

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

# Molgenis, life science databases at the push of a button

**Dr. Martijn Dijkstra**

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



umcg

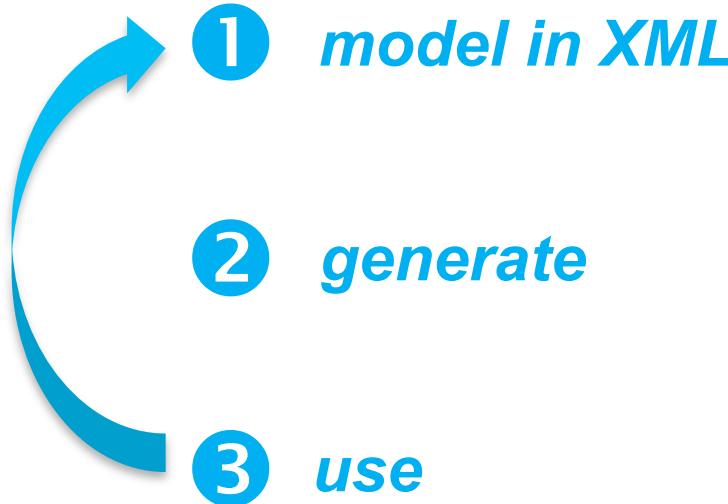


university of  
groningen

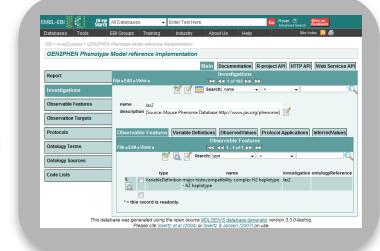
genomics coordination  
center

# Outline

- Molgenis Overview
- Demo Movie
  - model
  - generate
  - use
- Example application
- Analysis of big data
- Summary



<http://...>

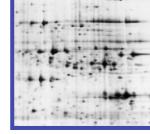
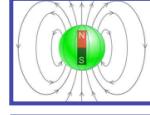
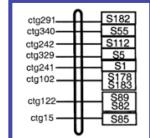
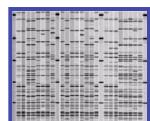
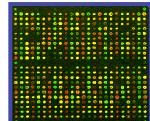


# Overview

# Challenge:



Species...



Experiments...



nbioassist

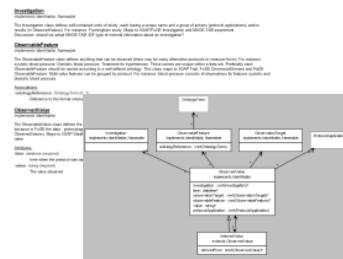


Projects ...

# Needed:

1

## Data models & Protocols



4

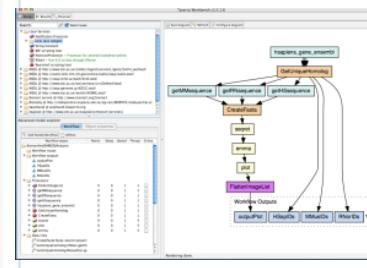
## Connect to statistics



```
find.investigation()  
102 downloaded  
  
obs<-find.observedvalue()  
43,920 downloaded  
  
#some calculation  
add.inferredvalue(res)  
36 added
```

5

## Connect to annotation pipelines



6

## Plugin rich analysis tools



2

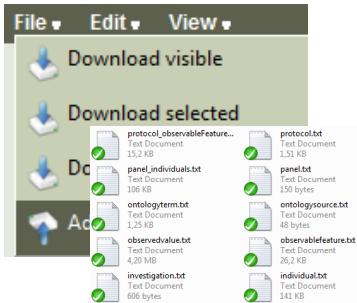
## Edit & trace your data

- Report
- Investigations
- Observable Features
- Observation Targets
- Protocols
- Ontology Terms

A screenshot of the EMBL-EBI ArrayExpress search interface. The top navigation bar includes 'All Databases', 'Enter Text Here', 'Go', 'Reset', 'Advanced Search', 'Give us feedback', 'Site Index', and 'RSS' feed icons. The main content area shows a search result for 'GEN2PHEN Phenotype Model reference implementation'.

3

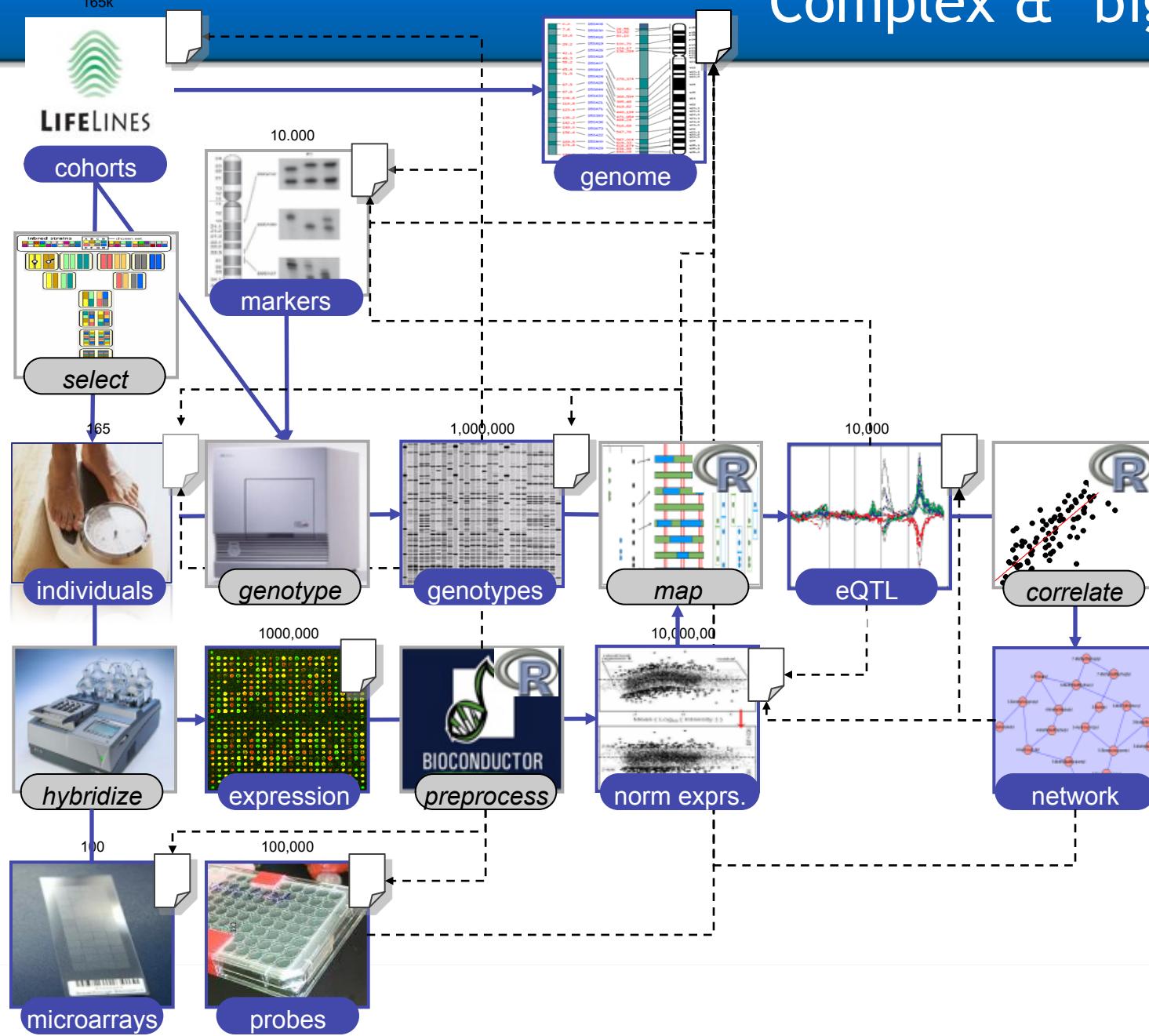
## Import/export



A screenshot of the MOLGENIS database generator interface. The top navigation bar includes 'Main', 'Documentation', 'R-project API', 'HTTP API', and 'Web Services API'. The main content area shows an 'Investigations' view for 'Jax2'. It includes tabs for 'Report', 'Investigations', 'Observable Features', 'Variable Definitions', 'ObservedValues', 'Protocol Applications', and 'InferredValues'. A detailed view of an 'Observable Features' record for 'VariableDefinition major histocompatibility complex H2 haplotype - h2 haplotype' is shown, with fields for 'name', 'description', and 'investigation'. A note at the bottom states: '\* = this record is readonly.'

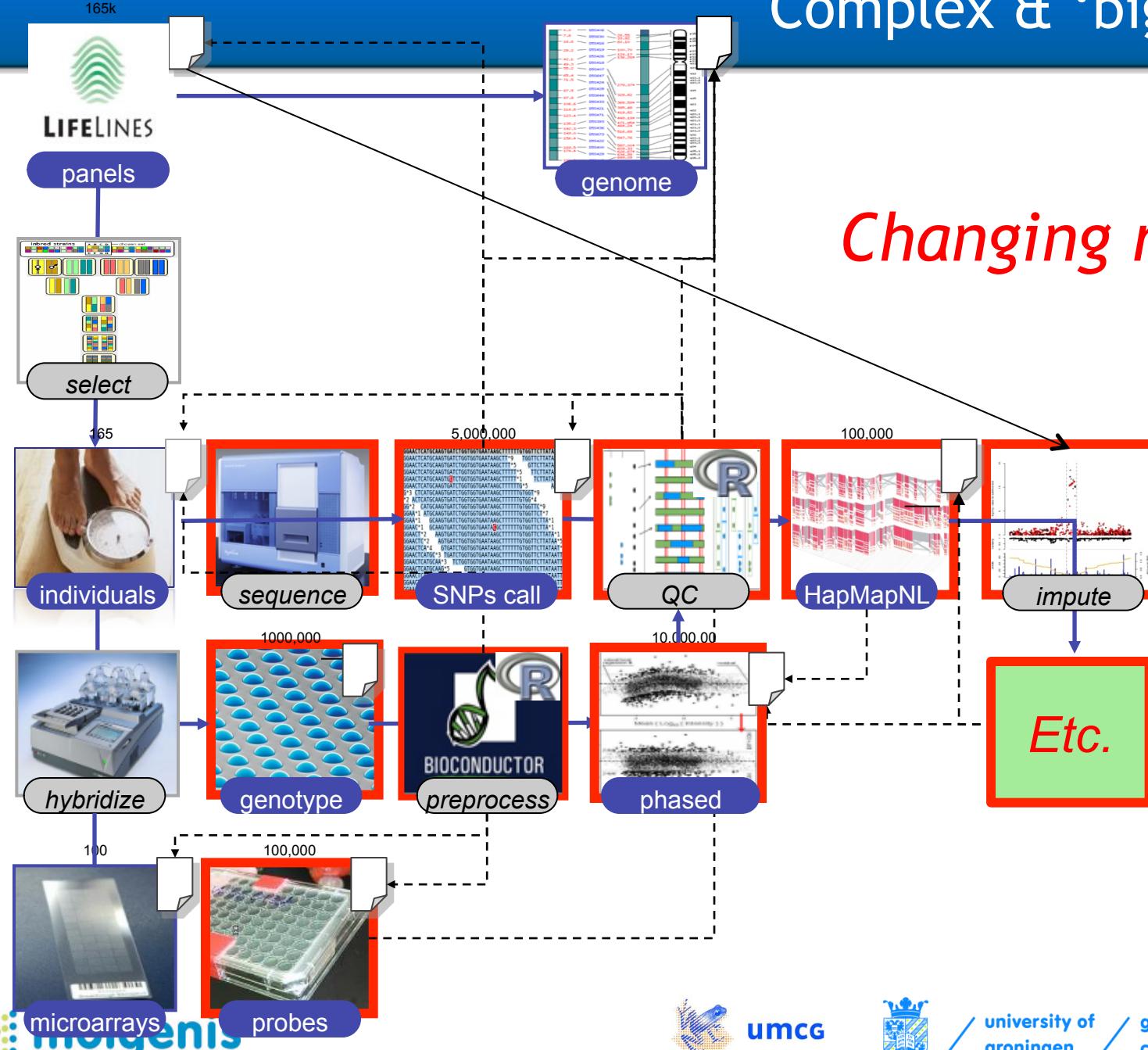
This database was generated using the open source [MOLGENIS database generator](#) version 3.3.0-testing.  
Please cite Swertz et al (2004) or Swertz & Jansen (2007) on use.

# Complex & ‘big data’

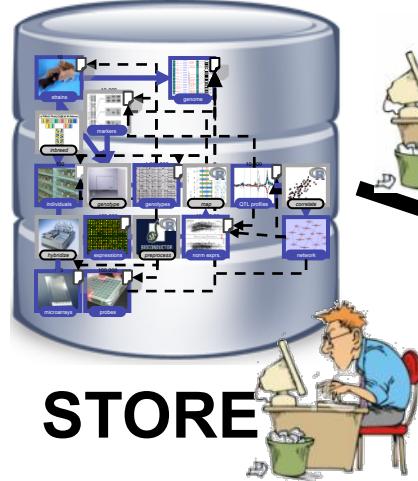


# Complex & ‘big data’

*Changing rapidly*



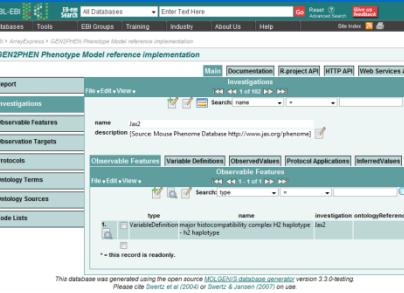
# One change → many places



## Logic

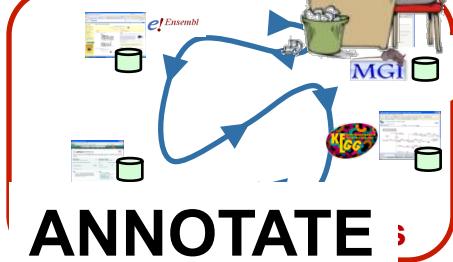
```
static void main(String[] args) throws Exception  
String path = args[0];  
final String expr = args[1];  
  
List l = new ArrayList();  
findfile(new File(path), new F() {  
    public boolean accept(String t) {  
        return t.matches(expr) || isZip(t);  
    }  
});  
  
List r = new ArrayList();  
for (Iterator it = l.iterator(); it.hasNext();)  
    File f = (File) it.next();  
    String fn = f.getName();  
    if (!fn.endsWith(".tar")) r.add(f);  
    if (isZip(fn)) new FileOutputStream(fn).write(  
        findZip(r.getByName(fn)).getBytes());  
        boolean accept(  
        return !match
```

## GUI



biologist

## ANNOTATE



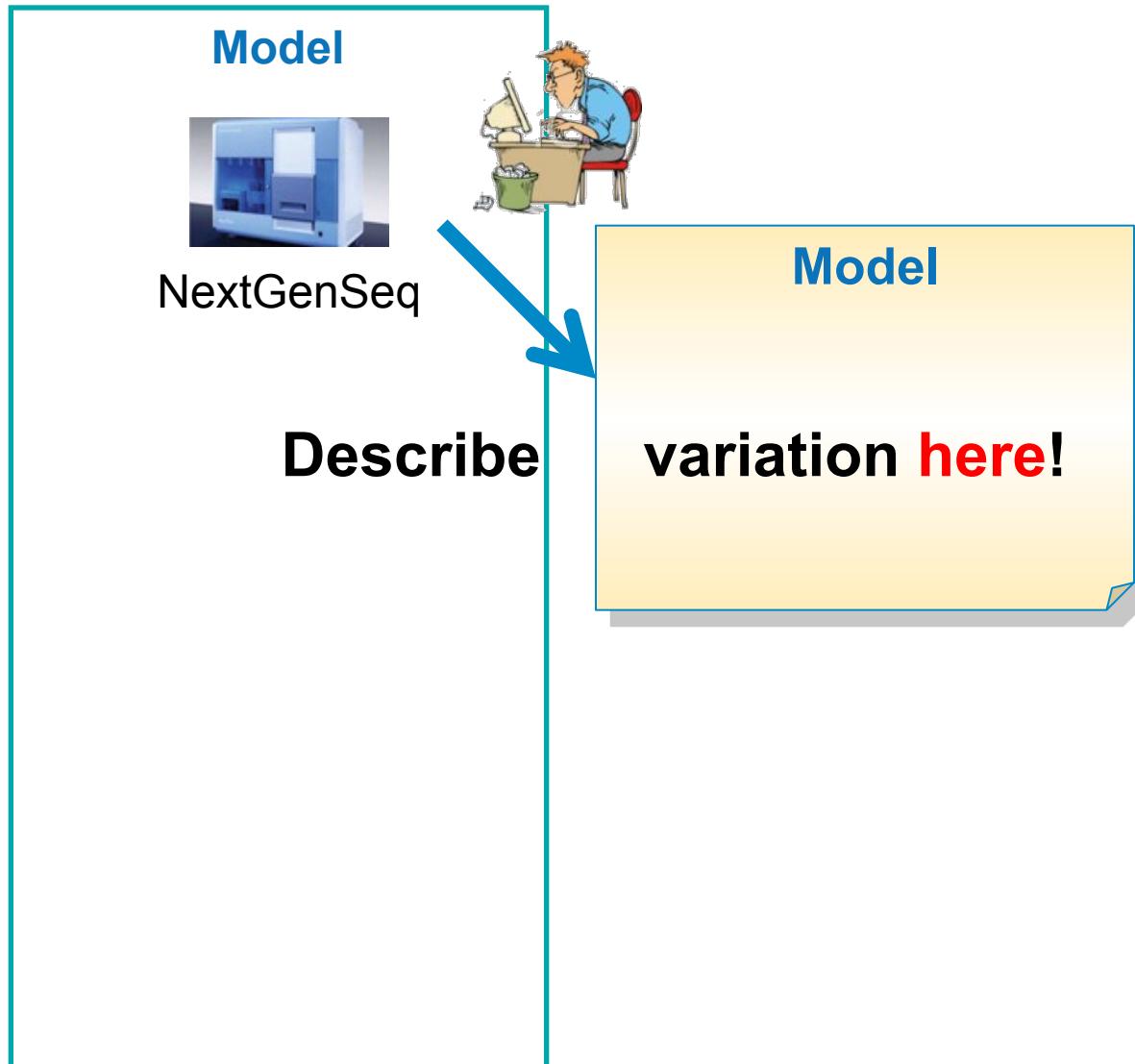
## ANALYSE

Service Requester



bioinformatician

# Model what is desired ...

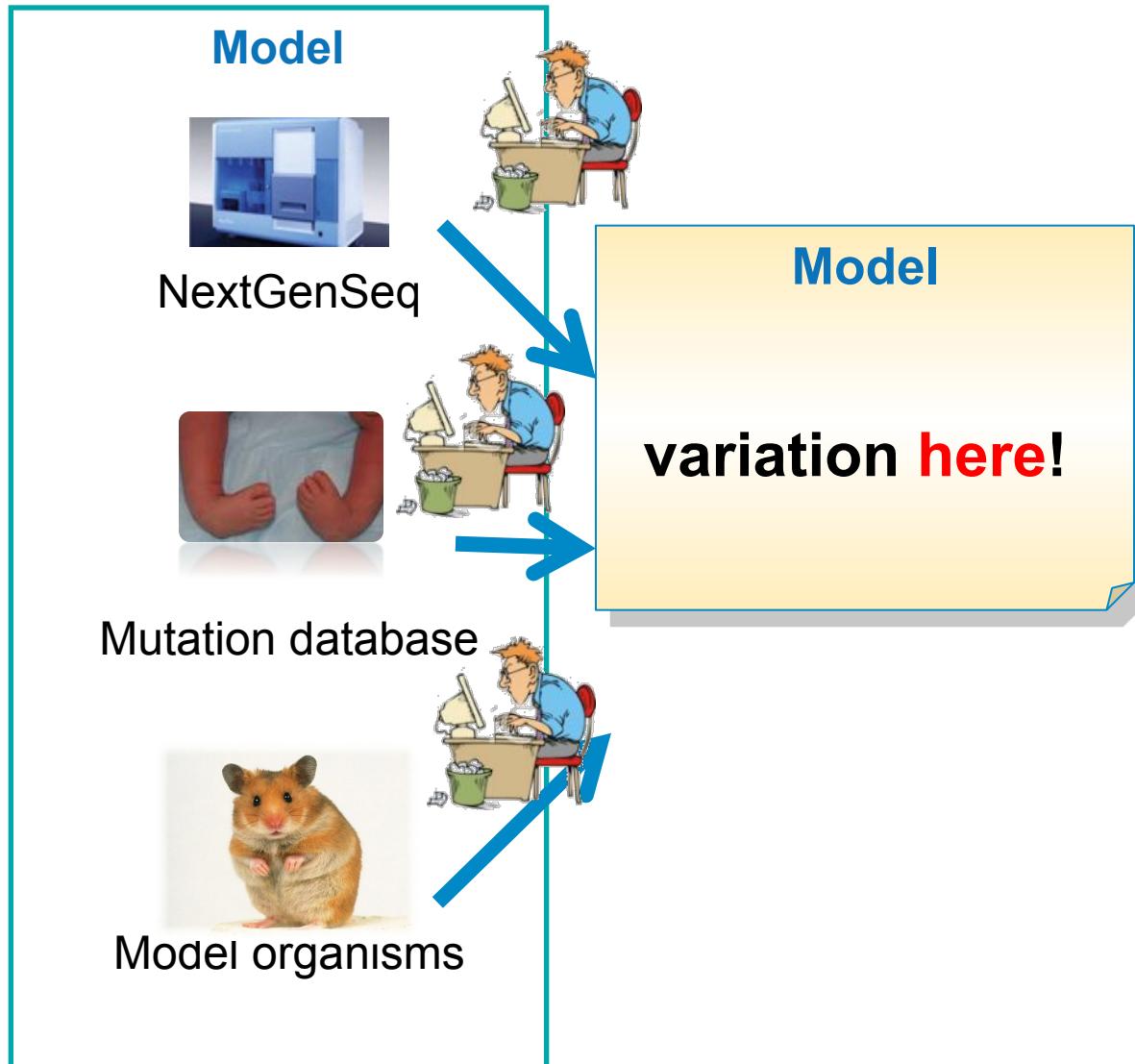


<http://www.molgenis.org>

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

Swertz, Dijkstra, vd Velde et al. (2010) BMC Bioinformatics 11(suppl 12):s12

# Model what is desired ...



<http://www.molgenis.org>

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

Swertz, Dijkstra, vd Velde et al. (2010) BMC Bioinformatics 11(suppl 12):s12

# Autogenerate the software

## Model in DSL



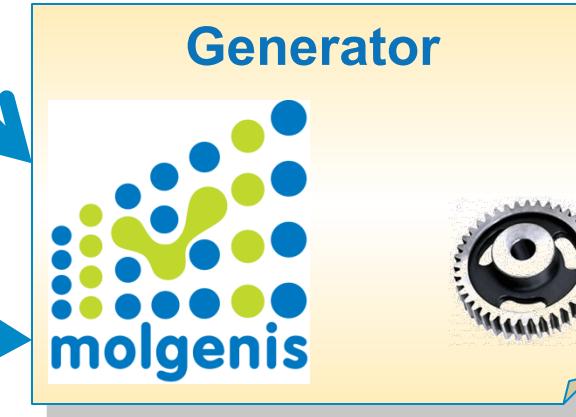
NextGenSeq



Mutation database



Model organisms



## Use generated software

Solexa Sequencer LIMS

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

Sampling  
Flowcell Preps  
Flowcells  
recipes  
sequence\_primers  
Genome Analyzer Runs  
Pipeline Runs  
Admin

Flowcell Preps

flowcells

Add new record

id  
flowcellid  
researcher\_id  
flowcell\_reagent  
reagent  
primer  
primer\_id  
sequence\_primer\_id

database of COL7A1 mutations

SearchDatabase UploadData Contact References Background Help Login

Find a specific mutation/variation  
Variation:   
Nucleotide No:   
Amino Acid No:

Find mutations/variations  
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  
Variant ID around 2000-2510 in tree null at location null. Details: Not set  
See full description at 2007-12-17 Animal Eventlog

Selected animal 5/5  
Type: Rodent  
AnimalID: 5  
Name: 3, 2000  
Status: Alive  
Breeding: Yes  
Species: Mus  
Sex: Female  
Litter: 1  
Location: Harlan Nederland  
Horn: Horn  
The parentGroup DelftPrickles  
Parents:

Apply event to selected animal(s)  
biometry | born | breeder | dies |  
feed | info | remove | sample |  
sexchange | selected | water | wean

Confirm event sample  
MMR: December 12, 2007  
confirm

<http://www.molgenis.org>

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

Swertz, Dijkstra, vd Velde et al. (2010) BMC Bioinformatics 11(suppl 12):s12

# Autogenerate the software

## Model in DSL



NextGenSeq



Mutation database



Model organisms



## Generator



repeat often

## Use generated software

### Solexa Sequencer LIMS

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

database of COL7A1 mutations

### Animal Observatory

<http://www.molgenis.org>

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

Swertz, Dijkstra, vd Velde et al. (2010) BMC Bioinformatics 11(suppl 12):s12

# Reuse the improvements

## Model in DSL



NextGenSeq



Mutation database



Model organisms



## Use generated software

Solexa Sequencer LIMS



Sampling  
Flowcell Preps  
Flowcells  
recipes  
sequence\_primers  
Genome Analyzer Runs  
Pipeline Runs  
Admin

database of COL7A1 mutations



Animal Observatory



<http://www.molgenis.org>

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

Swertz, Dijkstra, vd Velde et al. (2010) BMC Bioinformatics 11(suppl 12):s12

**OPINION**

# Beyond standardization: dynamic software infrastructures for systems biology

*Morris A. Swertz and Ritsert C. Jansen*

Abstract | Progress in systems biology is seriously hindered by the lack of suitable software infrastructures. Biologists need infrastructure that connects to work that is done in other laboratories, for example. Such infrastructure is helpful. However, the infrastructure must also accommodate the complexity of their biological system, but appropriate mechanisms for doing so are currently lacking. We argue that a minimal component of such infrastructure is a software tool called a generator, can be used to quickly generate dynamic software infrastructures that 'systems biologists really want'.



The screenshot shows the MOLGENIS homepage. At the top is a navigation bar with links for News, Documentation, Download, and ChangeLog. Below the navigation is a "ShareThis" button with a count of 41. The main content area features a large blue header "molgenis" with a yellow graphic of overlapping rectangles to its left. Below the header is a sub-header "Welcome to MOLGENIS". A paragraph of text describes MOLGENIS as a collaborative open source project. At the bottom of the page is a footer with social media icons for LinkedIn, Facebook, and Twitter.

MOLGENIS is a collaborative open source project on a mission to generate great software infrastructure for life science research. Each app in the MOLGENIS family comes with rich data management interface and plug-in integration of analysis tools in R, Java and web services.

<http://www.molgenis.org>

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

Swertz, Dijkstra, vd Velde et al. (2010) BMC Bioinformatics 11(suppl 12):s12

# Demo

*<http://www.molgenis.org/wiki/MolgenisDownload>*

# Example application

VOEDINGSDAGBOEK - Gezond

voedingsdagboek.nl

**VOEDINGSDAGBOEK**  
GEZOND SLANK MET JE EIGEN VOEDING

Check je gezondheid    Voeding    Log in    Over ons

Feedback

## Wil jij gezond slank zijn?

Dan is VOEDINGSDAGBOEK er voor jou! Je kunt hier bijhouden wat je eet. Wij geven je daarna inzicht in je voeding. Bovendien geven we je dieettips, afgestemd op je persoonlijke situatie. Zo helpen we je met gezond eten (blijvend!) naar een gezond gewicht. Dit noemen wij "Het nieuwe afvallen".

[Begin meteen »](#)

### Je bepaalt zelf

... wat je eet. Dat is *het nieuwe afvallen*. Houd je helemaal niet van eierkoeken? Wil je niet allerlei speciale ingrediënten in huis halen voor recepten die je eigenlijk helemaal niet lekker vindt? Bij VOEDINGSDAGBOEK hoeft dat ook helemaal niet! Je eet gewoon wat je normaal ook eet en krijgt op basis van dat eten een op maat gesneden advies. Door kleine aanpassingen in je eigen voedingspatroon zul je merken dat je snel gewicht verliest en je gezonder gaat voelen! **Yes, you can!**

### Gezond en blijvend afvallen

Een crash dieet van 1200 calorieën per dag lijkt snel tot resultaat te leiden. Maar schijn bedriegt, want al heel snel komt het jojo-effect om de hoek kijken. Om je streefgewicht te bereiken ?n te behouden is er wat anders nodig, namelijk een gezonde voeding. Een gezonde voeding bevat alle voedingsstoffen die je nodig hebt, maar houdt ook in dat je van niets te veel binnenkrijgt. Zo val je effectief af en houd je dit langdurig vol!

### Wetenschappelijk bewijs

Uit [onderzoek](#) blijkt dat je sneller afvalt als je een VOEDINGSDAGBOEK gebruikt.

Food records/week	Non-AA men (kg)	Non-AA women (kg)	AA men (kg)	AA women (kg)
0	-4.0	-4.0	-4.0	-4.0
2	-5.5	-5.0	-4.5	-4.5
4	-7.0	-6.5	-6.0	-5.5
6	-8.5	-7.5	-7.0	-6.5
8	-10.0	-8.5	-8.0	-7.0

Figure 4. Estimated effect of number of food records kept per week on weight change, by gender and race. <sup>a</sup> Evaluated at

### Aangesloten

- Diëtisten
- Restaurants
- Sportscholen

Wil je gezond uit eten? Afvallen in een sportschool? Of onder begeleiding van een diëtist? De bij ons aangesloten diëtisten, restaurants en sportscholen maken dat mogelijk. Door een database is het invoeren van gegevens erg eenvoudig.

# Voedingsdagboek.nl

- Model / lines of code
  - db.xml / 100 lines
  - ui.xml / 25 lines

```
1<?xml version="1.0" encoding="UTF-8"?>
2<molgenis>
3  <module name="demo">
4    <description>This is a demonstration module</description>
5    <entity name="Experiment">
6      <description>My Experiments</description>
7      <field name="Id" type="autoid" description=" autogenerated unique id" />
8      <field name="Name" unique="true" description="Unique name" />
9      <field name="Description" nullable="true" description="Optional description" />
10     </entity>
11    <entity name="Sample">
12      <description>My Samples</description>
13      <field name="Id" type="autoid" description=" autogenerated unique id" />
14      <field name="Name" unique="true" description="Unique name" />
15      <field name="Experiment" type="xref" xref_field="Experiment.Id" />
16      <xref_label>Name</xref_label>
17      <field name="Species" type="enum" enum_options="[hs,mm]" />
18      <field name="Created" type="date" auto="true" description="Automatic date" />
19    </entity>
20    <entity name="Data">
21      <field name="Id" type="autoid" description=" autogenerated unique id" />
22      <field name="Name" unique="true" description="Unique name" />
23      <field name="Experiment" type="xref" xref_field="Experiment.Id" />
24      <xref_label>Name</xref_label>
25      <field name="Samples" type="xref" xref_field="Sample.Id" xref_label="Name" />
26      <description>One or more samples</description>
27      <field name="Data" type="file" />
28    </entity>
29  </module>
30</molgenis>
```

# Voedingsdagboek.nl

- Model / lines of code
  - db.xml / 100 lines
  - ui.xml / 25 lines
- Generate
  - \*.sql / 1722 lines
  - \*.java / 46639 lines

```
1<?xml version="1.0" encoding="UTF-8"?>
2<molgenis>
3  <module name="demo">
4    <description>This is a demonstration module</description>
5    <entity name="Experiment">
6      <description>My Experiments</description>
7      <field name="Id" type="autoid" description=" autogenerated unique id" />
8      <field name="Name" unique="true" description="Unique name" />
9      <field name="Description" nullable="true" description="Optional descriptio...>
10     </entity>
11   <entity name="Sample">
12     <description>My Samples</description>
13     <field name="Id" type="autoid" description=" autogenerated unique id" />
14     <field name="Name" unique="true" description="Unique name" />
15     <field name="Experiment" type="xref" xref_field="Experiment.Id" xref_label="Name" />
16     <field name="Species" type="enum" enum_options="[hs,mm]" />
17     <field name="Created" type="date" auto="true" description="Automatic date" />
18   </entity>
19 </module>
20 <entity name="Data">
21   <field name="Id" type="autoid" descri...
22   <field name="Name" unique="true" desc...
23   <field name="Experiment" type="xref" ...
24     xref_label="Name" />
25   <field name="Samples" type="mref" xref...
26     description="One or more samples" />
27   <field name="Data" type="file" />
28 </entity>
29 </module>
30</molgenis>
```



<http://...>



# Voedingsdagboek.nl

- Model / lines of code
  - db.xml / 100 lines
  - ui.xml / 25 lines
- Generate
  - \*.sql / 1722 lines
  - \*.java / 46639 lines

```
1<?xml version="1.0" encoding="UTF-8"?>
2<molenis>
3  <module name="demo">
4    <description>This is a demonstration module</description>
5    <entity name="Experiment">
6      <description>My Experiments</description>
7      <field name="Id" type="autoid" description=" autogenerated unique id" />
8      <field name="Name" unique="true" description="Unique name" />
9      <field name="Description" nullable="true" description="Optional descriptio...>
10     </entity>
11   <entity name="Sample">
12     <description>My Samples</description>
13     <field name="Id" type="autoid" description=" autogenerated unique id" />
14     <field name="Name" unique="true" description="Unique name" />
15     <field name="Experiment" type="xref" xref_field="Experiment.Id" xref_label="Name" />
16     <field name="Species" type="enum" enum_options="[hs,mm]" />
17     <field name="Created" type="date" auto="true" description="Automatic date" />
18   </entity>
19 </module>
20 <molenis name="example">
21   <plugin name="molenis_header" type="plugins.header.MolenisHeader" />
22   <menu name="Main">
23     <form name="Experiments" entity="Experiment" sortby="Name">
24       <menu name="ExperimentMenu">
25         <form name="Samples" entity="Sample" />
26         <form name="Data" entity="Data" />
27       </menu>
28     </form>
29     <plugin name="ImportWizard" type="plugins.genericwizard.GenericWizard" />
30   </menu>
31 </molenis>
```



http://...



1 : 400!

# Voedingsdagboek.nl

- Model / lines of code
  - db.xml / 100 lines
  - ui.xml / 25 lines
- Generate
  - \*.sql / 1722 lines
  - \*.java / 46639 lines
- 5 custom plugins / 3682 lines
- Use

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE molgenis>
<module name="demo">
  <description>This is a demonstration module</description>
  <entity name="Experiment">
    <description>My Experiments</description>
    <field name="Id" type="autoid" description=" autogenerated" unique="true" description="Unique name" />
    <field name="Name" unique="true" description=" Unique name" />
    <field name="Description" nullable="true" description="Opmerkingen" />
  </entity>
  <entity name="Sample">
    <description>My Samples</description>
    <field name="Id" type="autoid" description=" autogenerated" unique="true" description="Unique name" />
    <field name="Name" unique="true" description=" Unique name" />
    <field name="Experiment" type="xref" xref_field="Experiment" xref_label="Name" />
    <field name="Species" type="enum" enum_options="[hs,mm]" />
    <field name="Created" type="date" auto="true" description="Geactiveerd op" />
  </entity>
  <entity name="Data">
    <field name="Id" type="autoid" description=" autogenerated" unique="true" desc="Gegevens" />
    <field name="Name" unique="true" desc="Gegevensnaam" />
    <field name="Experiment" type="xref" xref_label="Name" />
    <field name="Samples" type="xref" xref_label="Name" description="One or more samples" />
    <field name="Data" type="file" />
  </entity>
</module>
</molgenis>
```



Gezond slank  
60 kg



# Analysis of big data

# Challenge: big data and huge analyses



- Challenge 1
  - Workflows with dozens of analysis scripts
  - execute each script ~100x for different parameter values
  - 1000 - 10,000 scripts in total per workflow
- Challenge 2
  - Running on high-performance computational infrastructures
    - (Target) computer clusters, national GRID, ...
    - Different scheduling systems: *PBS*, ...
    - Different file management: *SRM*, ...
    - Different tool management: *module system*, ...
- Challenge 3
  - Monitoring execution, log files, data provenance

# Solution: Molgenis Compute - huge analyses straight from db

# Solution: Molgenis Compute - huge analyses straight from db

- Solution 1
  - Input: Few files
    - 1 workflow.csv
    - 1 parameters.csv
    - $n$  scriptTemplates.sh
  - Output: Generate 1000s of scripts
- Solution 2: transparent execution
  - *resource management*
  - *data management*
  - *tool management*
- Solution 3
  - Monitoring execution, log files, data provenance
- **Scientist can focus on the analysis only**



# Closing remarks

# All available as open source

Fork me on GitHub

**github** Search or Type a Command Explore Gist Blog Help mswertz Edit molgenis's Profile

**Repositories** Members Find a Repository... All Public Private Sources Forks Mirrors

**molgenis\_apps** Java ★ 25 ⚡ 27 MOLGENIS Advanced Application and Computation Framework for the Life Sciences Last updated 14 hours ago

**molgenis** Java ★ 27 ⚡ 26 The MOLGENIS Software generator tool for creating Dynamic Software Infrastructure used in the Life Sciences Last updated 17 hours ago

**molgenis\_test** Java ★ 7 ⚡ 9 The MOLGENIS Test suite, fork this if you plan to contribute to MOLGENIS apps, so you can test commits before sending them in Last updated 7 days ago

**molgenis\_distro** JavaScript ★ 5 ⚡ 4 The empty MOLGENIS Distribution, fork this and create your own MOLGENIS web application Last updated a month ago

**biosoftware platform** molgenis

The Netherlands, Europe <http://www.molgenis.org> Joined on Apr 28, 2012

4 public repos 0 private repos 23 members

<http://github.com/molgenis>  
<http://www.molgenis.org/>



umcg

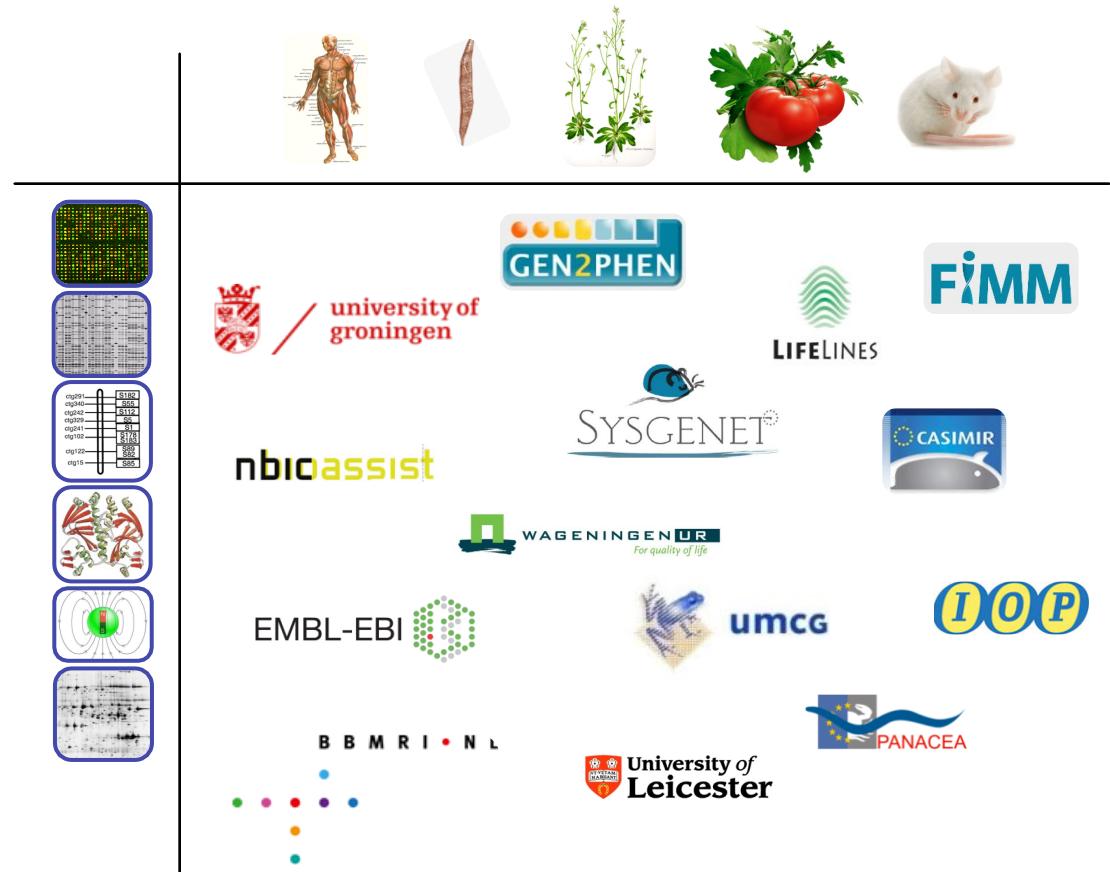


university of  
groningen

genomics coordination  
center

# Capture most variation in Life Sciences in one model

- Molgenis
  - Model
  - Generate
  - Use



# Acknowledgements

Joeri van der Velde  
 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  
 Martijn Dijkstra

*And more ...*

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)

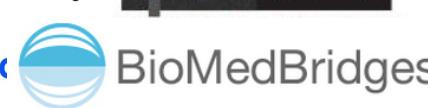
EU-CASIMIR consortium (Mm)  
 EU-BioMedBridges consortium (all)

NL Brassica Nutr. consortium (At)  
 Learning from Nature (At)

LifeLines (Hs)  
 TIFN (Hs)

BigGrid (info)  
 Target + CIT (info)

*And more...*



# Questions?

## Summary

- Dynamic database
- Rich web GUI
- Interfaces for analysis
- Integrated large scale analyses

## 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?

m.dijkstra.work@gmail.com

**molgenis** .org  
*Your database at the push of a button*

# Genomics Coordination Center, UMCG, Groningen

Biobanking

## Diagnostics & clinic

Core facility of ~15 programmers, postdocs and PhD students on a mission to research, develop and support high throughput life sciences with database and analysis e-infrastructures.



Large scale multi-omics

Human and model organism research: genotype 2 phenotype