Swertz *et al.* (2010) *Genome Biology* 9;11(3): R27.



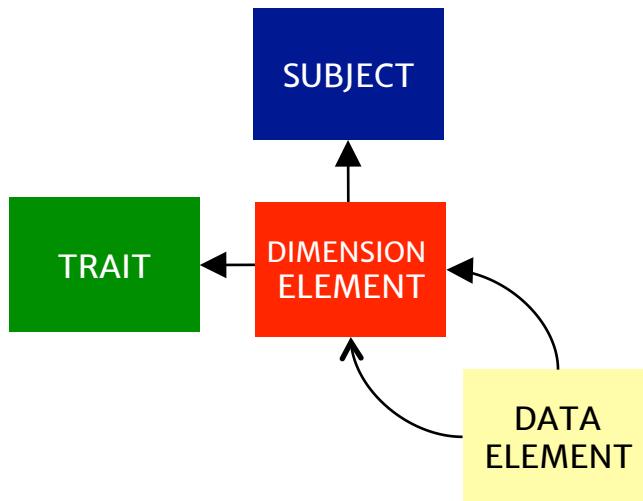
umcg



university of
groningen

genomics coordination
center

Using the XGAP model



1. model

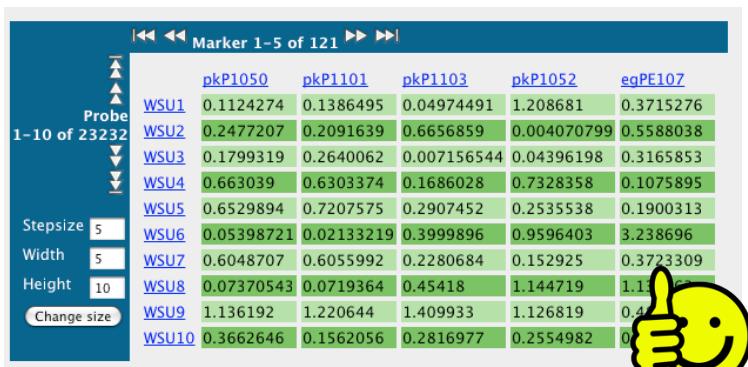
```
<field name="Feature" type="xref" xref_entity="ObservationElement"
      xref_field="id" xref_label="name"
      description="References the ObservableFeature that this observation was made on or observed."/>
<field name="Target" type="xref" xref_entity="ObservationElement"
      xref_field="id" xref_label="name"
      description="References the ObservationTarget that this feature was made on or observed."/>
</entity>
<entity name="ObservedValue" implements="Observation">
<description>
  Generic storage of values, relationships and optional ontology mapping of the value/relationship. Values can be atomic observations, e.g., length (feature) of individual 1 (target) = 179cm (value). Values can also be relationship values, e.g., extract (feature) of sample 1 (target) = derived sample (relation).
<br />
  Discussion: how to model sample pooling in this model?
<br />
  More Discussion: do we want to have type specific subclasses? No, because you can solve this by casting during querying?
</description>
<field name="ontologyReference" type="xref" xref_entity="OntologyTerm"
      nullable="true"
      description="(Optional) Reference to the ontology definition or 'code' for this value (recommended for non-numeric values such as codes)" />
```

4. import

	BPFZIL	PUSOL	ULY	I/M4	DPFOL
RIL1	A	A	B	B	B
RIL2	B	A	B	B	A
RIL3	B	B	A	B	A
RIL4	A	B	B	A	B
RIL5	A	B	A	A	B

2. generate

3. program
viewer plugin



Outcome: working application

Stable!

This screenshot shows the 'Probes' section of the WormQLT application. The top navigation bar includes links for Strains (panels), Chromosomes, Markers, Genes, Measurements, DerivedTraits, Probes, and Samples. A search bar and a page indicator (1 - 10 of 68,452) are also present. The main content area displays a table of micro-array probes, each with an ID, name, description, investigation ontology, reference, alternative identifiers, label, chromosome, and cM position. The first row is highlighted with a blue background. An orange box highlights the first row, and an orange arrow points from this box to the corresponding row in the 'Marker 1-5 of 121' table below.

1.	10505	WSU1	NA / SpotReport / last_match_NA / on_cumu_bp_start_0	Public				
2.	10505	WSU2	25A1.8 / cea2.c.00914 / last_match_60 / on_cumu_bp_start_10184580	Public		clec-87	I	
3.	10505	WSU3	21F3.6 / cea2.c.02677 / last_match_60 / on_cumu_bp_start_4912043	Public		F21F3.6	I	
4.	10505	WSU4	25H2.9 / cea2.c.02801 / last_match_60 / on_cumu_bp_start_10567120	Public		pas-5	I	
5.	10505	WSU5	56H1.4 / cea2.c.04344 / last_match_60 / on_cumu_bp_start_5741975	Public		PP-5	I	
6.	10505	WSU6	06O01.1 / cea2.c.04508 / last_match_60 / on_cumu_bp_start_7015970	Public		pdi-3	I	
7.	10505	WSU7	20F10.2 / cea2.c.06048 / last_match_60 / on_cumu_bp_start_10300315	Public		T20F10.2	I	

Stable!

This screenshot shows the 'Markers' section of the WormQLT application. The top navigation bar includes links for Strains (panels), Chromosomes, Markers, Genes, Measurements, DerivedTraits, Probes, and Samples. A page indicator (1 - 10 of 1,579) is shown. The main content area displays a table of genetic markers, each with an ID, name, description, investigation ontology, reference, alternative identifiers, label, chromosome, and cM position. The first seven rows are highlighted with a blue background. An orange box highlights the first row, and an orange arrow points from this box to the corresponding row in the 'Marker 1-5 of 121' table below.

1.	10347	pkP1050	CR_non_cumu_bp_pos_169018	Public				I -18
2.	10347	pkP1101	CR_non_cumu_bp_pos_992189	Public				I -17
3.	10347	pkP1103	CR_non_cumu_bp_pos_1881116	Public				I -11
4.	10347	pkP1052	CR_non_cumu_bp_pos_2818974	Public				I -6
5.	10347	egPE107	CR_non_cumu_bp_pos_3502476	Public				I -3
6.	10347	egPF101	CR_non_cumu_bp_pos_4338254	Public				I -1
7.	10347	pkP1054	CR_non_cumu_bp_pos_4845516	Public				I

Dynamic!

This screenshot shows a detailed view of a specific marker, 'Marker 1-5 of 121'. The top navigation bar includes links for Strains (panels), Chromosomes, Markers, Genes, Measurements, DerivedTraits, Probes, and Samples. A page indicator (1 - 10 of 23232) is shown. The main content area features a heatmap where rows represent different strains (WSU1 to WSU10) and columns represent the marker's values across various conditions. The heatmap uses a color scale from green (negative values) to red (positive values). On the left, there are controls for 'Step size' (5), 'Width' (5), and 'Height' (10), with a 'Change size' button. The top of the heatmap shows the probe names: pkP1050, pkP1101, pkP1103, pkP1052, and egPE107.

Marker 1-5 of 121					
	pkP1050	pkP1101	pkP1103	pkP1052	egPE107
WSU1	-0.1892	-0.1892	0.231	-0.8379	-0.9186
WSU2	-0.012	-0.012	0.1026	-0.2283	-0.4022
WSU3	0.0637	0.0637	0.2153	-0.1182	-0.1068
WSU4	0.0136	0.0136	0.1208	-0.1924	-0.1909
WSU5	0.054	0.054	0.1649	-0.1768	-0.1621
WSU6	0.0873	0.0873	0.1452	-0.0902	-0.0322
WSU7	-0.0529	-0.0529	-0.0248	0.0354	0.0405
WSU8	0.0629	0.0629	0.1488	-0.06	-0.1015
WSU9	0.0421	0.0421	0.0923	-0.254	-0.0614
WSU10	-0.0099	-0.0099	0.056	-0.0829	-0.0386

Results

Observ-OM model

So far..

- MOLGENIS toolbox
 - Good at rapidly generating databases for **stable** data
 - Values comparable within columns
- XGAP model
 - Good at **flexible homogeneous** data sets
 - All values in data set are semantically equivalent
- *But how about..*
 - Capture **flexible heterogenous** data
 - Detailed meta-data (value type, unit, ontologies..)
 - Data lineage by protocol & application thereof

Repeat

Can we use 'entity-attribute-value' modeling?

Individual	Gender	Age	Height	Body Weight	Diastolic_Blood_Pressure	Systolic_Blood_Pressure	Cervical_Carcinoma
Individual_1	female	56	180	65	90	140	yes
Individual_2	female	45	178	75	87	130	no
Individual_3	male	65	168	100	78	125	no
Individual_4	male	35	178	45	100	150	no
Individual_5	male	34	190	55	68	134	yes
Individual_6	female	24	185	67	76	134	yes
Individual_7	female	20	179	80	102	145	yes

The columns
should be
independent
(more flexible)

↓ 1. model

Values NOT all comparable:
Columns are NOT of the same type

```
<entity name="Locus" abstract="true">
  <description> position. Typical examples of such traits are genes,
  probes and markers. Common structure for entities that have a
  genomic</description>
  <field name="Chromosome" label="Chromosome" type="xref"
    xref_entity="Chromosome" xref_field="id" xref_label="name" nullable="true"
    description="Reference to the chromosome this
    position belongs to." />
  <field name="cM" label="cMPosition" type="decimal" nullable="true"
    description="genetic map position in centi morgan (cM)." />
  <field name="bpStart" label="Start (5')" type="long" nullable="true"
    description="numeric basepair postion (5') on the chromosome" />
  <field name="bpEnd" label="End" type="long" nullable="true"
    description="numeric basepair postion (3') on the chromosome" />
  <field name="Seq" type="text" nullable="true"
    description="The FASTA text representation of the sequence." />
  <field name="Symbol" type="varchar" nullable="true"
    description="todo" />
</entity>
<entity name="Chromosome" extends="ObservableFeature">
  <field name="orderNr" type="int" />
  <field name="isAutosomal" type="bool" description="Is 'yes' when number of chr
  <field name="bpLength" type="int" nullable="true" description="Lenght of the chromosome" />
  <field name="Species" label="Species" type="xref" xref_entity="Species"
    xref_field="id" xref_label="name" nullable="true"
    description="Reference to the species this
    chromosome belongs to." />
</entity>
```



Features become columns

- Feature ~= columns, protocols ~= tables

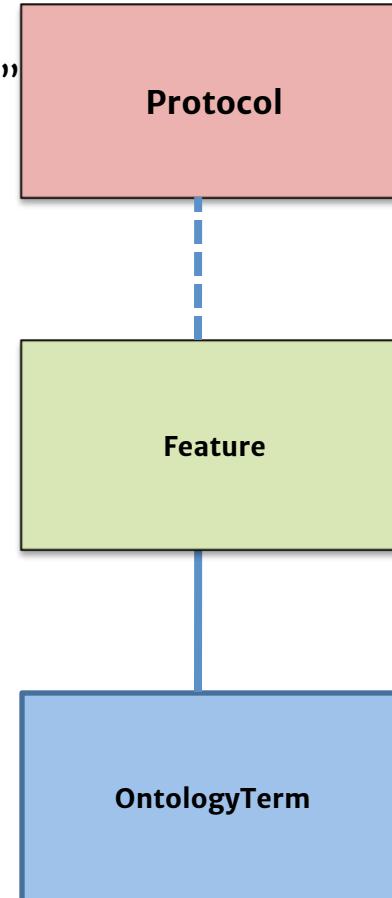
Protocol

Measuring “Gender”, “Age”, “Height”, “Blood pressure”
in some medically standardized way (ISO-54532)

Feature	Feature	Feature	Feature
“Gender”	“Age”	“Height”	“Blood pressure”
Categro.	Years	Centimeter	mmHg
M/F/O	Integer	Decimal	Decimal

Ontology
MeSH term
“Gender Identity”

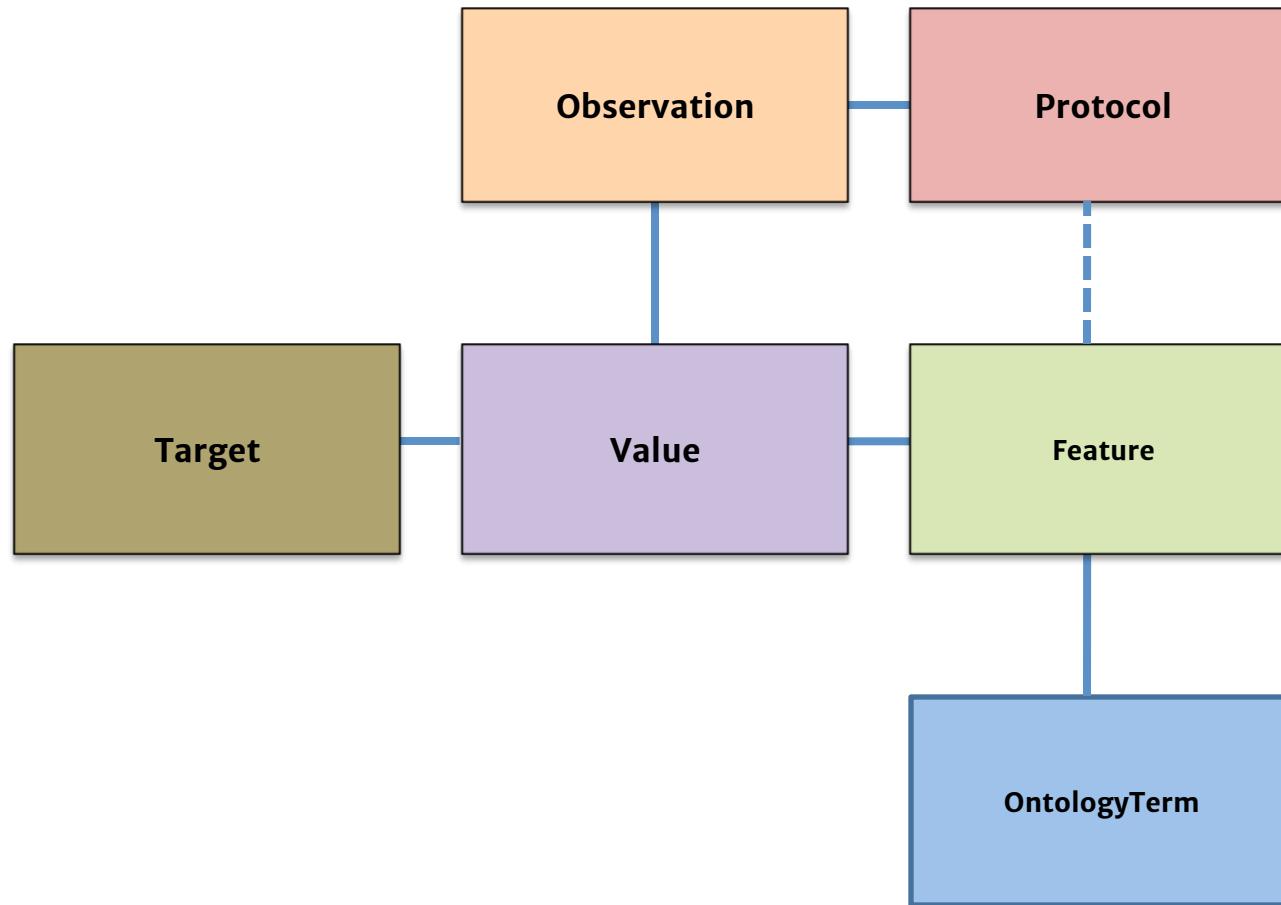
Ontology
MeSH term
“Blood pressure”



Adamusiak *et al.* (2012) *Human Mutation* 33(5):867-73

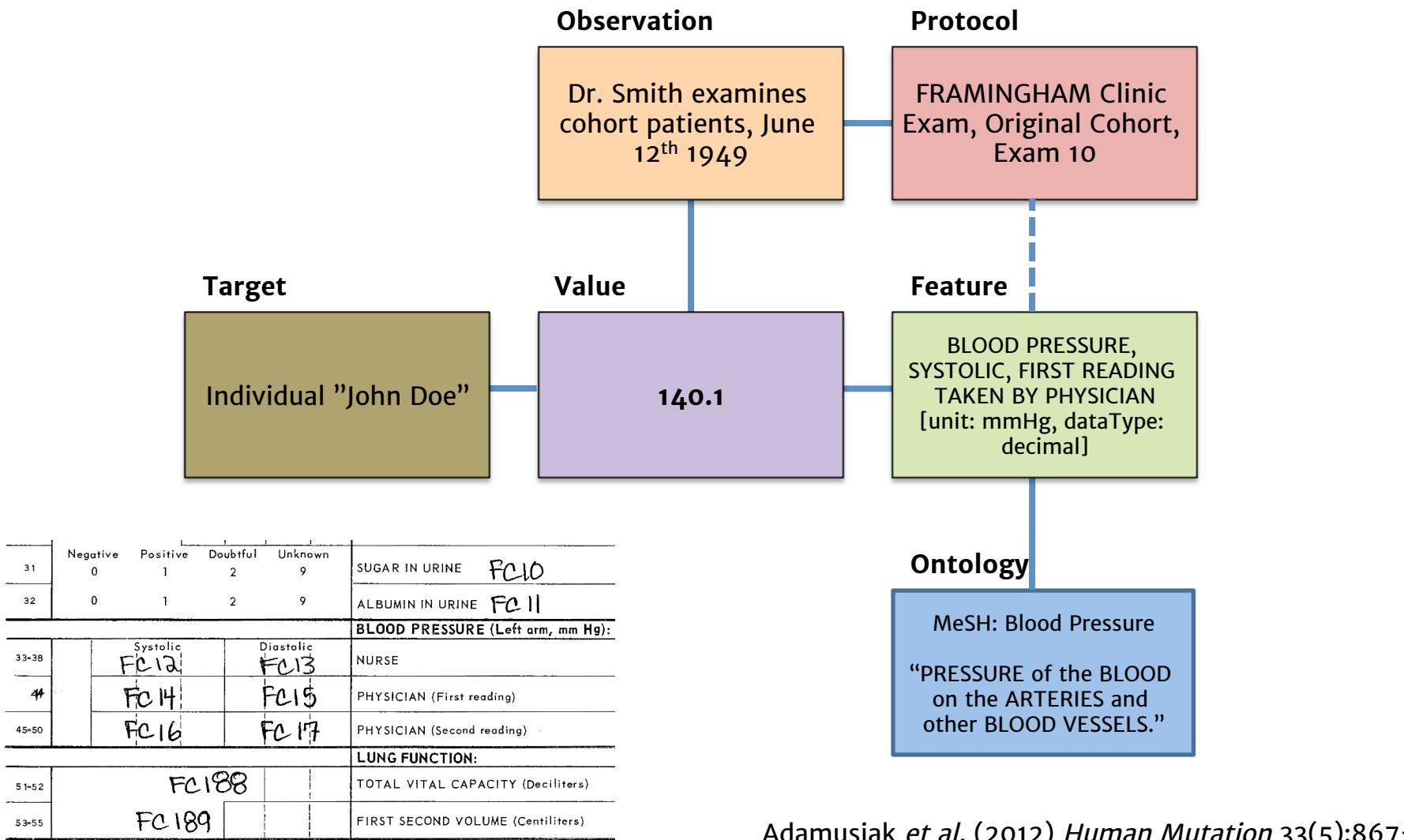
Observations become table rows

- Targets are patients, samples, groups, etc



Adamusiak *et al.* (2012) *Human Mutation* 33(5):867-73

Data example



Adamusiak *et al.* (2012) Human Mutation 33(5):867-73

Individual	Gender	Age	Height	Body_Weight	Diastolic_Blood_Pressure	Systolic_Blood_Pressure	Cervical_Carcinoma	Breast_Carcinoma
Individual_1	female	56	180	65	90	140	yes	no
Individual_2	female	45	178	75	87	130	no	yes
Individual_3	male	65	168	100	78	125	no	yes
Individual_4	male	35	178	45	100	150	no	yes
Individual_5	male	34	190	55	68	134	yes	yes
Individual_6	female	24	185	67	76	134	yes	no
Individual_7	female	20	179	80	102	145	yes	no
Individual_8	female	34	175	56	76	134	yes	yes
Individual_9	female	45	181	67	90	144	yes	yes
Individual_10	male	34	160	45	86	132	yes	yes
Individual_11	male	35	194	75	70	123	yes	yes
Individual_12	female	54	182	55	69	120	yes	yes
Individual_13	male	33	170	66	65	121	yes	yes
Individual_14	female	24	180	77	66	122	yes	yes
Individual_15	female	27	170	56	64	125	yes	yes
Individual_16	female	36	167	54	63	127	yes	yes
Individual_17	male	35	186	76	70	131	yes	yes
Individual_18	male	43	190	88	59	119	yes	yes
Individual_19	male	44	168	65	60	120	yes	yes
Individual_20	female	36	175	54	63	121	yes	yes

Ontologies

MeSH: Blood Pressure

"PRESSURE of the BLOOD on the ARTERIES and other BLOOD VESSELS."

Protocol

Gender ⓘ Age ⓘ Height ⓘ Body Weight ⓘ Diastolic_Blood_Pressure ⓘ Systolic_Blood_Pressure ⓘ Cervical_Carcinoma ⓘ Breast_Carcinoma ⓘ

Observation

Features

Individual ⓘ	Gender ⓘ	Age ⓘ	Height ⓘ	Body Weight ⓘ	Diastolic_Blood_Pressure ⓘ	Systolic_Blood_Pressure ⓘ	Cervical_Carcinoma ⓘ	Breast_Carcinoma ⓘ
Individual_1	female	56	180	65	90	140	yes	no
Individual_2	female	45	178	75	87	130	no	yes
Individual_3	male	65	168	100	78	125	no	yes
Individual_4	male	35	178	45	100	150	no	yes
Individual_5	male	34	190	55	68	134	yes	yes
Individual_6	female	24	185	67	76	134	yes	no
Individual_7	female	20	179	80	102	Values	yes	no
Individual_8	female	34	175	56	76	134	yes	yes
Individual_9	female	45	181	67	90	144	yes	yes
Individual_10	male	34	160	45	86	132	yes	yes
Individual_11	male	35	194	75	70	123	yes	yes
Individual_12	female	54	182	55	69	120	yes	yes
Individual_13	male	33	170	66	65	121	yes	yes
Individual_14	female	24	180	77	66	122	yes	yes
Individual_15	female	27	170	56	64	125	yes	yes
Individual_16	female	36	167	54	63	127	yes	yes
Individual_17	male	35	186	76	70	131	yes	yes
Individual_18	male	43	190	88	59	119	yes	yes
Individual_19	male	44	168	65	60	120	yes	yes
Individual_20	female	36	175	54	63	121	yes	yes

Building G2P applications

MOLGENIS software

Swertz *et al*,
BMC Bioinf. (2010)

<http://www.molgenis.org>



XGAP model

Swertz *et al*,
Genome Biology (2010)

<http://www.xgap.org>



Observ-OM model

Adamusiak *et al*,
Human Mutation (2012)

<http://www.observ-om.org>



EB Registry ◆

Van den Akker *et al*,
Human Mutation (2011)

<http://www.deb-central.org>



AnimalDB

Track and trace of animal life
events in research laboratories
<http://www.animaldb.org>



WormQTL ◆

- Panacea project, C. elegans data
- ~300 million measurements

Snoek, van der Velde, Arends & Li *et al*,
Nucl. Acids Res. (2013)

<http://www.wormqtl.org>



WormQTL



CropQTL: Learning From Nature
xQTL workbench



university of
groningen

genomics coordination
center

CropQTL

Learning From Nature project,
arabidopsis thaliana data

- 1400 plants
- SNP genotypes (~70 million values)
- Classical traits, e.g. flowering time

..more

Showcase EB Registry

EB Registry: Dystr. EB mutation database



International dystrophic eb Patient Registry

International registry of patients with dystrophic epidermolysis bullosa
and database of associated COL7A1 mutations

[Search](#)[Submit data](#)[Contact](#)[References](#)[Background](#)[News](#)[Login](#)[Search](#)

Welcome to the international registry of dystrophic epidermolysis bullosa (DEB) patients and associated COL7A1 mutations.

The International Dystrophic Epidermolysis Bullosa Patient Registry contains anonymised data on both published and unpublished DEB patients, as well as their associated COL7A1 mutations and genotypes, and clinical and molecular phenotypes.

The database currently contains 590 DEB patients, of which 71 unpublished, and 395 COL7A1 mutations. Search or browse below.

Search registry

Search by typing any search term in the search field, like cDNA (e.g. "3G>T") or protein (e.g. "Arg525Ter") notations of mutations, mode of inheritance (e.g. "dominant") or specific phenotypes (e.g. "severe generalized"). Search results are shown at bottom of page.

Enter search term:

[Advanced Search](#)

- Show mutations
 Show patients

[Search](#)[Display all mutations](#)[Display all patients](#)

Browse the COL7A1 gene

Click anywhere on this schematic representation of the COL7A1 gene to graphically browse the gene. With every click you will zoom in deeper on the COL7A1 gene. Mutated nucleotides are depicted in red. If the cursor is placed over the mutated nucleotide(s), the corresponding mutation is shown.

News

Patient and Mutation Update
14 Australasian patients added to Registry.
Jul 20, 2011
[More](#)

Article about DEB Registry online in Human Mutation.
Go to Pubmed to [view article](#)
Jun 16, 2011
[More](#)

Feature Update
Select type of search result tables.
Jun 9, 2011
[More](#)

Name change: International Dystrophic Epidermolysis Bullosa Patient Registry
Important update.
May 25, 2011
[More](#)

Patient and Mutation Update
22 Tunisian RDEB patients entered to database.
May 24, 2011
[More](#)

EB Registry: Dystr. EB mutation database

Phenotypic details for patient 'P10'

Characteristics

Age	20
Gender	m
Ethnicity	unknown
Deceased	yes
Cause of death	
MMP1 allele 1	
MMP1 allele 2	

target

feature

value

Cutaneous

Blistering	yes
Location	generalized
Hands	unknown
Feet	unknown
Arms	unknown
Legs	unknown
Proximal body flexures	unknown
Trunk	unknown
Mucous membranes	yes
Skin atrophy	yes
Milia	unknown
Nail dystrophy	yes

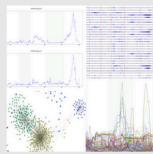
Protocol Application

Showcase WormQTL



WormQLT – Public archive and analysis web portal for natural variation data in *Caenorhabditis* spp.

WormQLT is an online scalable system for QTL exploration to service the worm community. WormQLT provides many publicly available datasets and welcomes submissions from other worm researchers.

[Find QTLs](#)[Genome browser](#)

	sp1P150	sp1P151	sp1P152	sp1P153	sp1P157
NS1	-0.1892	-0.1892	0.2331	-0.8379	-0.9336
NS2	0.0101	0.0101	0.0102	-0.2243	-0.0022
NS3	0.0367	0.0367	0.2333	-0.1132	-0.1668
NS4	0.0101	0.0101	0.0102	0.0102	-0.0009
NS5	0.0255	0.0255	0.0255	-0.0255	-0.0255
NS6	0.0091	0.0091	0.0091	0.0091	-0.0091
NS7	0.0538	0.0538	0.0538	0.0538	0.0538
NS8	0.0202	0.0202	0.0202	0.0202	-0.0005
NS9	0.0421	0.0421	0.0421	-0.234	-0.0034
NS10	0.0098	0.0098	0.0098	-0.0098	-0.0098

[Browse data](#)[Help](#)

What can you do?

- I want to search (e)QTLs for my trait or gene
 - 1. Go to [Find QTLs](#)
 - 2. Type the name or identifier of your trait or gene and press *Search*
 - 3. Put any relevant hits in the shopping cart
 - 4. Click *Plot cart* now and explore the results
- I want to know which genes have a QTL on my favourite position
 - 1. Go to [Genome browser](#)
 - 2. Add tracks from experiments of interest
 - 3. Navigate to your favourite location (tip: use *open in new window*)
 - 4. Collect significant probe identifiers from that region
 - 5. Use the identifiers to e.g. search with [Find QTLs](#)

<i>Phenotypes</i>	<i>Type of array</i>	<i>Sample size</i>	<i>Parental strains</i>	<i>Reference</i>	<i>Pubmed link</i>	<i>Growing temperature</i>	<i>Stage</i>	<i>Food</i>	<i>Medium</i>	<i>Dataset IDs</i>
Gene expression	Washington State University	2x40 RILs	CB4856; N2	Li et al. 2006; Mapping determinants of gene expression plasticity by genetical genomics in <i>C. elegans</i> .	17196041	16oC and 24oC	(72h at 16 and 40h at 24); L4	OP50	NGM Plate	37, 38
Gene expression	Affymatrix tiling array	60 RILs	CB4856; N2	Li et al. 2010; Global genetic robustness of the alternative splicing machinery in <i>Caenorhabditis elegans</i> .	20610403	24oC	(40h) L4	OP50	NGM Plate	n/a
Gene expression	Washington State University	36x3 RILs	CB4856; N2	Vinuela & Snoek et al. 2010; Genome-wide gene expression regulation as a function of genotype and age in <i>C. elegans</i> .	20488933	24oC	(40h, 96h and 214h) L4, Adult, Old	OP50	NGM Plate	3, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21
Gene expression	Agilent 4x44k microarrays	208 RILs	CB4856; N2	Rockman et al. 2010; Selection at linked sites shapes heritable phenotypic variation in <i>C. elegans</i> .	20947766	20oC	YA	OP50	NGM Plate	22, 34, 35, 36
Feeding curves RNAi exposure	n/a	56 RILs * 12 RNAi	CB4856; N2	Elvin & Snoek et al. 2011; A fitness assay for comparing RNAi effects across multiple <i>C. elegans</i> genotypes.	22004469	20oC	Multi-generational	n/a	Liquid S-medium	24, 32, 33
Life-history traits	n/a	80 RILs	CB4856; N2	Gutteling et al. 2007; Mapping phenotypic plasticity and genotype-environment interactions affecting life-history traits in <i>Caenorhabditis elegans</i> .	16955112	12oC and 24oC	Egg, L4, YA	OP50	NGM Plate	25, 26, 27
Lifespan and pharyngeal-pumping	n/a	90 NILs	CB4856; N2	Doroszuk et al. 2009; A genome-wide library of CB4856/N2 introgression lines of <i>Caenorhabditis elegans</i> .	19542186	20oC	All; synchronised	OP50	NGM Plate	4, 23, 28, 29, 30, 31
Lifespan, Recovery and reproduction after heat-shock	n/a	58 RILs	CB4856; N2	Rodriguez et al. 2012; Genetic variation for stress-response hormesis in <i>C. elegans</i> lifespan.	22613270	20oC and 35oC heat-shock	L4 and Adult	OP50	NGM Plate	39, 40
Gene expression	Washington State University	CB4856 and N2	CB4856; N2	Vinuela & Snoek et al. 2012; Aging Uncouples Heritability and Expression-QTL in <i>Caenorhabditis elegans</i> .	22670229	24oC	(40h, 96h and 214h) L4, Adult, Old	OP50	NGM Plate	41, 42, 43

Home Find QTLs Genome browser Browse data Help provide feedback: [comment](#)

Find QTLs

All data (175,366) seam cell

(for example: [measurement](#) (45), [Panel](#) (500), [Gene](#) (47,360), [Transcript](#) (55,782), [Chromosome](#) (8), [Probe](#) (68,452), [Sample](#) (1,630), [DerivedTrait](#) (12))

View

Found n Your results were limited to the first 100. Please be more specific.

Probe [AGIUSA14764 / clc-2](#) reports for [WBGene00000523 - WormBase](#)
C01C10.1 / C01C10.1 / wb|C01C10.1 / non_cumu_bp_start_743339 [...more](#)

Probe [AGIUSA16119 / ceh-1](#) reports for [WBGene00000428 - WormBase](#)
F16H11.4 / F16H11.4 / wb|F16H11.4 / non_cumu_bp_start_465353 [...more](#)

Probe [AGIUSA19594 / acn-1](#) reports for [WBGene00000039 - WormBase](#)
peptidase [C42D8.5.2] / C42D8.5.2 / C42D8.5 / wb|C42D8.5.2|w [...more](#)

Probe [AGIUSA41433 / gsp-1](#) reports for [WBGene0001747 - WormBase](#)
serine/threonine protein phosphatase [F29F11.6.1] / F29F11.6 [...more](#)

Probe [AGIUSA5476 / cul-2](#) reports for [WBGene00000837 - WormBase](#)

Ontological terms
GO:0016021-integral to membrane
GO:0005198-structural molecule activity
GO:0005923-tight junction
WBbt:0005733-hypodermis
WBbt:0005753-seam cell

Ontologies

Ontologies

Ontologies

Home [Find QTLs](#) [Genome browser](#) [Browse data](#) [Help](#) [provide feedback:](#) 

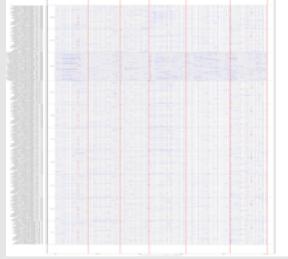
Find QTLs

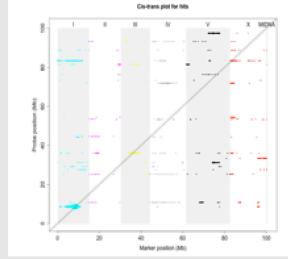
All data (175,366)

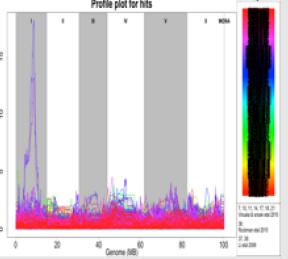
(for example: ctl, daf, pgp-7, gst-27, Y65B4BR, K02B12, WBGene00021562, WBGene00006727, acetylcholine, luciferase ...)
Gene hits, for example on [Geno Ontology](#) or anatomy terms, will show the probes and related terms for that gene.

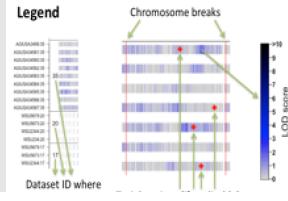
Results for my selected hits:

(get a [permanent link to these results](#))

Heatplot, click to enlarge: 

Cis-trans plot, click to enlarge: 

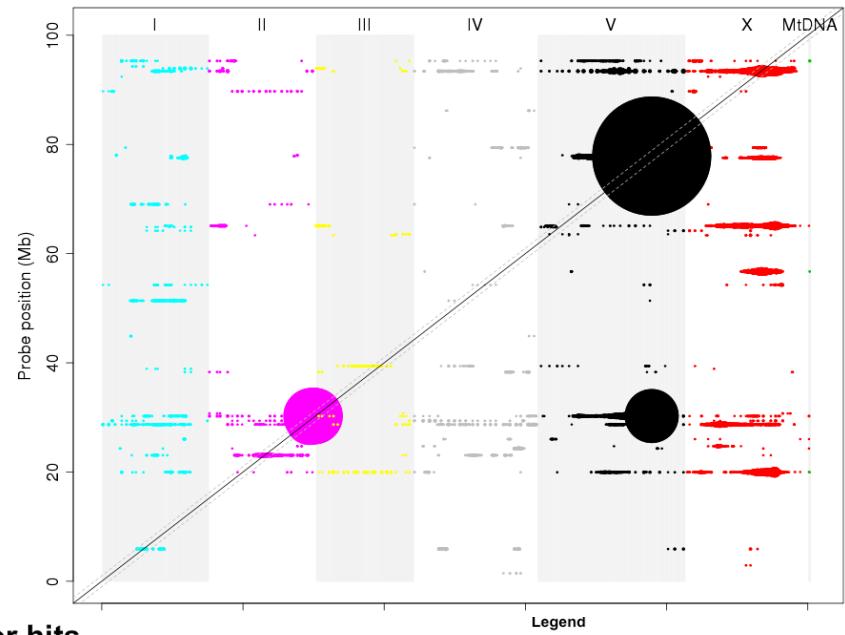
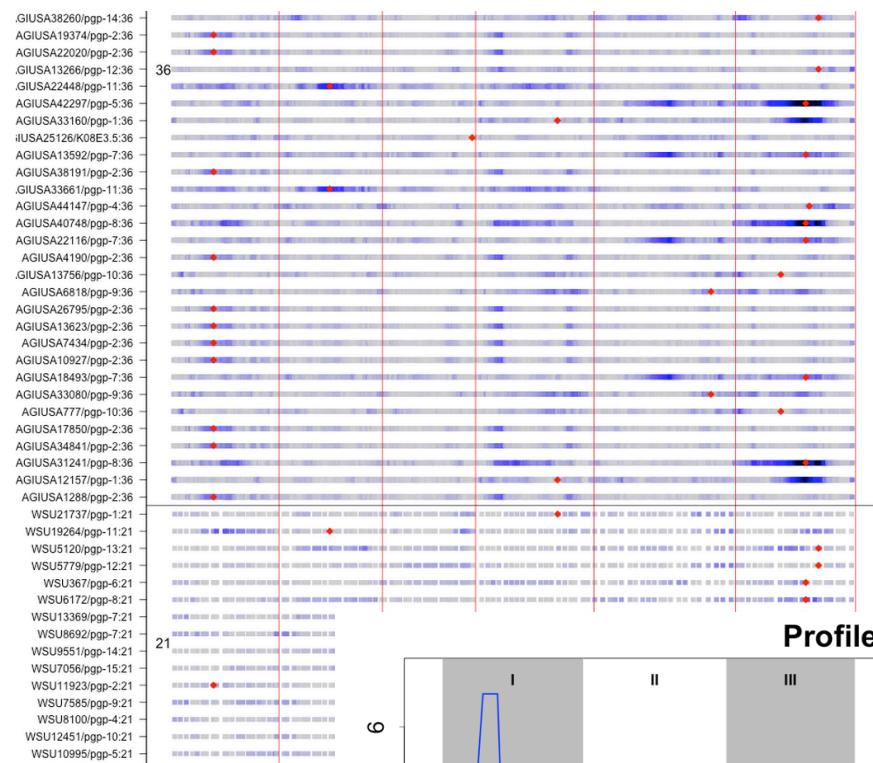
Profile plot, click to enlarge: 

Legend, click to enlarge: 

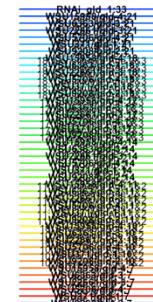
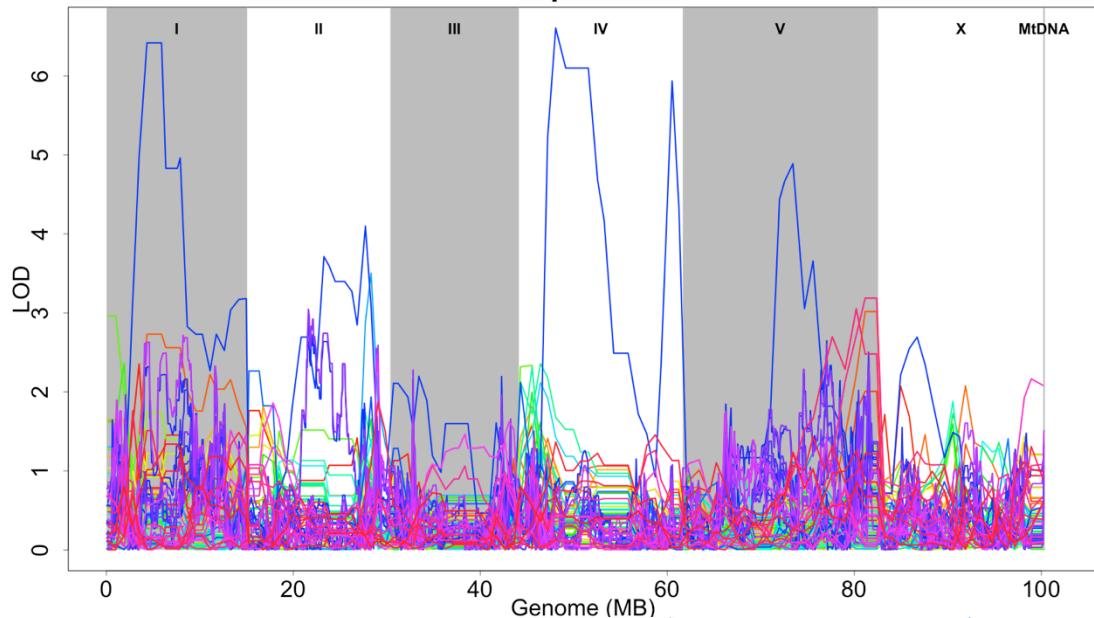
More downloads:

Get the [Cytoscape network](#) for this plot. ([how-to import](#))
Get the [Cytoscape nodes](#) for this plot. ([how-to import](#))
Note: includes **significant results only**. (LOD > 3.5)
Save both files. Import network (has LOD scores), then node attributes (chrom, bploc, dataset). [Example visualization](#)

Get the generated [source data](#) for these plots.
Get the generated [multiplot plot R script](#).
Get the generated [cistrans R plot script](#).
Get the generated [profile R plot script](#).

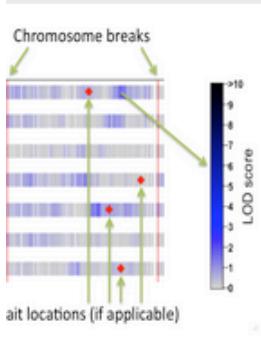


Profile plot for hits





Click to enlarge:



More downloads:

Get the [Cytoscape network](#) for this plot. ([how-to import](#))

Get the [Cytoscape nodes](#) for this plot. ([how-to import](#))

Note: includes **significant results only**. (LOD > 3.5)

Save both files. Import network (has LOD scores), then node attributes (chrom, bploc, dataset). [Example visualization](#)

Get the generated [source data](#) for these plots.

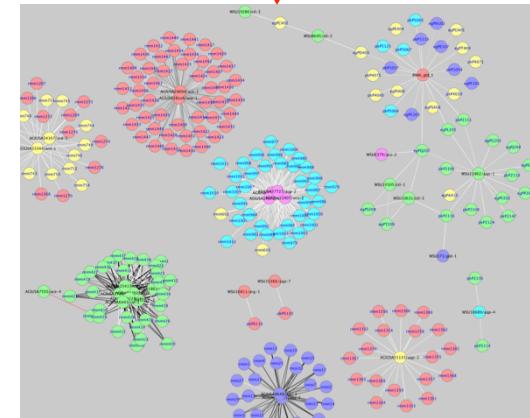
Get the generated [multiplot plot R script](#).

Get the generated [cistrans R plot script](#).

Get the generated [profile R plot script](#).

[/daf-1 \[explore deeper\]](#) – protein kinase [F29C4.1b] / F29C4.1b / F...
[/daf-3 \[explore deeper\]](#) – F25E2.5b.3 / F25E2.5 / wb|F25E2.5b.3|wb|...
[/daf-11 \[explore deeper\]](#) – R0240.3 / cea2 n 107079 / blast match 60

WSU6355/daf-3	trait	X	83377539	age12_int_gtl
pkP1050 marker	I	169018	age12_int_gtl	
pkP1101 marker	I	992189	age12_int_gtl	
pkP1103 marker	I	1881116	age12_int_gtl	
WSU7870/daf-16	trait	I	10763660	age23_int_gtl
egPD403 marker	IV	47198842	age23_int_gtl	
WSU3844/daf-3	trait	X	83371790	age3_gtl
egPK601 marker	X	91884289	age3_gtl	
pkP5071 marker	V	76788020	age3_gtl	
AGIUSA25467/daf-36	trait	V	71855469	rock_gtl
rmm743 marker	IV	52297746	rock_gtl	
rmm744 marker	IV	52364173	rock_gtl	
rmm745 marker	IV	52422112	rock_gtl	
rmm746 marker	IV	52482040	rock_gtl	
rmm747 marker	IV	52569872	rock_gtl	
rmm748 marker	IV	52584390	rock_gtl	
rmm749 marker	IV	52626947	rock_gtl	
rmm750 marker	IV	52705021	rock_gtl	
rmm751 marker	IV	52722889	rock_gtl	
rmm752 marker	IV	52809758	rock_gtl	



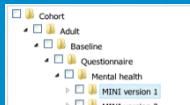
WSU6355/daf-3	trait	X	83377539	age12_int_gtl
pkP1050 marker	I	169018	age12_int_gtl	
pkP1101 marker	I	992189	age12_int_gtl	
pkP1103 marker	I	1881116	age12_int_gtl	
WSU7870/daf-16	trait	I	10763660	age23_int_gtl
egPD403 marker	IV	47198842	age23_int_gtl	
WSU3844/daf-3	trait	X	83371790	age3_gtl
egPK601 marker	X	91884289	age3_gtl	
pkP5071 marker	V	76788020	age3_gtl	
AGIUSA25467/daf-36	trait	V	71855469	rock_gtl
rmm743 marker	IV	52297746	rock_gtl	
rmm744 marker	IV	52364173	rock_gtl	
rmm745 marker	IV	52422112	rock_gtl	
rmm746 marker	IV	52482040	rock_gtl	
rmm747 marker	IV	52569872	rock_gtl	
rmm748 marker	IV	52584390	rock_gtl	
rmm749 marker	IV	52626947	rock_gtl	
rmm750 marker	IV	52705021	rock_gtl	
rmm751 marker	IV	52722889	rock_gtl	
rmm752 marker	IV	52809758	rock_gtl	

Current work

OmicsConnect running on Observ-OMX

Catalogue

Find data item and sample collections



Data

Filter individual data sets and download to Excel & SPSS

	pP1059	pP1101	pP1101	pP1052	pP1107
9502	-0.1892	-0.1892	0.231	-0.8179	-0.186
9502	-0.0192	-0.0192	0.1026	-0.2283	-0.0022
9502	0.0637	0.0637	0.2159	-0.1184	-0.1068
9502	0.0546	0.0546	0.2046	-0.1848	-0.0809
9502	0.054	0.054	0.1649	-0.1768	-0.1621
9502	0.0873	0.0873	0.1452	-0.0902	-0.0322
9502	-0.0529	-0.0529	-0.0248	0.0354	0.0405
9502	0.0608	0.0608	0.1486	0.0706	-0.1018

Compute

Run analysis workflows on big data compute infrastructure



GWAS Central

Explore summary level GWAS data



Protocol

CRFs, Questionnaires, Lab protocols, and assays

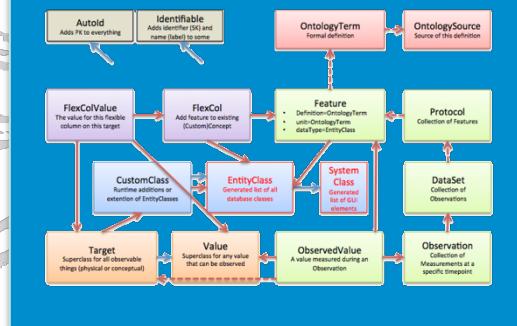


NGS

Next-Generation Sequencing



Core model



xQTL

Multi-omics association & visualization tools



Share

Friends, Groups and Permission management



Mutation

Explore genetic mutations and pathogenicity effects



Organization

Institutes, Departments, People, Locations & Containers



File

File storage and drivers for images and data

L	H	Z	D	G	S	T	M	C	P	A
4-BE	23	D9	R2	HGD3xpl shh						
4-2B	A3	B1	26	aPPaU H21						
E-97	DB	AB	F4	b63=H b64=A						
B-9B	DC	99	G2	*w68 H e						
B-6Z	99	18	EC	o'W'gFOPC						
B-5P	P4	CH	CB	491e1kr l						
B-9	66	65	E6	Ig3 trv H						
B-5F	F9	82	FC	Ko7 hs3-e						

Acknowledgements

Martijn Dijkstra
 Despoina Antonakaki
 Tomasz Adamuziak
 Rob Hastings
 Sirisha Gollapudi
 Guðmundur Thorisson
 Chao Pang
 Myles Byrne
 David van Enckvort
 Linda Mook
 Pieter Neerincx
 Ger Strikwerda
 Danny Arends
 Roan Kanninga
 Jan Bot
 George Byelas
 Yang Li
 Basten Snoek
 Noortje Festen
 Konrad Zych

And more ...

BBMRI • NL



Lude Franke
 Juha Muilu
 Anthony Brookes
 Helen Parkinson
 Vincent Ferreti
 Gert-Jan van Ommen
 Jan Jurjen Uitterdijk
 Ritsert C. Jansen
 Jan Kammenga
 Cisca Wijmenga
 Paul de Bakker
 Irene Nooren
 Rob Hooft
 Salome Scholtens
 Hans Hillege
 Ronald Stolk
 Morris Swertz

And more ...

NBIC/BioAssist consortium (bioinfo)
 BBMRI-NL catalogue group(Hs)
 CTMM/TraIT consortium (Hs)
 EU-GEN2PHEN consortium (Hs)
 EU-PANACEA consortium (Ce)
 EU-BioSHARE consortium (Hs)
 EU-CASIMIR consortium (Mm)
 EU-BioMedBridges consortium (all)
 NL Brassica Nutr. consortium (At)
 Learning from Nature (At)
 LifeLines (Hs)
 TIFN (Hs)
 BigGrid (info)
 Target + CIT (info)

And more...



LIFELINES

TIFOOD NUTRITION

Wrap-up

Summary

- Complex variation, geno-to-pheno
- Exploiting the data requires structure
- Best-of flexible and stable parts
- Support homo- and heterogeneous data

Read more

- MOLGENIS: <http://www.molgenis.org>
- MOLGENIS Compute: <http://www.molgenis.org/wiki/ComputeStart>
- xQTL: <http://www.xqtl.org>
- Adamusiak *et al* (2011) *BMC Bioinformatics*
- Akker *et al* (2011) *Human Mutation*
- Arends *et al* (2010) *Bioinformatics* 26: 2990-2992
- Brandsma *et al*, *Norsk Epidemiologi* 2012
- Snoeks *et al* (2013) *Nucleic Acids Res*
- Swertz *et al* (2010) *Genome Biology* 9;11(3): R27.
- Smedley *et al* (2008) *Briefings in bioinformatics* 9(6):532-44.
- Swertz & Jansen (2007) *Nature Reviews Genetics* 8, 235-243

Thank you!
Questions?

k.j.van.der.velde@umcg.nl

