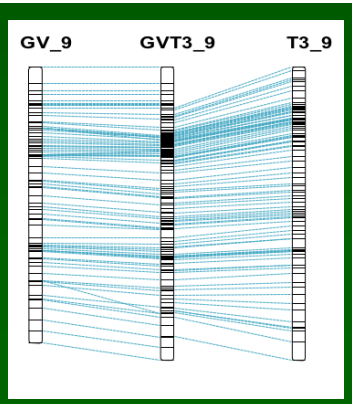
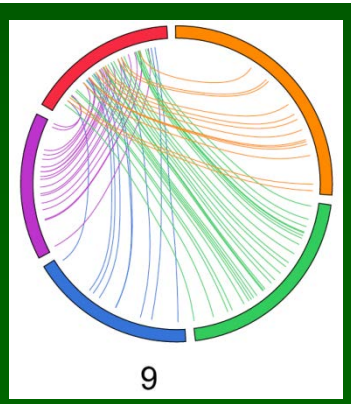
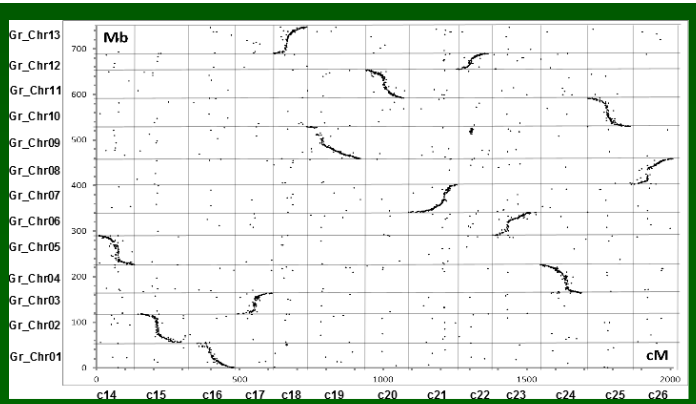
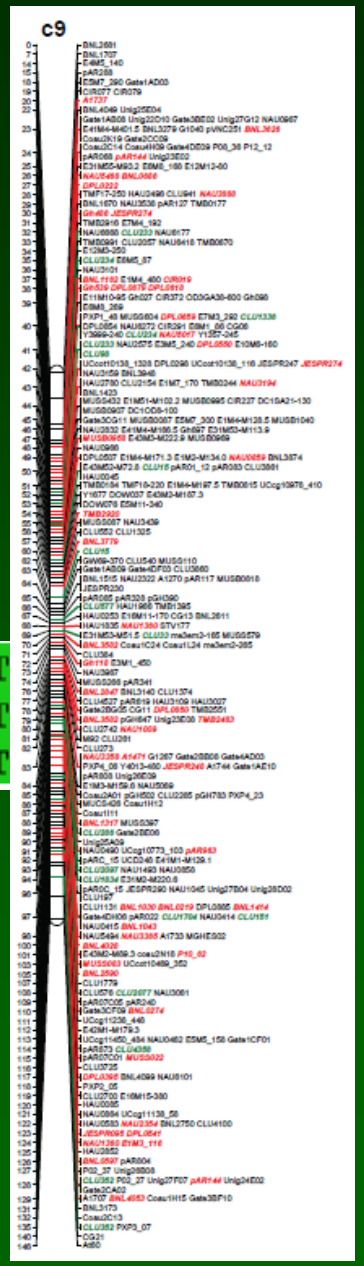
Molecular marker redundancy check and construction of a high density genetic map of tetraploid cotton

Jean-Marc **Lacape**









Raleigh ICGI, October 10, 2012

Background ICGI/2008 Anyang

Structural Genetics WG priorities

Numerous marker projects
(several 10s thousands)
SSRs: BNL, CIR, etc..
SNPs, etc …

→ 1- Marker nomenclature, synonymy and redundancy

Several genetic/QTL maps
(several 10s)
Interspecific (*Gh*x*Gb*)
Intraspecific (*Gh*x*Gh*)
Others ..

→ 2- Map convergence, and building of a consensus map

Genome sequencing
D-genome
(A- genome)
(AD-genome)

→ 3- Convergence genetic-physical

Background

Numerous marker projects (several 10s thousands)
SSRs: BNL, CIR, etc..
SNPs, etc …

Several genetic/QTL maps (several 10s)
Interspecific (*Gh*x*Gb*)
Intraspecific (*Gh*x*Gh*)
Others ..

Genome sequencing
D-genome
(A- genome)
(AD-genome)

Objectives

1- Curate redundancy among SSR markers

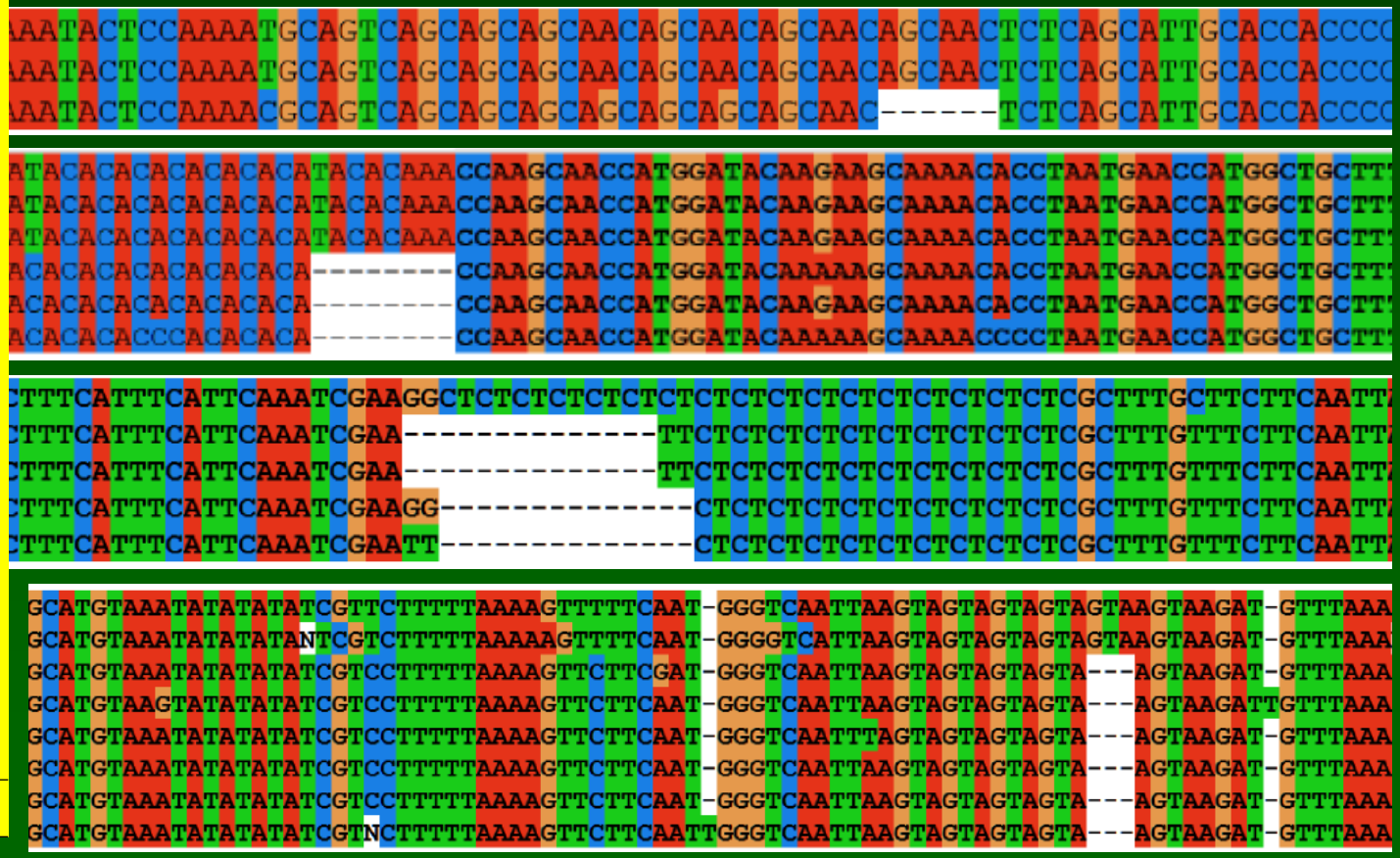
2- Build a High Density Consensus (**HDC**) map from 6 component maps

3- Conduct genetic-physical map alignments

1- Marker redundancy

Library	No seq. In library
BNL	379
CIR	392
CM	53
DOW	52
DPL	200
Gh	700
HAU	3383
JESPR	309
MGHES	84
MON	2568
MUCS	617
MUSB	1316
MUSS	554
NAU	3249
NBRI	2233
PGML	308
STV	192
TMB	754
total	**17343**

- ►17343 markers from CMD (18 different libraries) as of April 2012
- ► pair-wise sequence alignment (Smith-Waterman algorithm)
- ► similarity threshold 90%



1- Marker redundancy

► at 90% sequence similarity threshold, **5741 markers**,

Lib	BNL	CIR	CM	DOW	DPL	Gh	HAU	JESPR	MGHES	MON	MUCS	MUSB	MUSS	NAU	NBRI	PGML	STV	TMB
BNL	9	10	1		4	8	4	4		16	2			4	1			3
CIR		6				7	2	3		8				2	1	1		40
CM			5		1	3		28		3								1
DOW				1	1	3		1										
DPL					1	2	5			2				5	3			3
Gh						103	5	29	1	27	1	1		4	10			1
HAU							646	3	49	244	145	1	206	872	136	55	46	3
JESPR								14	4	11				4	2			1
MGHES									3	40	6		10	52	3	2		
MON										183	14	2	22	182	56	6	8	6
MUCS											54		162	114	28	22		
MUSB												168		12	1			
MUSS													5	224	26	19		
NAU														593	123	50	115	
NBRI															146	5	4	1
PGML																16	2	3
STV																	12	
TMB																		35

1- Marker redundancy

► Level of redundancy within libraries

Lib	BNL	CIR	CM	DOW	DPL	Gh	HAU	JESPR	MGHES	MON	MUCS	MUSB	MUSS	NAU	NBRI	PGML	STV	TMB
BNL	9	10	1		4	8	4	4		16	2			4	1			3
CIR		6				7	2	3		8				2	1	1		40
CM			5		1	3		28		3								1
DOW				1	1	3		1										
DPL					1	2	5			2				5	3			3
Gh						103	5	29	1	27	1	1		4	10			1
HAU							646	3	49	244	145	1	206	872	136	55	46	3
JESPR								14	4	11				4	2			1
MGHES									3	40	6		10	52	3	2		
MON										183	14	2	22	182	56	6	8	6
MUCS											54		162	114	28	22		
MUSB												168		12	1			
MUSS													5	224	26	19		
NAU														593	125	50	115	
NBRI															146	5	4	1
PGML																16	2	3
STV																	12	
TMB																		35

► Within library, redundancy was highest (>10%) in Gh, MUSB, NAU and HAU

1- Marker redundancy

- Level of redundancy between libraries

Lib	BNL	CIR	CM	DOW	DPL	Gh	HAU	JESPR	MGHES	MON	MUCS	MUSB	MUSS	NAU	NBRI	PGML	STV	TMB
BNL	9	10	1		4	8	4	4		16	2			4	1			3
CIR		6				7	2	3		8				2	1	1		40
CM			5		1	3		28		3								1
DOW				1	1	3		1										
DPL					1	2	5			2				5	3			3
Gh						103	5	29	1	27	1	1		4	10			1
HAU							646	3	49	244	145	1	206	872	136	55	46	3
JESPR								14	4	11				4	2			1
MGHES									3	40	6		10	52	3	2		
MON										183	14	2	22	182	56	6	8	6
MUCS											54		162	114	28	22		
MUSB												168		12	1			
MUSS													5	224	26	19		
NAU														593	123	50	115	
NBRI															146	5	4	1
PGML																16	2	3
STV																	12	
TMB																		35

- Between libraries, matches amongst most libraries

1- Marker redundancy

Library	No seq. In library	Total cross library (pair-wise)
BNL	379	57
CIR	392	74
CM	53	37
DOW	52	5
DPL	200	26
Gh	700	102
HAU	3383	1776
JESPR	309	90
MGHES	84	167
MON	2568	647
MUCS	617	494
MUSB	1316	17
MUSS	554	669
NAU	3249	1763
NBRI	2233	400
PGML	308	165
STV	192	175
TMB	754	62
total	**17343**	**6726**

► our sequence-based redundancy check aimed at recovering **identical-by-sequence** SSR markers (cross-PCR-amplification), thus recovering as « redundants »:

- homoeologs, paralogs and other copies
- sequences with mutiple SSRs
- alleles etc …

► It is **necessary** to account-for marker/locus redundancy for (1) across- map alignments and (2) consensus map construction

2- A High Density Consensus map

Six selected interspecific Gh x Gb genetic maps (only map positions, **no raw codings**)

Map code	Parents	Gene-ration	References
T3	TM-1 x 3-79	RIL	Yu et al (2012)
GV	Guazuncho 2 x VH8	BC_1-RIL	Lacape et al (2009)
PK	Palmeri x K101	F_2	Rong et al (2004)
TH	TM-1 x Hai 7124	BC_1	Guo et al (2008)
E3	Emian22 x 3-79	BC_1	Yu et al (2011)
CH	CRI 36 x Hai 7124	F_2	Yu et al (2007)

► standardize marker names (BNL119=BNL0119)

► verify orientation

► acronyms **T3**, **GV**, **PK**, **TH**, **E3** and **CH**

2- A High Density Consensus map

Six maps were saturated with combination of different markers types (SSRs as predominant)

Map code	Gene-ration	Pop size	Total	cM	No. of mapped markers				
					RFLP	AFLP	SSR	SNP	Others
T3	RIL	186	2072	3380			1825	247	
GV	BC1-RIL	215	1745	3637	190	715	781		59
PK	F2	57	2584	4448	2459		124		1
TH	BC1	138	2247	3541		71	1865	10	301
E3	BC1	141	2316	4419			2311		5
CH	F2	186	1080	4418		93	690		297
Total			12044		2649	879	7596	257	663

2- A High Density Consensus map

Six maps with good level of connections: as many as 10086 bridges (same chromosome) altogether

No bridges (same chromosome)

Before redundancy curation

Map	T3	GV	PK	TH	E3	CH	Total Betw.
T3	105	422	63	408	837	351	2081
GV		37	221	350	323	329	1645
PK			76	69	64	79	496
TH				66	929	283	2039
E3					193	315	2468
CH						31	1357

Probably over-estimated because of redundancy

2- A High Density Consensus map

Following redundancy check, groups of suspected redundant markers **(2357 clusters with up to 14 SSR)** were assigned a collective name '**CLU**' (cluster)

HAU1430
NAU3074
STV069

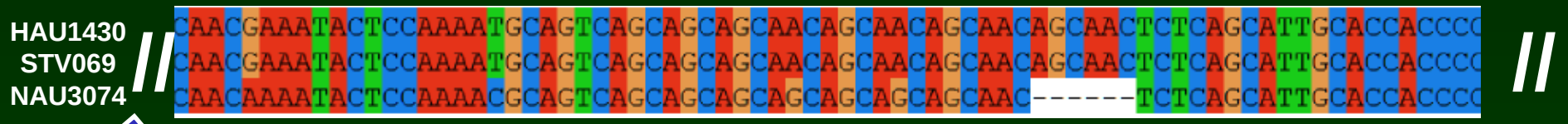
► 3 markers renamed as **CLU1057**

JESPR240
HAU2108
MGHES10
NAU3533
NAU5239

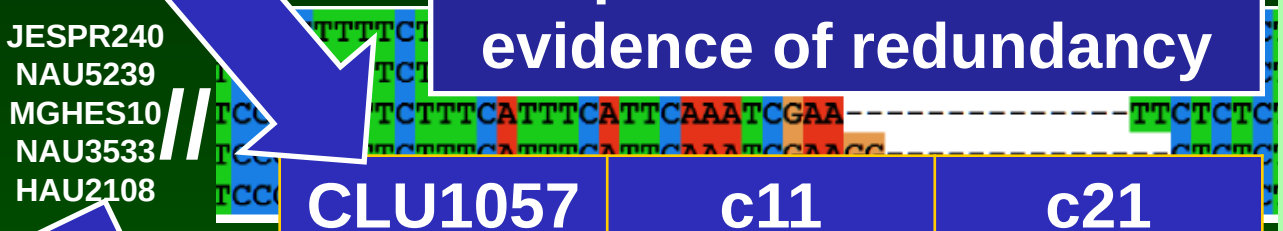
► 5 markers as renamed **CLU1355**

► **Additional evidence of redundancy from map data**

► CLU1057



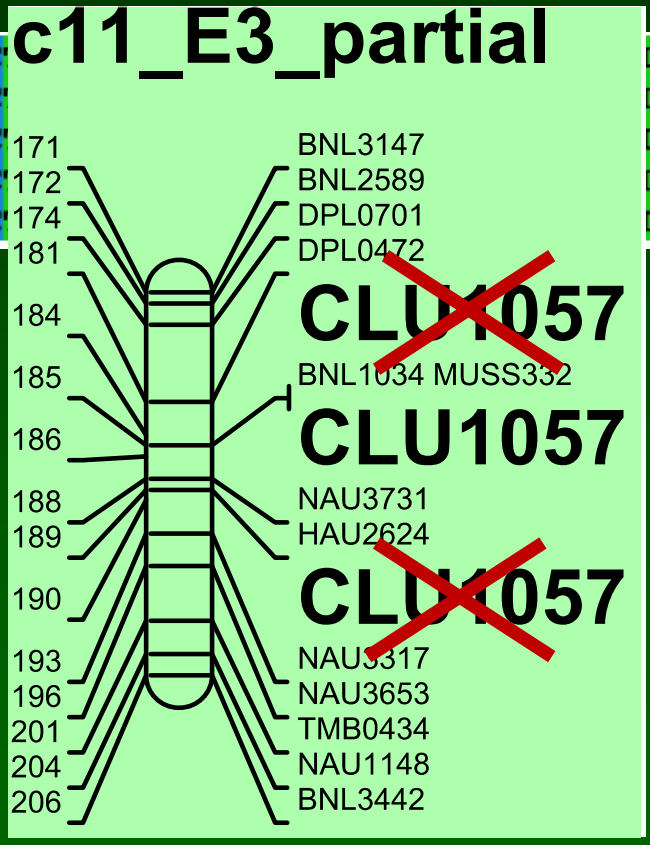
► CLU1355



Map data as additional evidence of redundancy

CLU1057	c11	c21
HAU1430	E3	
STV069	E3, T3	T3
NAU3074	E3, TH	E3, TH

CLU1355	c1	c15
JESPR240	T3, TH	T3, TH
NAU3533	E3, TH	E3
MGHES10	T3, CH, GV	T3, CH, GV



2- A High Density Consensus map

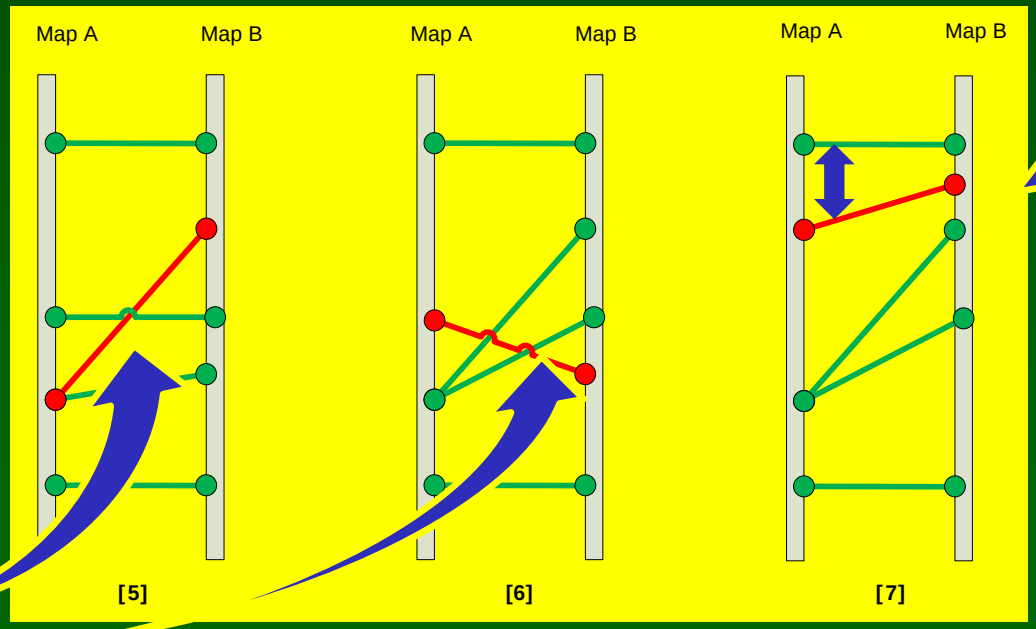
Redundancy check (replacing marker names by a collective 'CLU' name) resulted in:

► additionnal **new** correspondances: 2 redundant markers mapped on 2 different maps, including erroneous paralogs

► numerous **spurious** correspondances generated by co-localized redundant markers on the same map

► need for a curation of duplicates both internally and between maps

► **all inversions between all possible pairwise comparisons had to be solved**



2- A High Density Consensus map

Marker redundancy and map curation resulted in a decrease of the number of bridges (5578 between instead of 10086)

No bridges (same chromosome)

Before redundancy curation

Map	T3	GV	PK	TH	E3	CH	Total
T3	105	422	63	408	837	351	2081
GV		37	221	350	323	329	1645
PK			76	69	64	79	496
TH				66	929	283	2039
E3					193	315	2468
CH						31	1357

After redundancy curation

T3	GV	PK	TH	E3	CH	Total
13	258	37	226	398	180	1099
	17	179	196	161	178	972
		75	34	35	39	324
			19	583	142	1181
				32	[illegible]	1320
					12	682

2- A High Density Consensus map

After map curation

Map code	Total	No unique loci	No loci for map integration
T3	2072	677	1977
GV	1745	778	1671
PK	2584	1352	2559
TH	2247	796	1892
E3	2316	480	1814
CH	1080	33	417
Total	**12044**	**4116**	**10330**

Map integration
(using *Biomercator v3*)

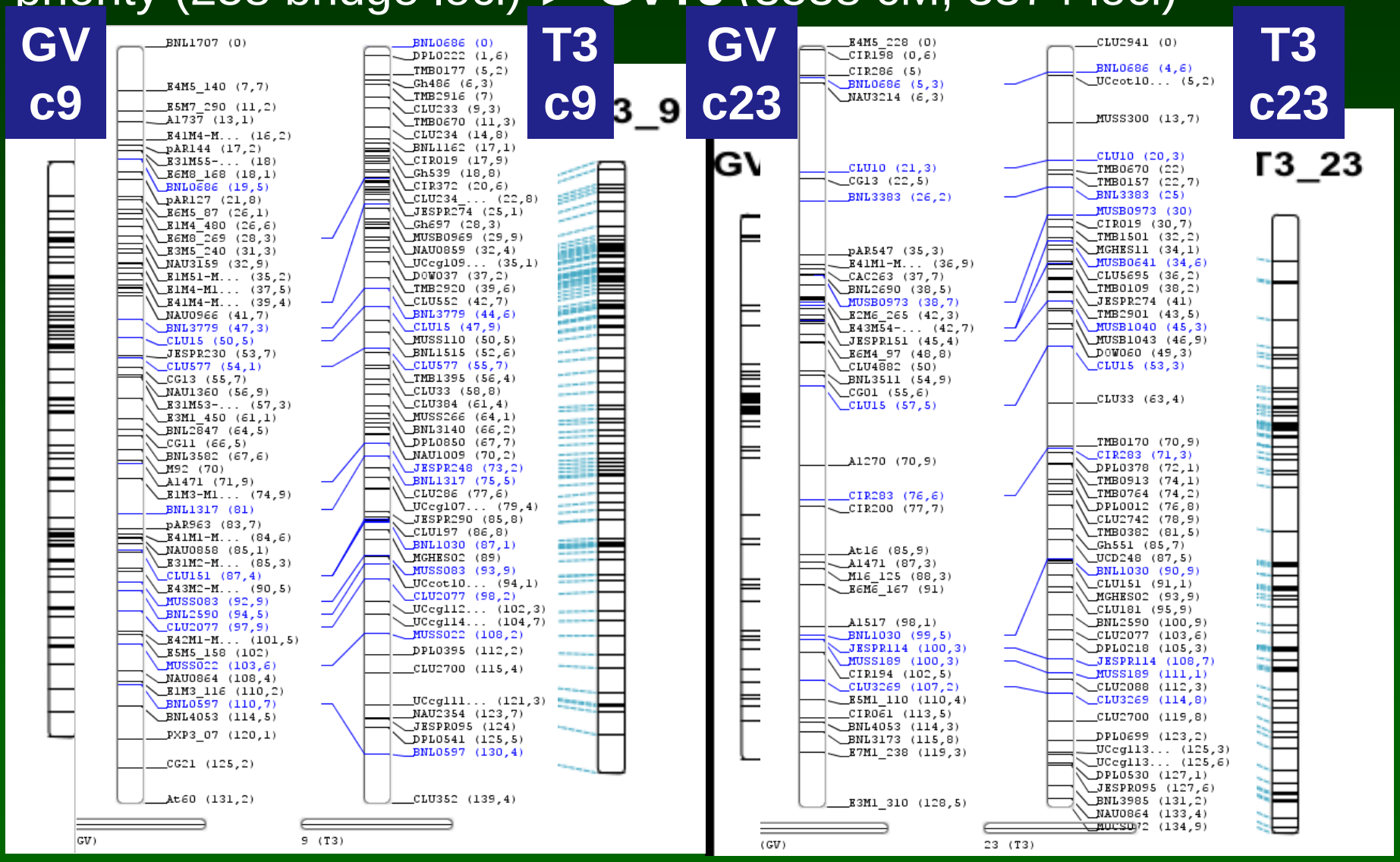
With 2 maps, GV and T3, given «higher confidence», a HDC map was build in **2 steps**:

► **Step 1**: integration of GV and T3, **no reference** (result=GVT3)

► **Step 2**: iterative projections of PK, TH, E3 and CH with **GVT3 as fixed reference**

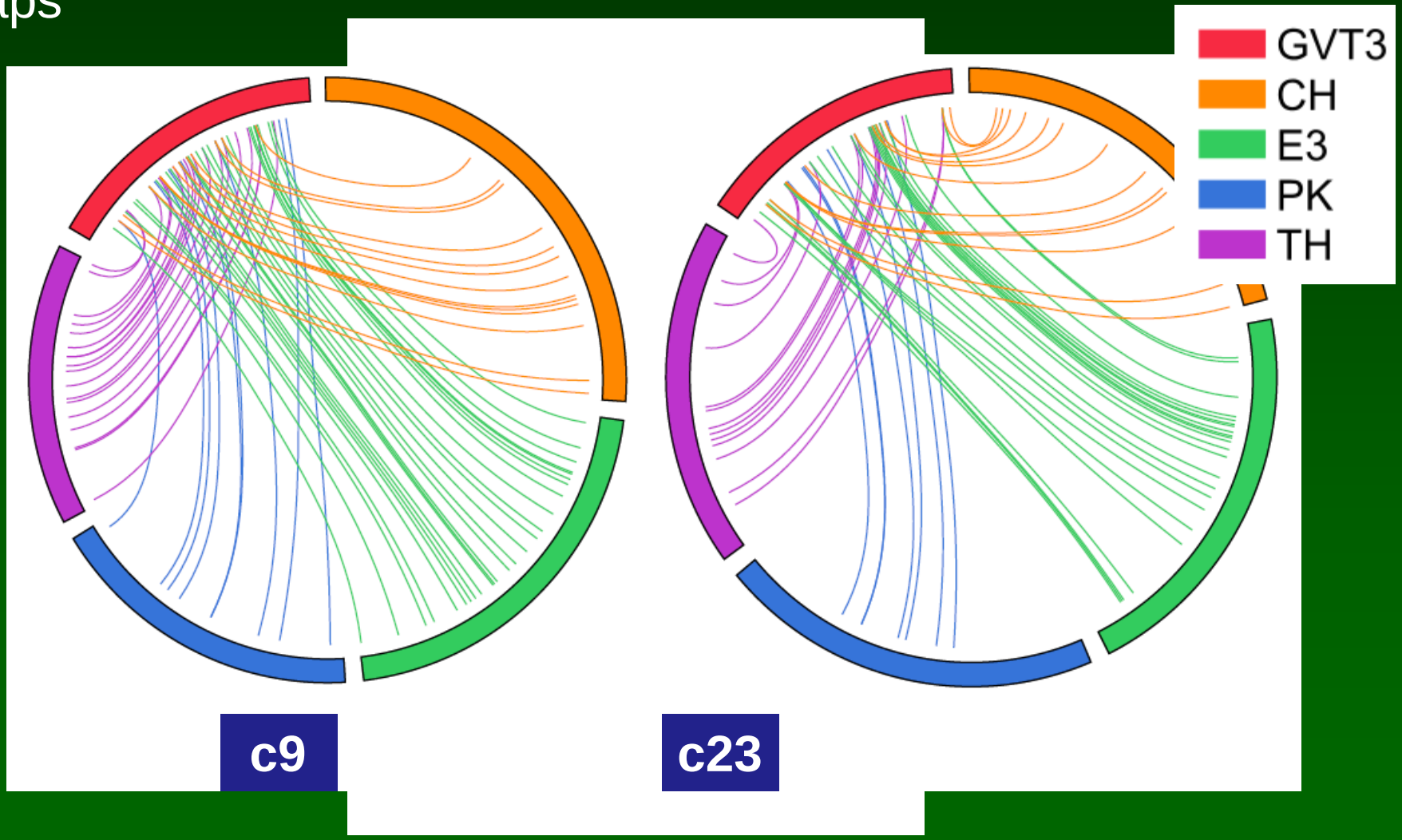
2- A High Density Consensus map

Step 1 (*Biomercator*): integration of GV+T3, 2 maps of 1st priority (258 bridge loci) ► **GVT3** (3538 cM, 3374 loci)

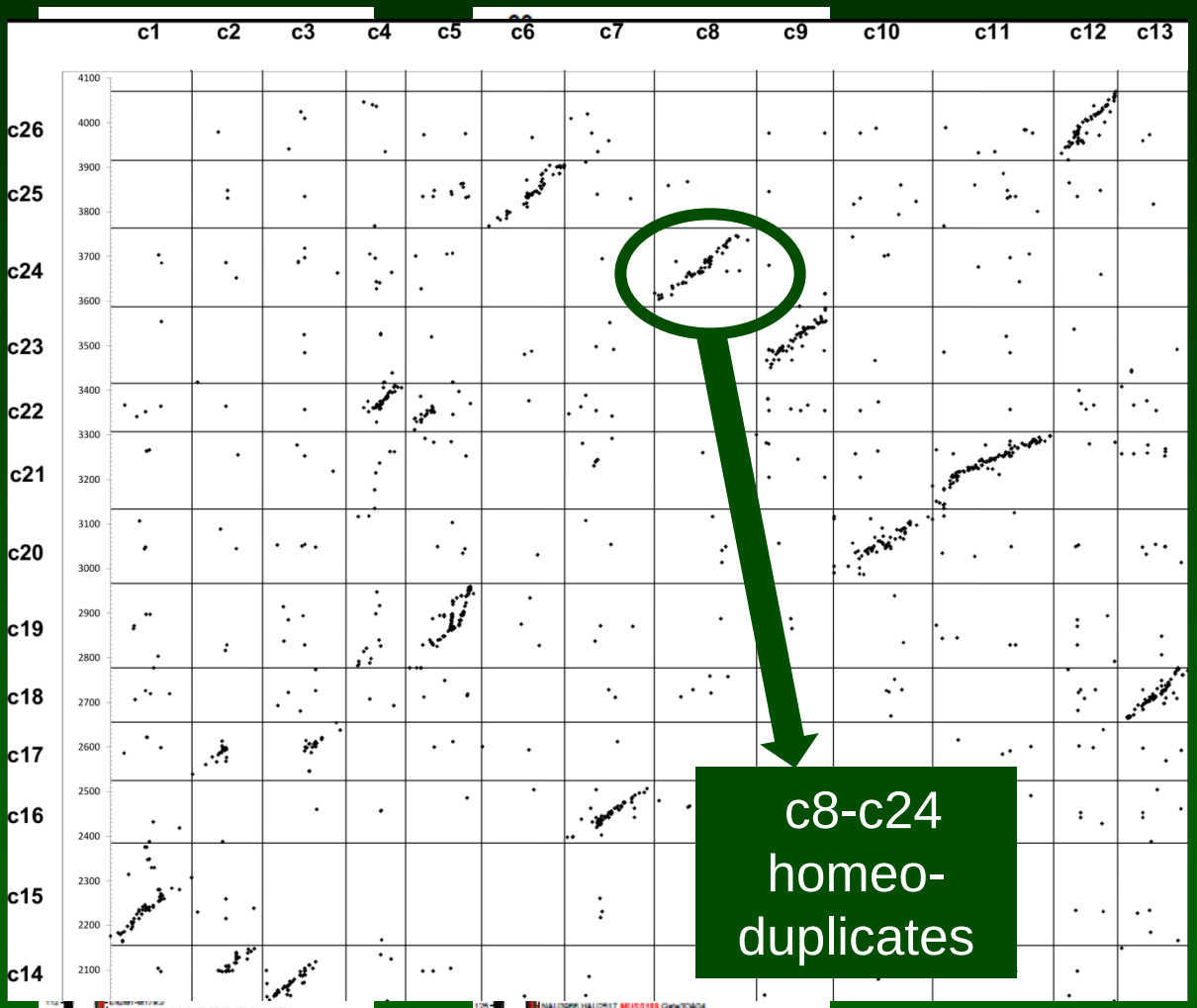


2- A High Density Consensus map

Step 2 *(Biomercator):* Consensus of GV+T3 used as « fixed » framework for projecting markers from 4 other component maps



2- A High Density Consensus map



HDC map represents:

- ► 8254 loci (317/chr)
- ► 6669 « unique » markers, 3450 SSRs (727 as 'CLU'), 1894 RFLPs
- ► 4070 cM, 2 loci/cM
- ► a single dense region per chrom.
- ► 1292 multi-copy markers, 64% homeo-duplicates

que markers of the HDC map are
a sequence (genomic, cDNA)

3- Convergence HDC map / *G raimondii*

► alignment of the sequences of :

- the 4744 markers mapped on the HDC map and
- the 13 scaffolds (750 Mb) of *G raimondii* (www.phytozome, v2.1, **Sept.2012**)

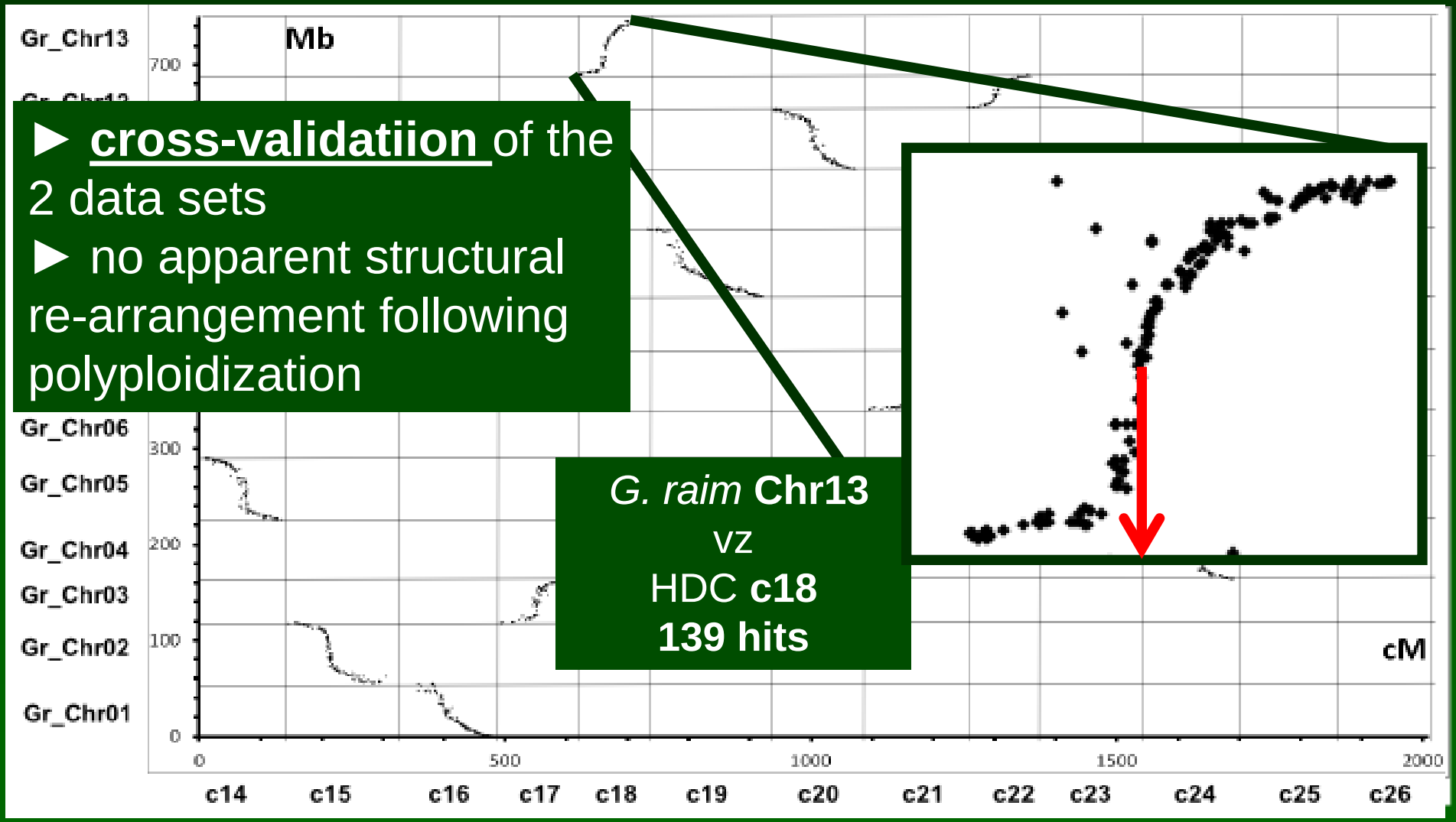
► using **BBMH ,** Best Blast (Bidirectional) Mutual Hit approach, **(**BLASTN, 1e-5)

► the **3607 single hits** (4592 loci) included 2182 hits (2369 loci) for the D chromosomes of HDC map (c14-c26), or **182 anchor loci per chromosome**

The sequences originate from different ploidy/evolutionary context (1) from a **tetraploid** (markers on the tetraploid HDC map) and (2) from a **diploid** context (*G raimondii* genome)

3- Convergence HDC map / *G raimondii*

► **excellent** convergence ► ratio kb/cM highly variable (average 400 kb/cM) ; S-shaped curve (high ratio in centromeric regions)



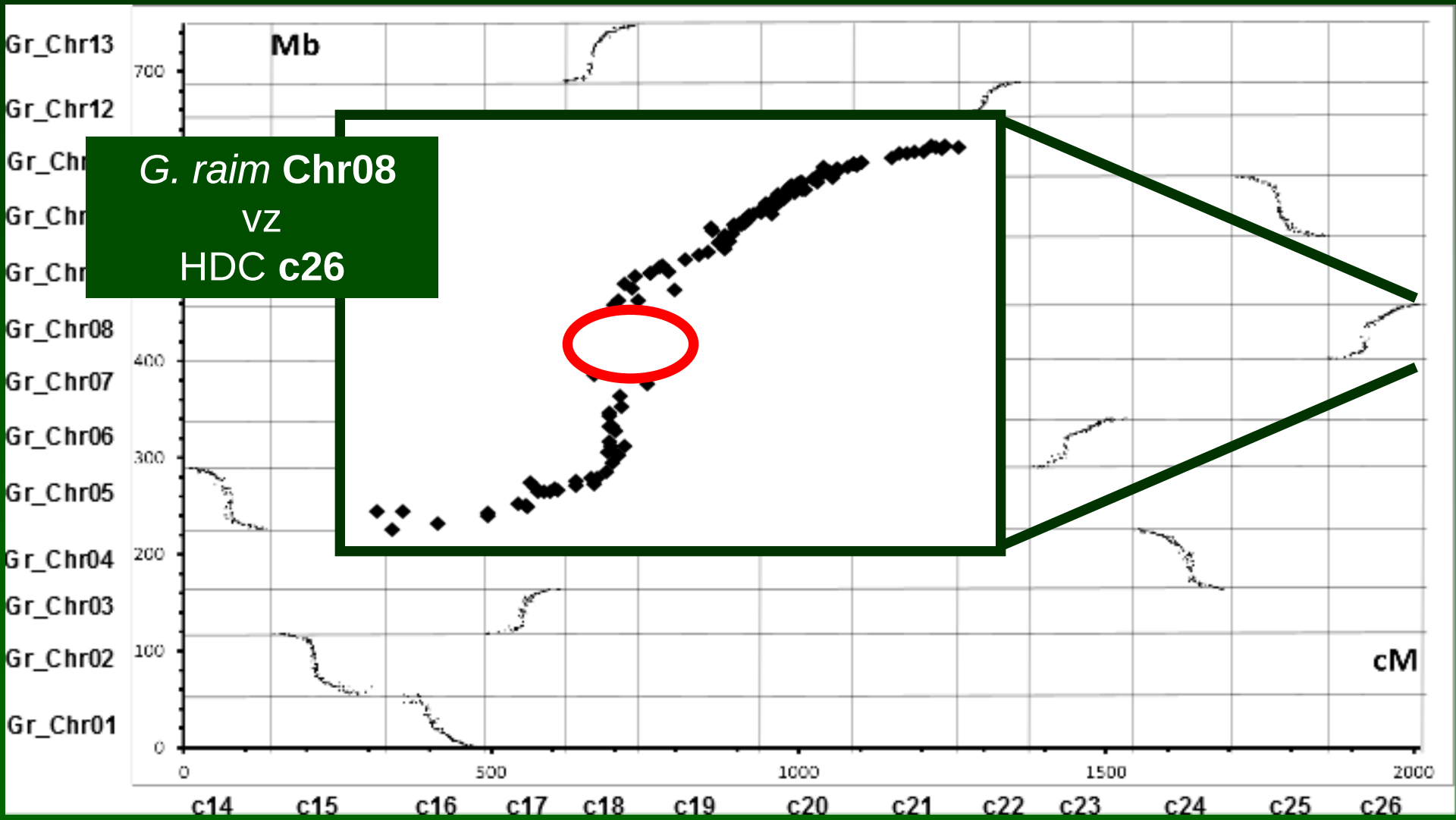
► **cross-validatiion** of the 2 data sets

► no apparent structural re-arrangement following polyploidization

G. raim **Chr13**
vz
HDC **c18**
139 hits

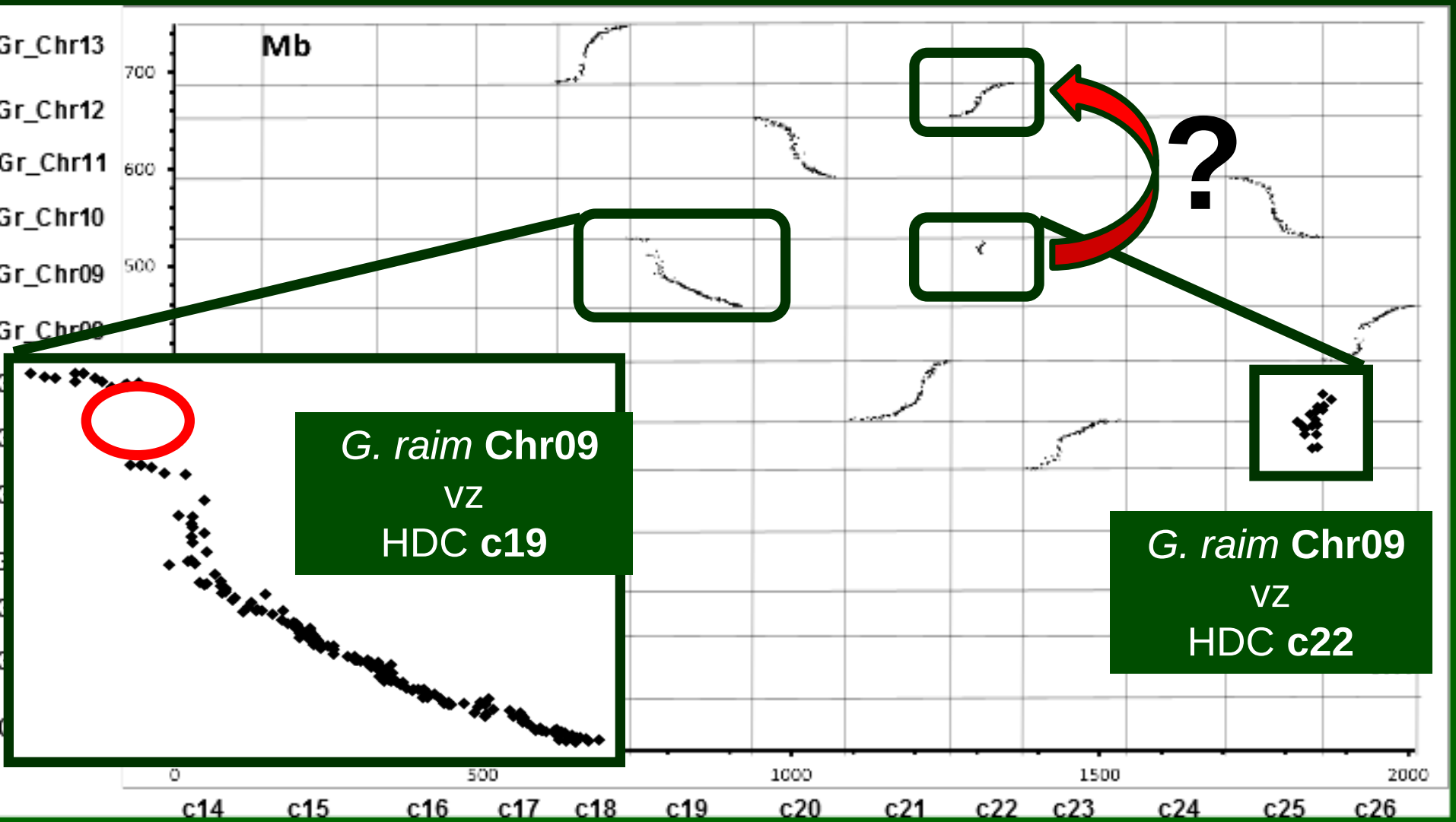
► **<u>after</u>** clustering (5 colinear pairs): possible incongruency

► an internal region of Gr_Chr08 with not hit on genetic map



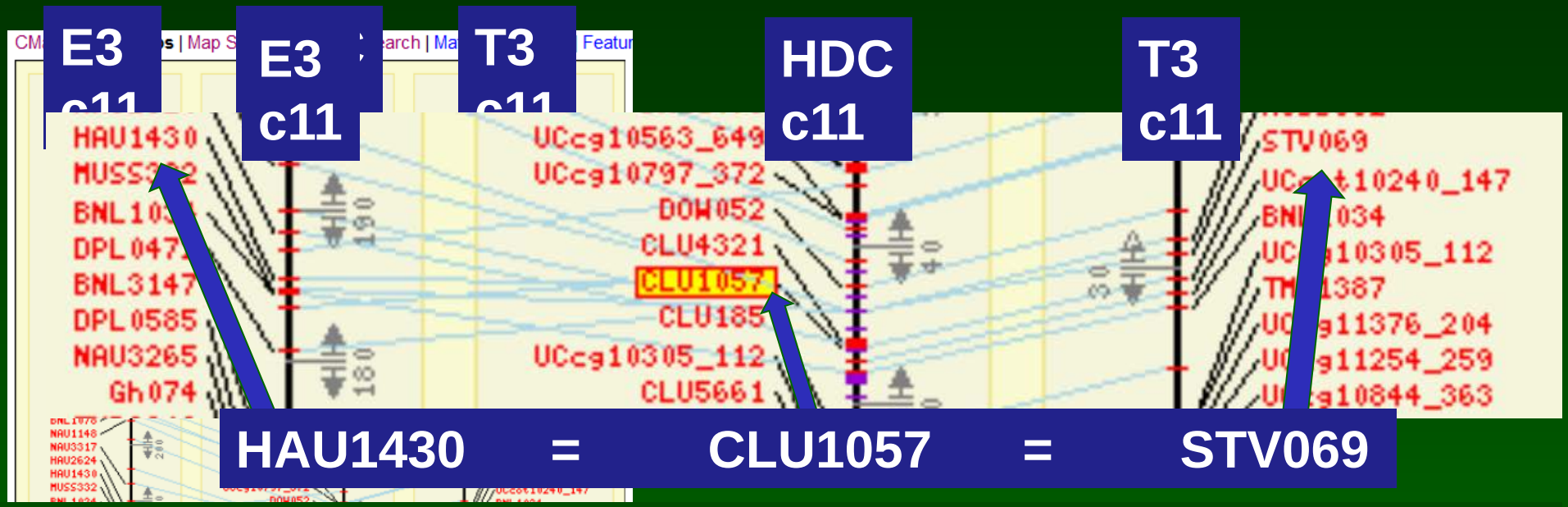
► **after** clustering (5 colinear pairs) : possible incongruency

► dual correspondance of Gr_Chr09 with c19 and c22



last but not least …

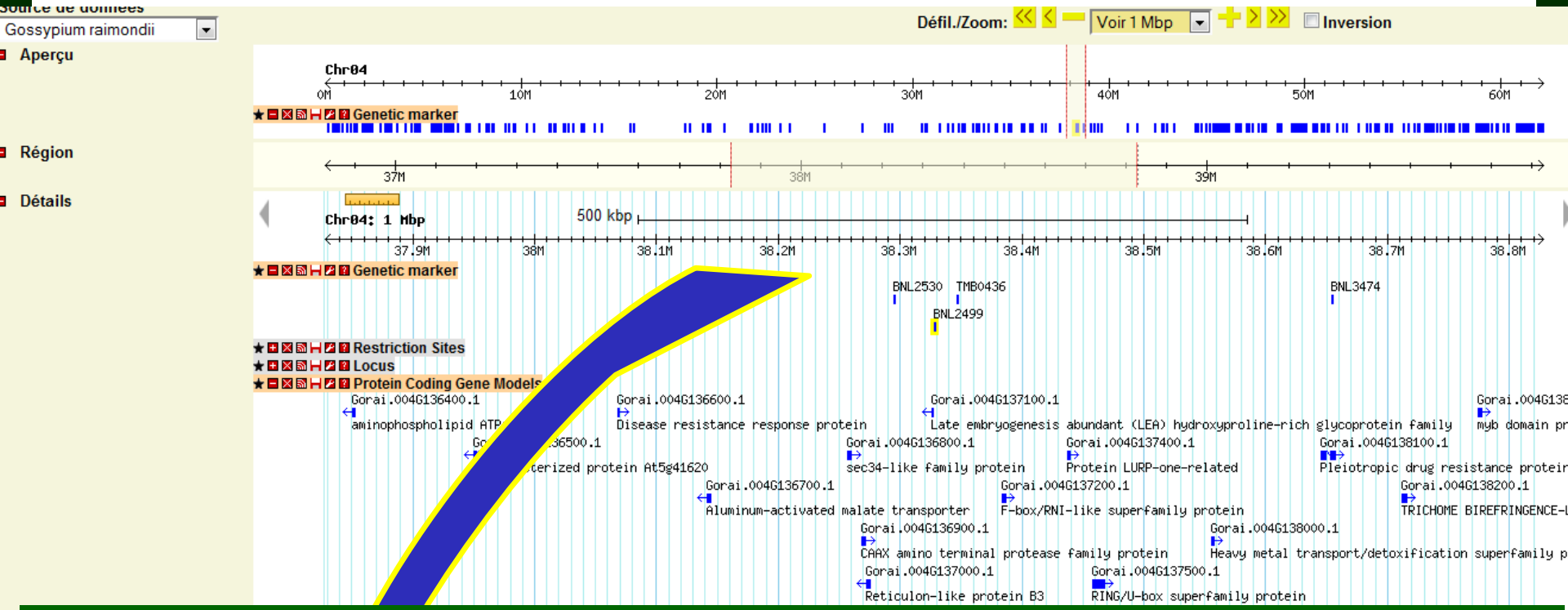
New data on TropGeneDB/Cmap (www.tropgenedb.cirad.fr)



► CLU markers serve as bridges between genetic maps (thanks to aliases/feature names associated with markers)

IN ADDITION :

► added (TropGeneDB, Cmap) the ***G raimondii*** v2.1 **physical map** and correspondances/HDC ► more QTLs (**metaQTLs**, **expression eQTLs hotspots**) ► created links from markers (Cmap) to a genome **browser** of *G raimondii*



► (list of) candidate gene(s) in QTL confidence interval

tropgenedb.cirad.fr/

►Any request for additional map and QTL data to be included in **TropgeneDB**, email to:

marc.lacape@cirad.fr or chantal.hamelin@cirad.fr

► **Perspective**: exchange with CottonGen

This research is published in *PLosONE:*

OPEN ACCESS Freely available online PLOS | ONE

A High Density Consensus Genetic Map of Tetraploid Cotton That Integrates Multiple Component Maps through Molecular Marker Redundancy Check

Anna Blenda[1,2], David D. Fang[3], Jean-François Rami[4], Olivier Garsmeur[4], Feng Luo[5], Jean-Marc Lacape[4]*

1 Department of Genetics and Biochemistry, Clemson University, Clemson, South Carolina, United States of America, **2** Department of Biology, Erskine College, Due West, South Carolina, United States of America, **3** Cotton Fiber Bioscience Research Unit, USDA-ARS-SRRC, New Orleans, Louisiana, United States of America, **4** CIRAD, UMR AGAP, Montpellier, France, **5** School of Computing, Clemson University, Clemson, South Carolina, United States of America

Acknowledgments and Contributors

Acknowledged are

- Andrew **Paterson** and JGI for early access to *G raimondii* scaffolds
- Johann **Joets** and Olivier **Sosnowski** for early access to *Biomercator v3*

Contributors:

- Anna **Blenda** (Clemson Un.)
- David **Fang** (ARS)
- Feng **Luo** (Clemson Un.)
- Jean-François **Rami** (Cirad)
- Olivier **Garsmeur** (Cirad)
- Chantal **Hamelin** (Cirad)
- Gaëtan **Droc** (Cirad)

Thank you, Merci ...