Genetical Genomics:

High Throughput Modeling, Mapping and Exploration of Big Data

28 september - CBSG Arabidopsis and Brassica

Danny Arends, GBIC – RUG
Joeri v/d Velde, GBIC – RUG & GCC - UMCG
Ronny Joosen, Plant science - WUR
Wilko Luchterink, Plant science - WUR
Pjotr Prins, Nematoid science - WUR
Ritsert Jansen, GBIC - RUG
Karl Broman, Dept of biostatistics - WISC

university of groningen

CBSG 2012
Centre for BioSystems Genomics

Challenge outline

Finding regulators of

* RNA transcription / gene regulation
* Protein abundance / degradation
* Metabolite abundance / flow

Almost complete genetic maps

Sample size <<< Variables

BUT

Tools are available

* QTL analysis
* PCA analysis
* Correlations
* Machine learning

Scattered: java, perl, php, webservices, r, c, etc

Not optimized: Single trait to a handfull

Our approach

Managed data storage
* Molgenis & XGAP

Smarter algorithms
* R/QTL

CPU power / Memory
* Cluster computing

Visualizations

Molgenis

Rapid prototyping:

* Data storage
* User interfaces

Added benefit:

* Connections to mayor analysis platforms
* Plugin system
* Data import and export

- Molecular Genetics Information System (MOLGENIS): alternatives in developing local experimental genomics databases.
- Beyond standardization: dynamic software infrastructures for systems biology

XGAP

Data model for Genetical Genomics

Based on the FUGE and MIAME

Plugins:

* Analysis
* Exploration

	LerKas1	LerKas2	LerKas3	LerKas4	LerKas6	LerKas
10.00000		0.00081	0.00079	0.00111	0.0001	0.0008
9.96000	0.00096	0.00099	0.00087	0.00015	0.00081	0.
9.92000	0.00087	0.00095	0.00085	0.00016	0.0008	0.
9.88000	0.00085	0.00121	0.00103	0.00018	0.00098	0.
9.84000	0.00066	0.00113	0.00106	0.00029	0.00091	0.
9.80000	0.00097	0.00114	0.00104	0.00027	0.0008	0.
9.76000	0.00093	0.00106	0.00087	0.00013	0.0009	0.
9.72000	0.00111	0.00087	0.00112	0.00014	0.00092	0.
9.68000	0.00086	0.00103	0.00091	0.00017	0.00096	0.
9.64000	0.00092	0.00102	0.00093	0.00016	0.00096	0.

	NGA59	SNP5		SNP107		SNP251		M1.10		SNP100	
LerKas1	B	B	B	A	A	A	A	A	A	NA	A
LerKas2	A	A	A	A	A	A	A	A	A	A	A
LerKas3	A	A	A	A	A	A	A	A	A	A	A
LerKas4	A	A	A	A	A	A	A	A	A	A	A
LerKas5	A	NA	NA	NA	A	NA	NA	NA	NA	NA	NA
LerKas6	B	B	A	A	A	A	A	A	A	A	A
LerKas7	B	B	B	B	B	B	B	B	B	B	B
LerKas8	A	A	A	A	A	A	B	B	B	A	A
LerKas9	A	A	B	B	B	B	B	B	A	A	A
LerKas10		B	B	A	A	A	A	A	A	A	A

XGAP: a uniform and extensible data model and software platform for genotype and phenotype experiments Genome biology 2010

R/qtl : Our analysis platform

Open-source software for **R**

Optimized

Multiple experimental crosses

Different statistical methods

* Non-parametric
* Parameric
* Bayesian (qtl/bim)

Parallel computing

Divide and conquer

Scaling up on:

* Multiple CPU (Duo,Quad,Hexa)
* Across networks (Cluster, Grid)
* Globally (Amazone/Rackspace Cloud)

Application (I/V)

Creation of a population

* WUR, JAX

Measure

* Polymorpic markers
* Traits

Creating and QC of the map

Mapping regulation

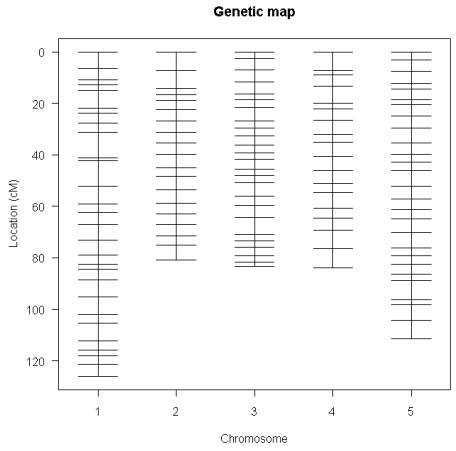
Application (I/V)

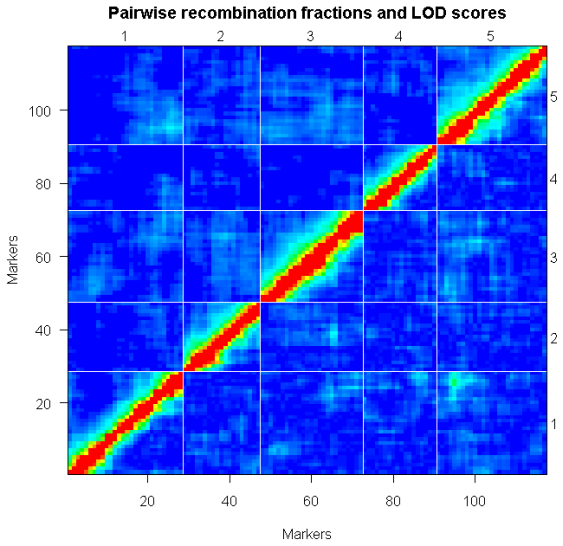
* Initial data comes in
* Reformat into XGAP format
* Upload in generated database
  * Solve data inconsistencies (goto 1)
* Create basic statistics and heatmaps
* Map on cluster using plugin
* Download QTL data and explore in detail

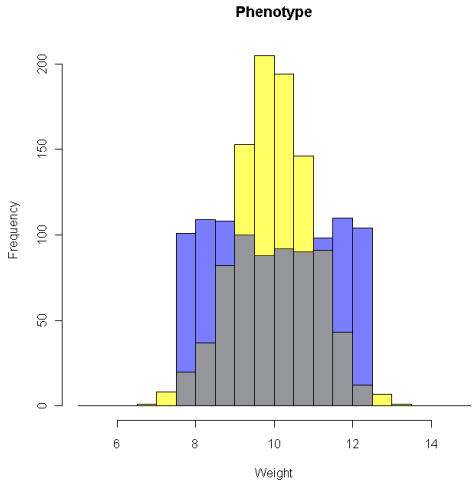
Application (II/V)

Marker segregation / location

Phenotype distribution







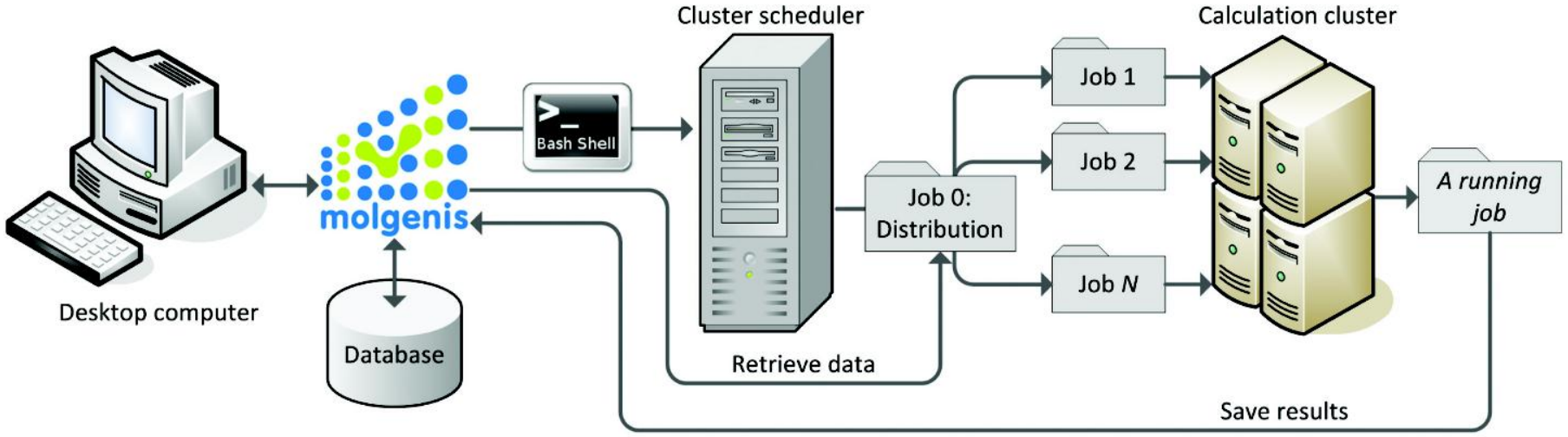
university of groningen

CBSG 2012 Centre for BioSystems Genomics

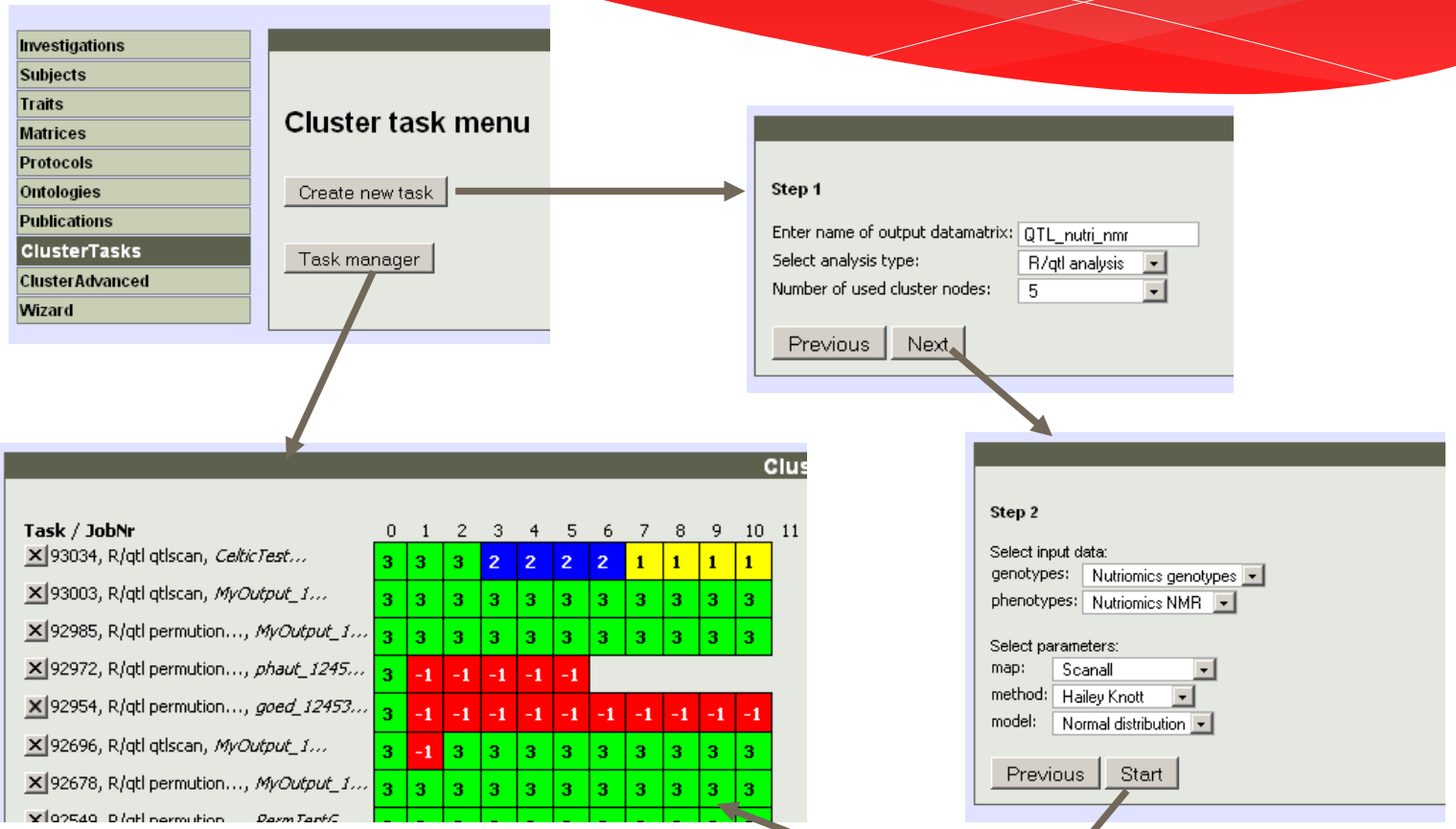
WAGENINGEN UNIVERSITY
WAGENINGEN UR

Application (III/V)

Millipede cluster



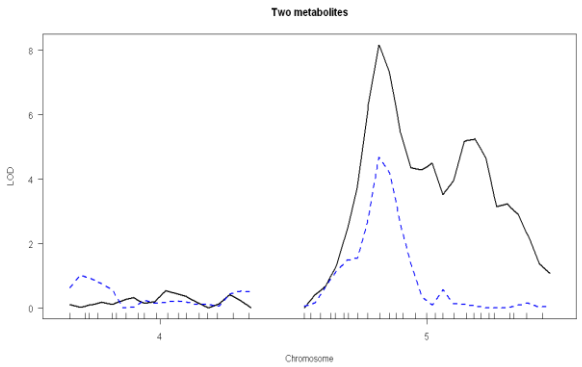
Application (IV/V)

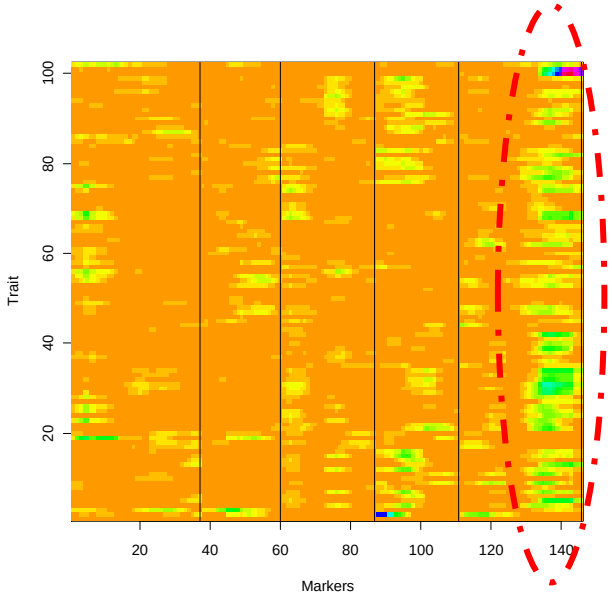


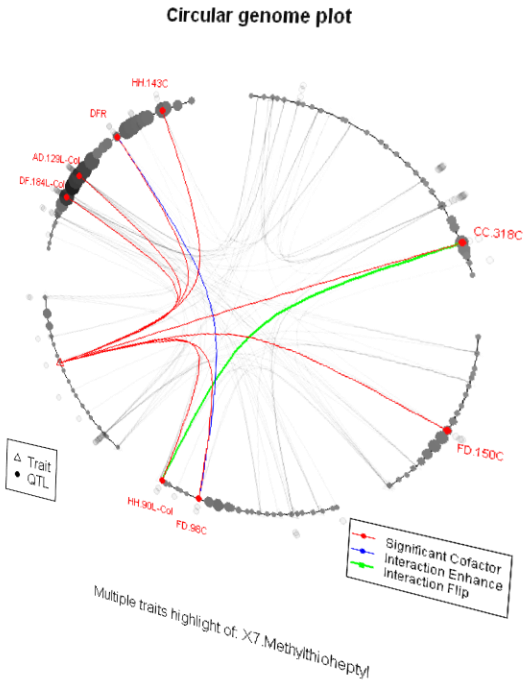
Application (V/V)

Get QTL data from Molgenis

* Explore in more detail







university of groningen

CBSG 2012 Centre for BioSystems Genomics

WAGENINGEN UNIVERSITY
WAGENINGEN UR

Conclusions

Big data problem is solvable

* Discipline
* New / old optimized tools

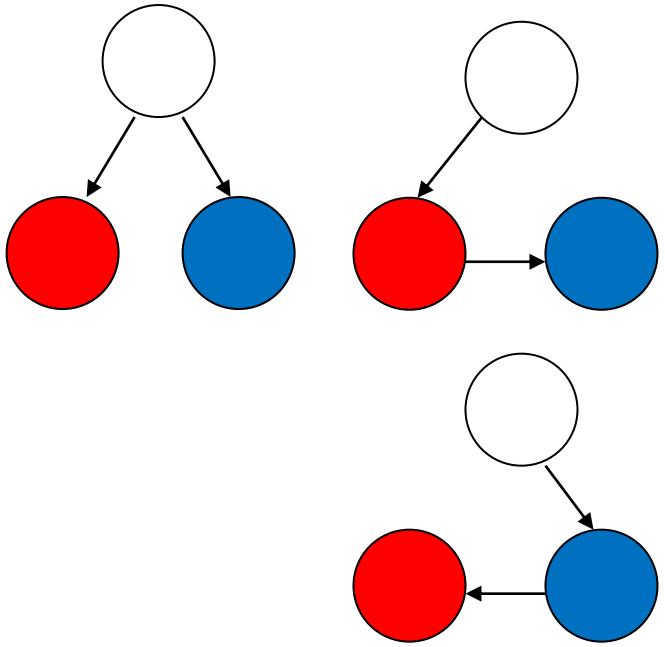
Getting the big picture

* New visualizations
* Ways to show others

Future perspectives / current work

Removing limitations:

* Towards high throughput
* Extending Multiple QTL mapping
  * 4-way, 8-way, Magic and CC



Adding new features:

* Causal inference
* Differential environmental analysis
* Interactive plots

Publications

Used

* Bioinformatics 2003 R/qtl: QTL mapping in experimental crosses.
* Bioinformatics. 2004 Molecular Genetics Information System (MOLGENIS): alternatives in developing local experimental genomics databases.
* Nature 2007 Beyond standardization: dynamic software infrastructures for systems biology
* *Springer 2009 A Guide to QTL Mapping with R/qtl* by K. Broman
* Genome biology 2010 XGAP: a uniform and extensible data model and software platform for genotype and phenotype experimentsBOSC paper Molgenis
* BOSC 2010 MOLGENIS: rapid prototyping of bio software at the push of a button

Pre-submission / Review

* R/qtl: High throughput Multiple QTL Mapping (together with Pjotr Prins and Karl Broman)
* XGAP and Cluster computing (Together with Pjotr Prins and Joeri v/d Velde)

Under construction

* MQMloop together with Ronny Joosen and Wilco Ligterink
* Seed quality in *Arabidopsis* and *Tomato* together with Ronny Joosen and Wilco Ligterink