# GABI GAIN Data INtegration And Mining Tool

Alex Finck, Jürgen Gremmels, Robert Wagner, Birgit Kersten, Björn Usadel

## Max Planck Institute of Molecular Plant Physiology

www.gabiPD.org gabipd@mpimp-golm.mpg.de

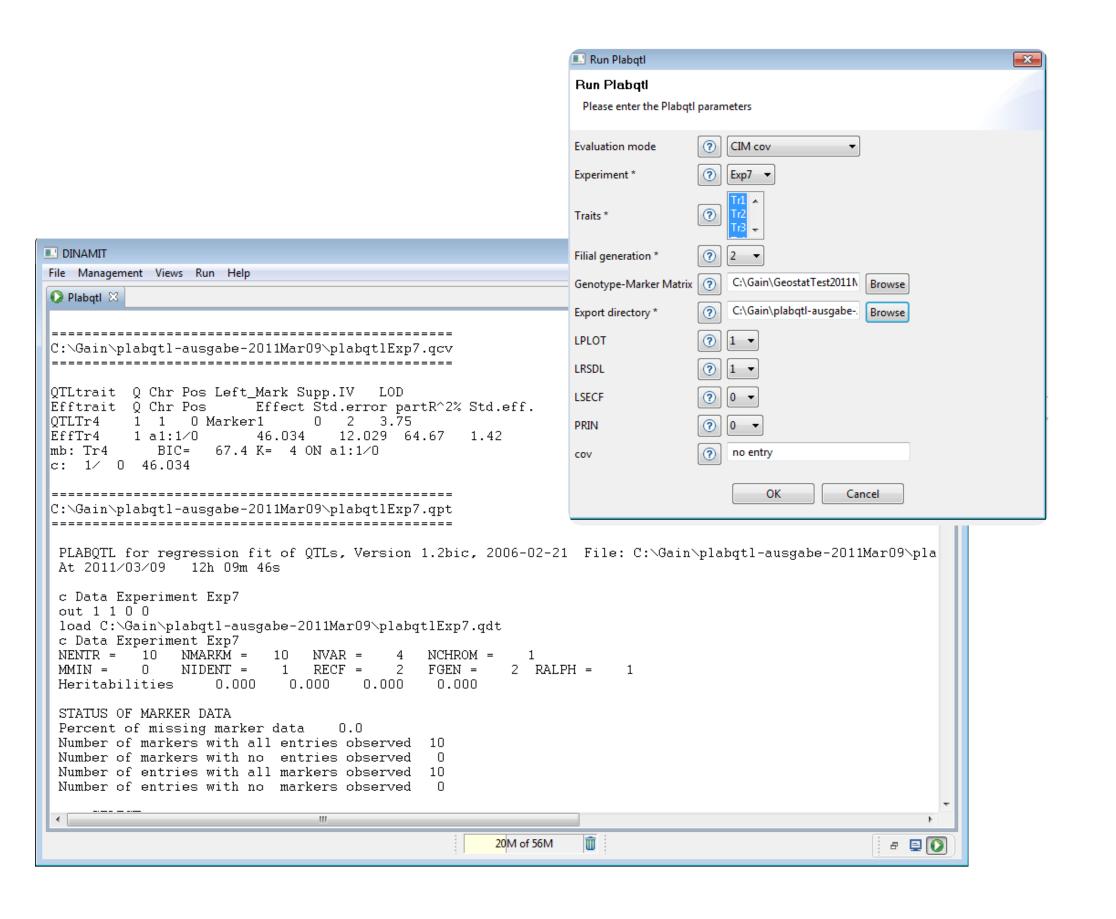
#### Introduction

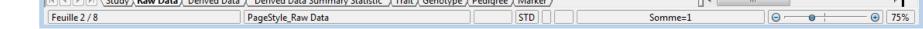
Modern plant breeding research programs produce a huge amount of genomic and phenotypic data. Therefore efficient data mining becomes indispensable to exploit the generated information and to support the improvement of experimental strategies and the optimization of process design. In the scope of the GABI GAIN project, DINAMIT (Data INtegration and Mining Tool) is being developed for the integration of standardized data produced in genomics-based plant breeding. A database schema has been built into DINAMIT to store integrated breeding data of different crops in an efficient and flexible way. A corresponding exchange format, GAIN-Tab, has been defined to ease the integration of new data into DINAMIT. Based on that data, DINAMIT provides a connection with specialized software for the analysis of results of plant breeding experiments. For formats different from GAIN-Tab converters in the form of R-Scripts have been developed. Finally, the development of the graphical user interface DINAMIT provides a user with management functionality on the data, different tabular and graphical visualizations as well as the integration of various tools for planing and analysis of results of geostatistical experiments. Below we present results concerning the integration of plabqtl [1] and R Scripts for geostatistical analysis.

## **GAIN-Tab**

The DINAMIT graphical user interface (GUI) implements general purpose consitency checking for the GAIN-Tab format. GAIN-Tab can easily be generated using spreadsheet tools (Excel, OpenOffice), as shown on Figure 1, or using dedicated automatic converters. Datasets are subsequently validated by the DINAMIT GUI when they get imported.

								[		×				
File Management View	s Run	Help												
Import •	迠 Gain	-Tab												
Export •														
Create +	ER MIT	N ania tak avanantadDarian 20	11Mar08.xls - OpenOffice.org Calc											
8- Preferences			Forma <u>t O</u> utils Donnée <u>s</u> Fe <u>n</u> ét	re Aid <u>e</u>										
Quit		• 😕 🗔 👒 📝 🔒 🖴		\$ 5.	e - (	🗟 🛃 🧸 i 🖻	2∕2   ∰1 (	2 🖻 🗟	9					
			$[0  \bigcirc  \mathbf{G}  I  \underline{S}  \equiv  \exists$											
	; Up	<b>γ</b> <i>f</i> <sub>x</sub> Σ =				.000.0			- ·					
		A B	C D	E		G	н	I		К		M	N	_
	1 5	Study-ID Experiment	Genotype-ID Entry	Check	Year	Location	Replicate	Treatment	FPN	Block	Plot Size	Row	Column	_
	2	1test	148	1	0	2008 HOH		1None		1	1NA		1	_
	3	1 test	2002	1	1	2008 HOH		1None		2	1NA		2	
	4	1test	179	1	0	2008HOH		1None		3	1NA		3	
	5	1 test	986	1	0	2008HOH		1None		4	1NA		4	_
	6	1 test	118	1	0	2008HOH		1None		5	1NA		5	_
	7	1test 1test	984 260	1	0	2008HOH 2008HOH		1None 1None		6	1 NA 1 NA		6	_
	9	1test	258	1	0	2008HOH		1None		8	1NA		8	_
	10	1 test	2001	1	1	2008HOH		1None		9	1NA		9	_
	11	1 test	1049	1	0	2008HOH		1None		10	1NA		10	
	12	1 test	914	1	0	2008 HOH		1None		1	2 NA		11	
	13	1 test	121	1	0	2008 HOH		1None		2	2NA		12	
	14	1 test	134	1	0	2008 HOH		1 None		3	2NA		13	
	15	1 test	1023	1	0	2008 HOH		1None		4	2NA		14	
	16	1 test	2001	1	1	2008HOH		1 None		5	2NA		15	
	17	1test	2002	1	1	2008HOH		1None		6	2NA		16	
	18	1 test	502	1	0	2008HOH		1None		7	2NA		17	
😑 Console 🔀	19	1 test	1030	1	0	2008HOH		1None		8	2NA		18	
Console	20	1 test	333 417	1	0	2008HOH		1None		9	2NA ZNA		19 21	_
console	21	1 test 1 test	782	1	0	2008HOH 2008HOH		1None 1None		2	7 NA 7 NA		21	_
	22	1 test	2002	1	1	2008HOH		1None		3	7 NA		23	-
	23	1 test	896	1	0	2008HOH		1None		4	7NA		24	-
	25	1test	423	1	0	2008HOH		1None		5	7 NA		25	_
	26	1 test	759	1	0	2008HOH		1None		6	7 NA		26	_
	27	1 test	661	1	0	2008HOH		1None		7	7 NA		27	
	28	1 test	7	1	0	2008HOH		1None		8	7 NA		28	
	29	1 test	2001	1	1	2008HOH		1None		9	7 NA		29	
	30	1 test	743	1	0	2008HOH		1 None		10	7 NA		30	
	31	1test	49	1	0	2008HOH		1 None		1	8 NA		31	
	32	1 test	698	1	0	2008HOH		1None		2	8NA		32	
	33	1 test	2002	1	1	2008HOH		1None		3	8 NA		33	
	34	1 test	474	1	0	2008HOH		1None		4	8 NA		34	_
	35	1 test	521	1	0	2008HOH		1None		5	8 NA		35	
		Study Raw Data Deriv												





*Figure 1:* Definition of a GAIN-Tab dataset (foreground) and import into DINAMIT (background)

## QTL detection with plabqtl

In DINAMIT a configuration of plabqtl is specified by the user interactively using a dialog shown in *Figure 2*. A user can configure plabqtl to search for QTLs using Simple Interval Mapping (SIM) and Composite Interval Mapping (CIM) with or without cofactors as advised by the plabqtl documentation. Input data corresponding to a GAIN-Tab dataset originally chosen by the user is then extracted automatically from the built-in database of DINAMIT and results are presented to the user as shown on *Figure 2*.

#### **Geostatistical analysis of plant breeding experiments**

R scripts developed in the context of Gabi GAIN WP C [2] have been integrated into DINAMIT. Thus DINAMIT allows a geostatistical analysis of plant breeding experiments performed using either a replicated alpha design or an unreplicated augmented design. A unified dialog allows a user to choose between the two different designs. Here model dependent consistency checks on a GAIN-Tab dataset originally chosen by the user can also be specified. After confirmation of the dialog and consistency of the data has been asserted results are presented to the user as shown on *Figure 3*.

Figure 2: Configuration dialog for plabqtl (upper right) and corresponding results (lower left)

💷 Run Geostatistic

Block identity

Contiguity

factor(geno)2 factor(geno)3

ctor(geno)21 factor(geno)22 factor(geno)23 factor(geno)24

Structure: Gaussian spatial correlation

4.27504011

(Intercept) Residual

File Management Views Run Help

4.66202874

0.34689103

Formula: ~1 | bloc]

Formula: ~1 | block Parameter estimate(s)

Number of Groups: 36

Number of Observations: 144

andom effects

508016

og-restricted-likelihood: -53.79528 ixed: y ~ factor(geno) + z1 + z2 - 3

Norking directory

xport file name

3.71290793

4.18370034

1 28 177.1224 254.7015 -60.56118

Run Geostatistic

Please enter the Geostatistic parameters

AUGMENTED

BLOCK\_ONLY

IMPLICIT -

iain\GeostatTest2011Mar08 Browse

OK

factor(geno)4

4.74409183

Run Geostatistic

Run Geostatistic

Contiguity

Working directory '

Export file name

Browse

4.17250022

0.02983157

Cancel

Please enter the Geostatistic paramete

ALPHA

BLOCK\_ONLY

IMPLICIT -

OK

IMPLICIT

st2011Mar08 Browse

Cancel

🐴 🗆 🗆

## Availability

DINAMIT is provided as a locally installable application for partners within the GABI GAIN project to optimize genomics-based plant breeding

#### **References and Publications**

[1] Utz & Melchinger (1996), 'PLABQTL: A program for composite interval mapping of QTL', Journal of Agricultural Genomics 2, 1.
[2] Müller, S. & Piepho (2010), 'Arrangement of check plots in augmented block designs when spatial analysis is used', Plant Breeding 129, 581-589.

Geostat\$corAR1 Geostat\$corSpher	2 29 170.4782 250.8281 -56.23911 1 vs 2 8.644157 0.0033 3 29 169.3995 249.7493 -55.69973	
Geostat\$corExp	4 29 170.4782 250.8281 -56.23911	
Geostat\$corGaus	5 29 165.5906 245.9404 -53.79528	
•		
	23M of 48M	

logLik Test L.Ratio p-value

*Figure 3:* Configuration dialog for geostatistical analysis of datasets produced using an alpha design (upper right) or an augmented design (middle) and corresponding results (below)

#### Acknowledgments

This work is supported by the German Federal Ministry of Education and Research BMBF (GABI-FUTURE grant 0315072C)