Analysis and characterization of the repetitive sequences of *T. aestivum* chromosome 4D

Romero J.R., Garbus, I., Helguera M., Tranquilli G., Paniego N., Caccamo M., Valarik M., Simkova H., Dolezel J., Echenique V.

CERZOS
CONICET
U N S

The aim of this work was the analysis and characterization of the repetitive sequences of *Triticum aestivum* chromosome 4D arms obtained by flow sorting and sequenced using Next Generation Sequencing technologies (NGS)

Methods

Sequences from chromosome 4D

- Contigs were assembled through Newbler into 8141 and 7077 scaffolds for 4DS and 4DL, respectively, ranging from 2000 to 47795 bp.

Identification of repetitive elements

- Repetitive sequences were identified using RepeatMasker
- Sequence comparisons of the signatures of each family of TE were performed using the alignment program cross_match
- The interspersed repeat database used was mips-REdat_v9.0p, hosted by the MIPS at PlantsDB (42,000 sequences)
- To reduce redundancies, the sequences with >=95% identity over >=95% length coverage were clustered, taking the longest element as representative.
- The classification was performed according to the following hierarchy suggested by IWSGC (http://wheat.pw.usda.gov/ITMI/Repeats/gene_annotation.pdf): Class, Subclass, Superfamily.

CERZOS
CONICET
U N S

Identification and annotation of LTR retrotransposons

- Scaffolds were scanned for LTR retrotransposons using LTR-FINDER and LTR_STRUC. The LTR-Finder program was used with default parameters with the following exceptions: the minimum LTR size was set to 100 and the minimum distance of LTRs (internal domain) was set at 1000 bp.

- The Arabidopsis thaliana , Brachypodium distachyon, Oryza sativa, Sorghum bicolor and Zea mays databases were used to predict the primer binding sites.

- The output sequences were manually checked to confirm the boundary of the LTRs and target site duplications

- The output candidate LTR retrotransposons were extracted from the scaffolds and manually inspected for incorrectly predicted sequences and to determine the precise boundaries of the retroelements.

- Candidate LTR retrotransposons were clustered using CD-HIT and further BLAST searched against MIPS in order to distinguish between LTR retrotransposons that belong to previously annotated families and non-annotated sequences.

- For each identified retrotransposon, both LTRs were aligned using ClustalX. Transposon-associated genes were identified with BLASTX searches on NCBI.

- Annotation of LTR retrotransposons was performed according to Wicker et al. (2007).

Table 1. Repetitive elements identified on the 4D chromosomal arms of *Triticum aestivum* (var. Chinese Spring). For each element class the number of elements (#), the length of the sequence occupied by these elements (length) and the percentage of the sequence that is covered by repetitive elements (%) are given.

Element class	4DS			4DL		
	#	Length	%	#	Length	%
Retroelements	20607	19344019	50.12	15107	14212381	52.61
DNA transposons	9835	5969965	15.47	6296	3030966	11.22
Small RNA	69	22419	0.06	55	14356	0.05
Satellites	21	2468	0.01	16	2622	0.01
Simple repeats	1106	47817	0.12	750	31119	0.12
Low complexity	1293	83621	0.22	1104	74281	0.27
Unclassified	1849	539049	1.40	1258	342679	1.27
Total	**34780**	**26009358**	**67.4**	**24586**	**17708404**	**65.55**

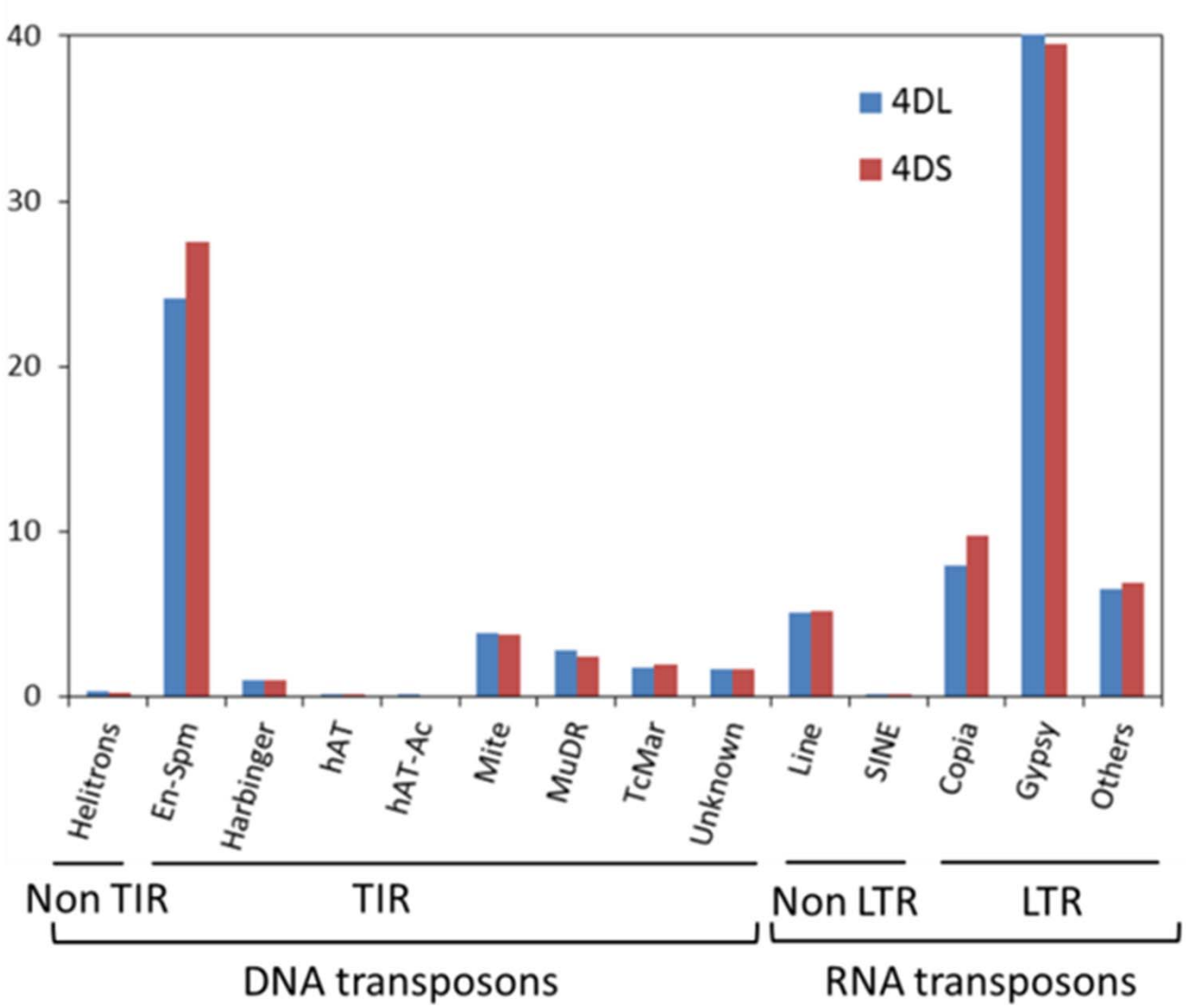

Figure 1. Distribution of DNA and RNA transposon superfamilies on 4D chromosome arms. Bars represent the percentage of each superfamily with respect to the total DNA or RNA transposable elements.

Table 1: Satellites identified in 4D chromosomal arms from *T. aestivum.* Satellites were classified according to the nucleotide composition of the repetitive motif.

4DS		4DL	
Secale_cereale_D1100	7	trep73	5
trep73	5	Secale_cereale_D1100	4
trep67	2	trep67	3
trep69	2	trep69	3
Beta_vulgaris_150	1	atr0015	1
Hordeum_vulgare_Crep1	1	Hordeum_vulgare_Crep1	1
Hordeum_vulgare_Crep2	1		
Poaceae_Af	1		
Triticum_aestivum_320	1		

Low complexity elements

4DS		4DL	
A-rich	199	A-rich	144
AT_rich	542	AT_rich	525
C-rich	49	C-rich	33
CT-rich	56	CT-rich	35
GA-rich	49	GA-rich	45
GC_rich	171	GC_rich	68
G-rich	42	G-rich	25
polypurine	1	polypyrimidine	1
polypyrimidine	4	T-rich	237
T-rich	200		

Microsatellites

4DS		4DL	
(TC)n	99	(TC)n	67
(CA)n	74	(CA)n	61
(GA)n	71	(TG)n	55
(TG)n	69	(GA)n	49
(TA)n	61	(TA)n	36
(CCG)n	56	(TTC)n	24
(CGG)n	38	(TTG)n	22
(GAA)n	32	(CCG)n	20
(CAA)n	23	(CAT)n	18
(TTG)n	22	(CAA)n	17
(GGA)n	21	(CGG)n	17
(CAT)n	20	(TAA)n	14
(CTG)n	20	(GAA)n	13
(TAA)n	19	(GGA)n	12
(TTA)n	19	(TTA)n	12
(TTC)n	19	(TGG)n	10
(ATG)n	18	trep1706	10
(TGG)n	15	(TAAAAA)n	8
trep1706	15	(TATG)n	8
(CCA)n	13	(TTTC)n	8
(TCC)n	13	(ATG)n	7
(CATG)n	10	(TAG)n	7
(TAAA)n	10	(TCTA)n	7
(TCTA)n	10	(TTTA)n	7
(CAG)n	9	(CAG)n	6
(CGA)n	9	(TCC)n	6
(TTTTA)n	9	(TTTTA)n	6

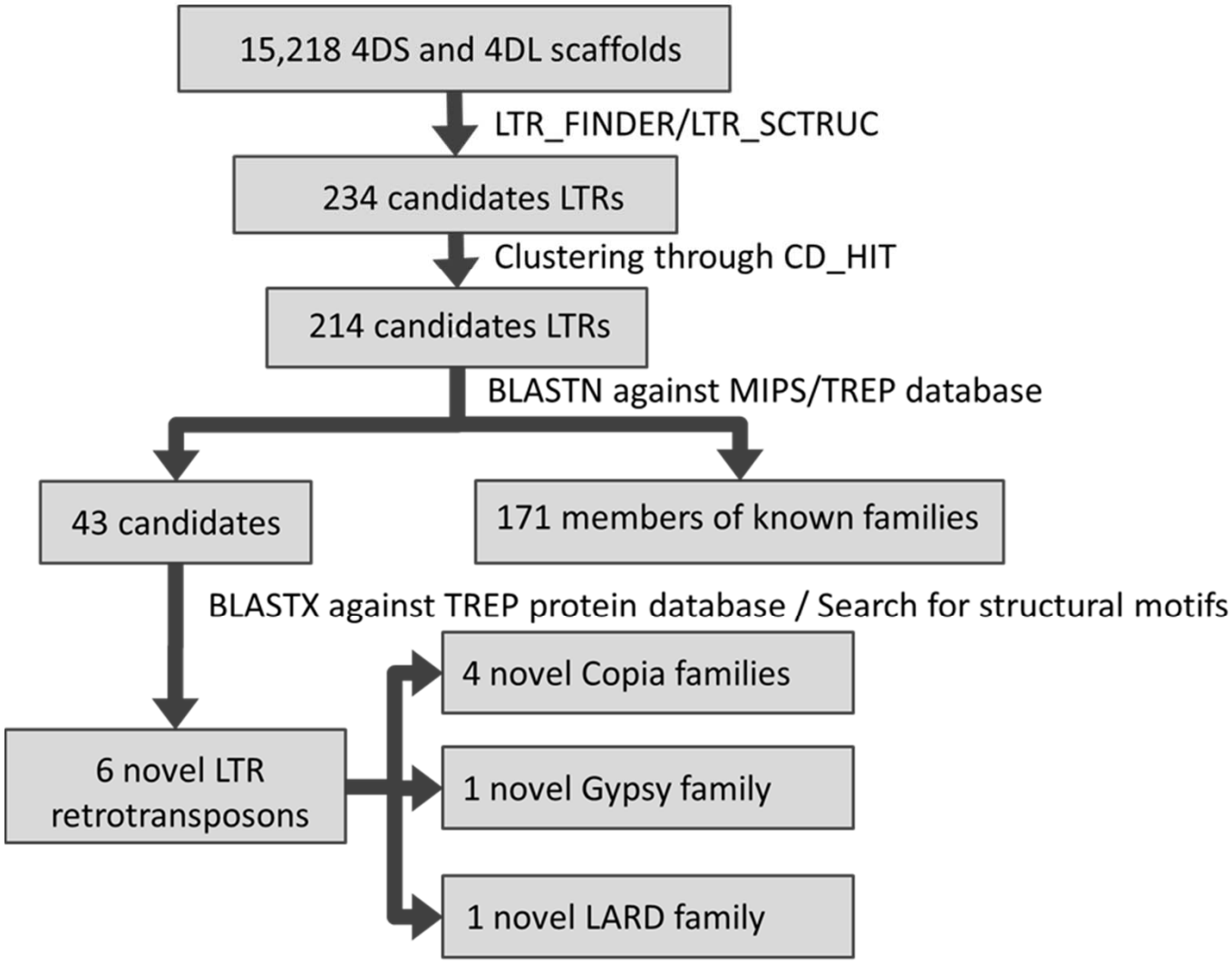

Figure 2. Annotation of novel LTR retrotransposons. The procedure was performed according to Wicker et al. (2007).

Description of the 6 LTR retrotransposon candidates identified on 4D chromosome scaffolds. The number of genomic repetitions for each candidate LTR retrotransposon was estimated by searching against the *T. aestivum* chromosome arm contigs deposited in the URGI database (# in genome). LTR: Long terminal repeat; TSD: target site duplication; PBS: primer binding site; PPT: polypurine tract. The last column indicates the presence (+) or absence (-) of retrotransposon proteins when BLAST searched against the TREP protein database.

SCAFFOLD	program	LTR retrotransposon size	# in genome	LTR region similarity	5'-LTR size	3'-LTR size	Insertion time (years x 10^{-9})	TSD	Orientation	PBS	PPT
4DSscaffold00165	STRUC / FINDER	5132	42	0.978	215	215	0,13	GAGGC	-	Lys_TT	GCCTCCCTCTTCCTC
4DSscaffold00708	FINDER	4597	40	0.977	131	131	0,05	-	-	SerTGA	CCATCTTCTTCCTCC
4DSscaffold00808	STRUC	2644	33	0.935	685	680	1,13	ACATT	+	AGTGGTATCAGAGCT GAGGTTGCTCA	ATAGCTTCGTTCCAAGAAGGAG GGGA
4DSscaffold01251	STRUC / FINDER	5279	21	0.969	321	321	0,13	CTGTC	-	SerGCT	TCTCCTGGTCCTCCC
4DLscaffold00928	FINDER	4842	757	0.942	138	139	0,86	-	+	MetCAT	GATACTGCGGGGGGA
4DLscaffold06475	FINDER	1898	121	0.960	375	375	0,18	-	-	MetCAT	TCATCCTCTCGCCCT

BLASTX analysis of the proteins encoded by the candidate LTR retrotransposons. Nucleotide scaffold sequences were BLAST searched against the TREP protein database.

SCAFFOLD	TREP protein code[1]	LTR retrotransposon associated[2]	Identity[3]	Conservative substitutions[4]	Coverage[5]
4DSscaffold00165	PTREP238 (1515 aa)	TREP3154 Copia, RLC_Olivia_42j2-1	64%	78%	84%
4DSscaffold00708	PTREP238 (1515 aa)	TREP3154 Copia, RLC_Olivia_42j2-1	54%	67%	75%
4DSscaffold00808	PTREP63 (1520 aa)	TREP98 Gypsy, RLG_Cereba_AY040832-1	32%	51%	22%
4DSscaffold01251	PTREP249 (1536 aa)	TREP3203 Gypsy, RLG_Latidu_10k23-1	32%	49%	93%
4DLscaffold00928	PTREP120 (1121 aa)	TREP2012 Copia, RLC_Zenia_AY853252-1	58%	74%	85%
4DLscaffold06475	PTREP64 (1717 aa)	TREP99 Gypsy,RLG_Cereba_AY040832-2	37%	58%	20%

Table references:

[1]: code of the protein that showed the highest identity to the scaffold. Its length is indicated in parenthesis; [2]: the code and name of the retrotransposon associated with the mentioned proteins; [3]: percentage of identity of the alignments; [4]: percentage of conservative substitutions, i.e., the aligned amino acids are not identical but both side chains have similar biochemical properties. [5] percentage of the protein sequences that aligned with the scaffold sequence.

SCAFFOLD	4DSscaffold00165	4DSscaffold00708	4DSscaffold00808	4DLscaffold00928
Family	Jose	Viviana	Gabriel	Facundo
Superfamily	Copia	Copia	Copia	Copia
Class	Retrotransposon	Retrotransposon	Retrotransposon	Retrotransposon
Order	LTR retrotransposon	LTR retrotransposon	LTR retrotransposon	LTR retrotransposon
Insertion	RLC_Jose_4DSscaffold00165-1	RLC_Viviana_4DSscaffold00708-1	RLC_Gabriel_4DSscaffold00808_1	RLC_Facundo_4DLscaffold00928-1
Structural description	Autonomous retrotransposon	Autonomous retrotransposon	Non autonomous retrotransposon	Autonomous retrotransposon
Other members (software)	RLC_Jose_AOCO010237191-1*	RLC_Viviana_CALP010001681-1* RLC_Viviana_AOCO010088545-1* RLC_Viviana_AOCO010057678-1*	RLC_Gabriel_AOCO010200744-1*	RLC_Facundo_AOCO010674073-1
Other members (wgs BLAST)	RLC_Jose_AOCO010066794-1* RLC_Jose_AOCO010066793-1 RLC_Jose_AOCO010233966-1 RLC_Jose_CALP010001044-1	RLC_Viviana_AOCO010608854-1 RLC_Viviana_AOCO010690661-1* RLC_Viviana_CALP010028816-1 RLC_Viviana_AOCO010256456-1 RLC_Viviana_CALP010306555-1 RLC_Viviana_CALP010308935-1 RLC_Viviana_CALP010045119-1 RLC_Viviana_CALP011468448-1 RLC_Viviana_CALP011191886-1 RLC_Viviana_CALP010103792-1 RLC_Viviana_CALP010067260-1 RLC_Viviana_CALP010446884-1 RLC_Viviana_CALP010317155-1 RLC_Viviana_CALP010163694-1 RLC_Viviana_AOCO010339138-1 RLC_Viviana_CALP010008428-1 RLC_Viviana_CALP010357964-1 RLC_Viviana_CALP011255795-1 RLC_Viviana_CALP011598660-1 RLC_Viviana_CALP010236408-1 RLC_Viviana_CALP011765918-1 RLC_Viviana_CALP010332330-1 RLC_Viviana_AOCO010608853-1 RLC_Viviana_CALP011001485-1 RLC_Viviana_CALP010093698-1 RLC_Viviana_CALP011112138-1 RLC_Viviana_CALP011000205-1 RLC_Viviana_CALP010736972-1 RLC_Viviana_CALP010121574-1 RLC_Viviana_CALP012884601-1 RLC_Viviana_CALP011662724-1 RLC_Viviana_CALP011659433-1 RLC_Viviana_CALP012348547-1 RLC_Viviana_CALP012055131-1 RLC_Viviana_CALP010644204-1 RLC_Viviana_CALP010464375-1 RLC_Viviana_CALP010224113-1 RLC_Viviana_CALP012501138-1 RLC_Viviana_CALP013294378-1 RLC_Viviana_AOCO010440397-1 RLC_Viviana_CALP011828547-1 RLC_Viviana_AOCO010057678-1 RLC_Viviana_CALP010165878-1 RLC_Viviana_CALP011654708-1	RLC_Gabriel_CALP010466105-1 RLC_Gabriel_CALP010124160-1 RLC_Gabriel_CALP010723531-1 RLC_Gabriel_CALP013409816-1 RLC_Gabriel_CALP012559367-1 RLC_Gabriel_CALP013361237-1 RLC_Gabriel_CALP010625057-1 RLC_Gabriel_CALP010054491-1 RLC_Gabriel_CALP010532956-1	RLC_Facundo_AOCO010191627-1 RLC_Facundo_AOCO010653612-1 RLC_Facundo_AOCO010317688-1 RLC_Facundo_AOCO010105934-1 RLC_Facundo_AOCO010339256-1 RLC_Facundo_CALP010435951-1 RLC_Facundo_AOCO010636834-1 RLC_Facundo_CALP010660128-1 RLC_Facundo_CALP010480293-1 RLC_Facundo_CALP010527190-1 RLC_Facundo_AOCO010325142-1 RLC_Facundo_AOCO010438938-1 RLC_Facundo_AOCO010392677-1 RLC_Facundo_AOCO010503571-1 RLC_Facundo_CALP010431080-1 RLC_Facundo_AOCO010379024-1 RLC_Facundo_AOCO010799006-1 RLC_Facundo_AOCO010285134-1 RLC_Facundo_AOCO010558470-1 RLC_Facundo_CALP010781314-1 RLC_Facundo_AOCO010215994-1 RLC_Facundo_CALP010775656-1 RLC_Facundo_AOCO010803562-1 RLC_Facundo_AOCO010864957-1 RLC_Facundo_AOCO010416663-1 RLC_Facundo_CALP010862251-1 RLC_Facundo_CALP010951044-1 RLC_Facundo_AEOM01280901-1 RLC_Facundo_CALP010877442-1 RLC_Facundo_AOCO010664168-1 RLC_Facundo_AOCO010665250-1 RLC_Facundo_AOCO010483387-1 RLC_Facundo_CALP011080528-1 RLC_Facundo_CALP010889023-1 RLC_Facundo_AOCO010392678-1 RLC_Facundo_AOCO010021572-1 RLC_Facundo_AOCO010121158-1 RLC_Facundo_AOCO010383958-1 RLC_Facundo_CALP010778523-1 RLC_Facundo_CALP010981426-1 RLC_Facundo_AOCO010222056-1 RLC_Facundo_CALP011213595-1 RLC_Facundo_AOCO010558476-1 RLC_Facundo_CALP011003546-1 RLC_Facundo_AOCO010239865-1

Description of the internal structure of the novel LTR retrotransposons.

SCAFFOLD	4DSscaffold01251	4DLscaffold06475
Family	Francisco	Victoria
Superfamily	Gypsy	Gypsy
Class	Retrotransposon	Retrotransposon
Order	LTR retrotransposon	LTR retrotransposon
Insertion	RLG_Francisco_4DSscaffold01251-1*	XXX_Victoria_scaffold06475-1
Structural description	Autonomous retrotransposon	Non Autonomous retrotransposon (LARD)
Other members (software)	RLG_Francisco_AOCO010454749-1* RLG_Francisco_AOCO010144477-1*	See table 5
Other members (wgs BLAST)	RLG_Francisco_CALP010003649-1* RLG_Francisco_CALP010061300-1 RLG_Francisco_CALP010153829-1 RLG_Francisco_CALP010289309-1 RLG_Francisco_CALP010532755-1 RLG_Francisco_AEOM01028590-1 RLG_Francisco_CALP010296358-1 RLG_Francisco_CALP010313935-1 RLG_Francisco_AEOM01048149-1 *RLG_Francisco_CALP010061300-1* RLG_Francisco_CALP010594168-1 RLG_Francisco_CALP010450630-1 RLG_Francisco_CALP010619063-1 RLG_Francisco_CALP010691093-1 RLG_Francisco_CALP010903443-1 RLG_Francisco_CALP011534613-1 RLG_Francisco_CALP010535643-1 RLG_Francisco_CALP010635319-1 RLG_Francisco_CALP011637439-1 RLG_Francisco_CALP012062839-1 RLG_Francisco_CALP011856515-1 RLG_Francisco_CALP011637278-1 RLG_Francisco_AEOM01071161-1 RLG_Francisco_CALP012632193-1 RLG_Francisco_CALP012053594-1 RLG_Francisco_CALP013813545-1 RLG_Francisco_CALP013495848-1 RLG_Francisco_AEOM01094980-1 RLG_Francisco_CALP013685092-1 RLG_Francisco_CALP010869653-1 RLG_Francisco_CALP010090108-1 RLG_Francisco_CALP013648311-1	RLx_Victoria_CALP010064643-1 RLx_Victoria_AOCO010372201-1 RLx_Victoria_AOCO010278508_1

CONCLUSIONS

The present work reveals the complete landscape of the repeatome of this chromosome.

We identified six novel LTR retrotransposon families

The percentage of repetitive sequences reported here is lower than the currently accepted value for the whole genome. This suggests that the chromosome composition is not constant across the genome. By contrast, chromosome 4D had more repetitive sequences than that reported in the diploid contributors to hexaploid wheat, which probably reflects the extensive expansion due to the stress caused by the hybridization events that led to modern wheat.

In accordance with results obtained for other grasses, CACTA/En-Spm and Gypsy were the most abundant DNA transposons and retrotransposons, respectively, suggestive of their conserved roles in genome regulation.

In spite of the extensive research performed in Triticeae genomes and the high number of reported elements, the fact that six new elements could be identified in a single-chromosome analysis indicates that new families probably remain to be described.

Session Name: Poster Session 3: Toward the Whole Genome Sequence
Discussion Time & Date: 19:00-20:00
Monday, September 9, 2013
Place: F203+F205+Foyer, Pacifico Yokohama
Program Number: P3-2