Chromosome 4D survey sequencing analysis: current progress towards the understanding of its structural organization

Helguera M[1], Rivarola M[2] , Martis M[3], Vanzetti L[1], Garbus I[4], Leroy P[5], Clavijo B[6], Romero JR[4], Gonzalez S[2], Tabbita F[7], Bonafede M[7], Cativelli M[7], Valarik M[8], Simkova H[8], Wright J, Cáccamo M, Dolezel J[8], Feuillet C[5], Mayer Klaus[3], Tranquilli G[7], Paniego N[2], Echenique V[4]

[1]Instituto Nacional de Tecnología Agropecuaria/EEA Marcos Juárez Ruta 12 S/N 2580 Marcos Juárez Argentina,

[2]Instituto Nacional de Tecnología Agropecuaria/Instituto de Biotecnología 1686 Hurlingham, Bs As Argentina,

[3]Munich Republic, Information Center for Protein Sequences/Institute for Bioinformatics and Systems Biology Helmholtz Zentrum Munich German Research Center for Environmental Health 85764 Neuherberg Germany,

[4]Dpto. Agronomía and CERZOS/CONICET CCT Bahía Blanca, 8000 Bahía Blanca Argentina,

[5]UMR INRA-UBP 1095 Domaine de Crouelle 234, Avenue de Brézet 63100 Clermont-Ferrand, France,

[6]The Genome Analysis Centre Norwich Research Park Colney Norwich, NR4 7UH UK,

[7]Instituto Nacional de Tecnología Agropecuaria/Instituto Recursos Biológicos 1686 Hurlingham Buenos Aires Argentina,

[8]Institute of Experimental Botany Sokolovska 6 CZ-77200 Olomouc Czech Republic.

Int. Wheat Genome Sequencing Consortium Workshop, Yokohama (Japan), 7 September 2013

The sequence

✓Two batches of 4DS and 4DL lyophilized DNA provided by J. Dolezel group

✓a first batch of 4DS and 4DL DNAs was shotgun sequenced using a 454 NGS platform (4 runs, SE reads).

✓a second batch of 4DS and 4DL DNA was shotgun sequenced in 2 runs, LMP reads 3kb long.

SE reads, statistics

	4DL	4DS	Total	Run
Reads	785724	748009	1533733	1
Bases	260746095	251739388	512485484	
Length Average	332	336	334	
Reads	781863	776506	1558369	2
Bases	307435063	304359662	611794725	
Length Average	393	392	392	
Reads	771414	737454	1508868	3
Bases	305146131	288067556	593213687	
Length Average	395	391	393	
Reads	835874	719510	1555384	4
Bases	330794673	287923022	618717695	
Length Average	396	400	398	
Reads	793718,75	745369,75	1539088,5	Average
Bases	301030491	283022407	584052898	
Length Average	379	379,75	379,25	
Reads	3174875	2981479	6156354	Total
Bases	1204121962	1132089628	2336211591	

data equivalent to ≈ 3.6x chromosome coverage (4.9x+2.9)

LMP reads, statistics

	4DL	4DS	Total	Run
Reads		1.435.833	1.435.833	1
Bases		534.113.457	534.113.457	
Length Average		372	372	
Reads	1.385.664		1.385.664	2
Bases	508.862.344		508.862.344	
Length Average	367,23		367,23	
Reads			2.821.497	Total
Bases			1.042.975.801	

data equivalent to ≈ 1.5x chromosome coverage (2.3x+1.2x)

SE+LMP ≈ 5.1x chromosome coverage (7.2x+4.1x)

Finding the best assemblies

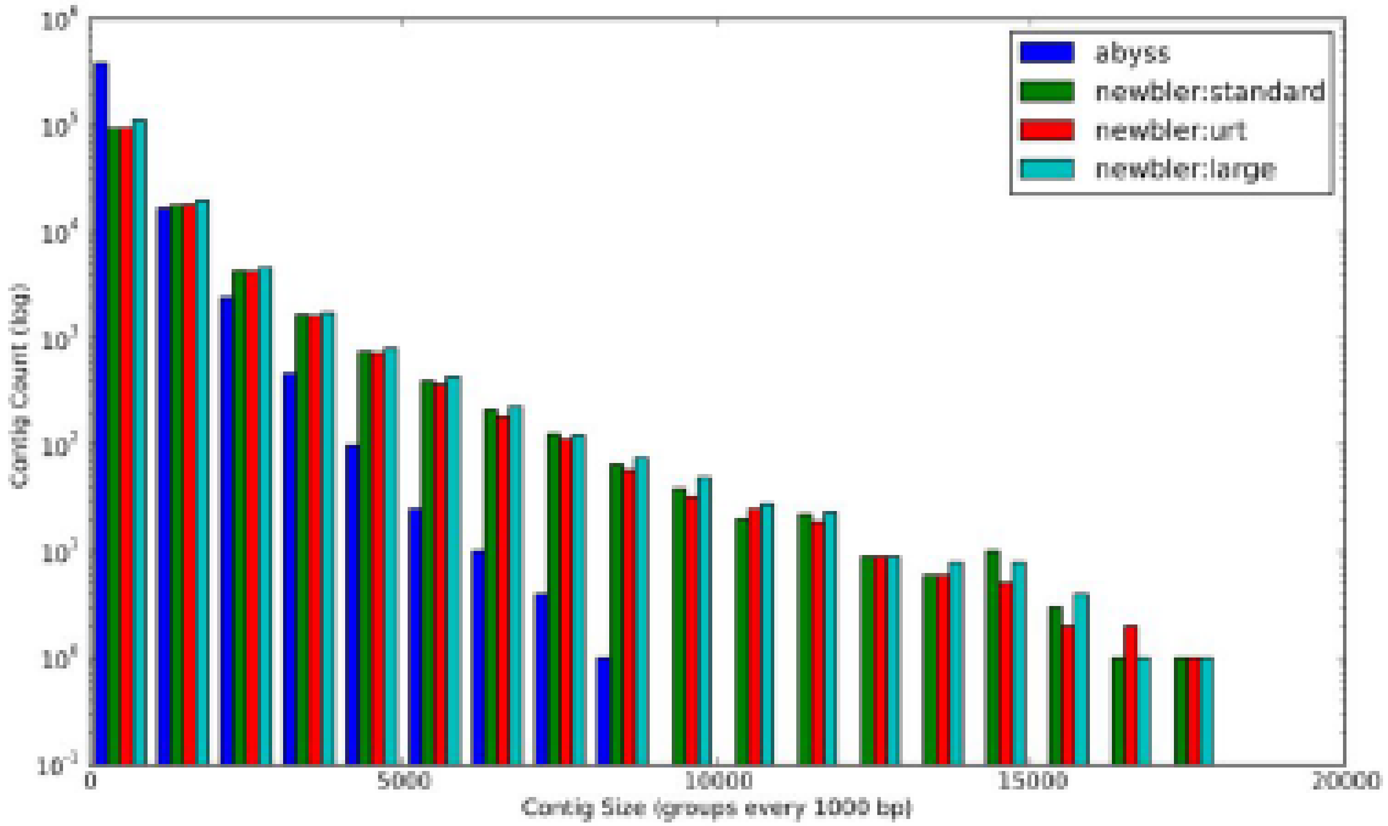


Figure 2: Contig size histogram for 4DS assemblies

best results with Newbler large in building of contigs and scaffolds

Statistics of obtained contigs and scaffolds

	contigs		scaffolds	
Measures	**4DS**	**4DL**	**4DS**	**4DL**
Total sequences	140607	204259	8141	7077
Total base pairs [Mbp]	103	120	38	27
Smallest length [bp]	100	100	1369	1530
Largest length [bp]	26237	21080	47795	22479
Average contig size [bp]	733.7	585.6	4741	3817
N50 [bp]	1132	807	5517	3998

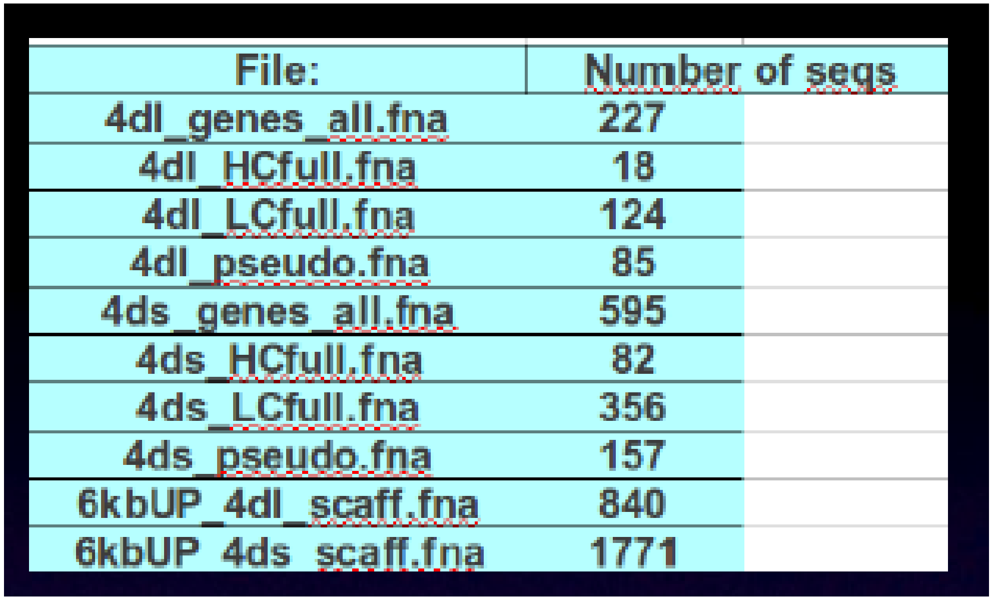
Validating scaffolds with diverse sources of data

Source of data	Total Bp/markers	Matched Scaffolds
Ae tauschii 4D anchored scaffolds (Jia et al. 2013)	180Mb	7410 (49%)
Ae t 4D anchored markers (Luo et al. 2013)	7Mb	
Binned 4D ESTs (Miftahudin et al. 2004)	603 ESTs	
Ae t total genome scaffolds (ex. 4D anchored)	4.05Gb	7700 (50%)
Ae t anchored scaf. in chr. diff 4D	(1.72Gb)	(1844 -12%)
other		108 (1%)

BLASTN with word-size=50 and e-value < 10^{-10}

Gene identification and annotation

1771= 18Mb (4DS) + 840= 6.8Mb (4DL) scaffolds larger than 6kb were analyzed with TriAnnot platform (thanks P. Leroy!)

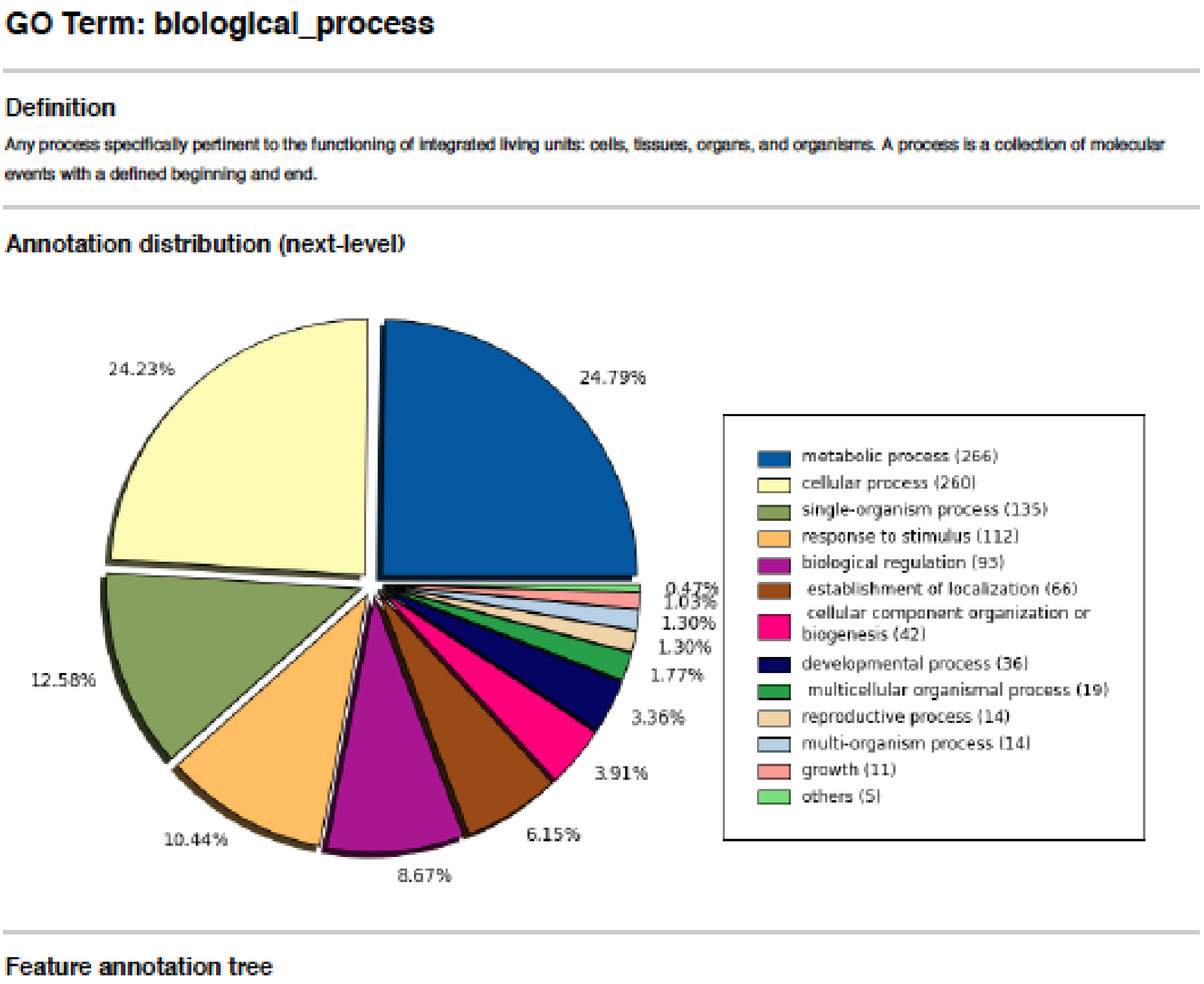


File:	Number of seqs
4dl_genes_all.fna	227
4dl_HCfull.fna	18
4dl_LCfull.fna	124
4dl_pseudo.fna	85
4ds_genes_all.fna	595
4ds_HCfull.fna	82
4ds_LCfull.fna	356
4ds_pseudo.fna	157
6kbUP_4dl_scaff.fna	840
6kbUP_4ds_scaff.fna	1771

full hit coverage >70% , pseudo hit cov. 70-50%
HC clear evidences start, stop, in-exon junctions

INTA

CONICET



GO Term: biological_process

Definition

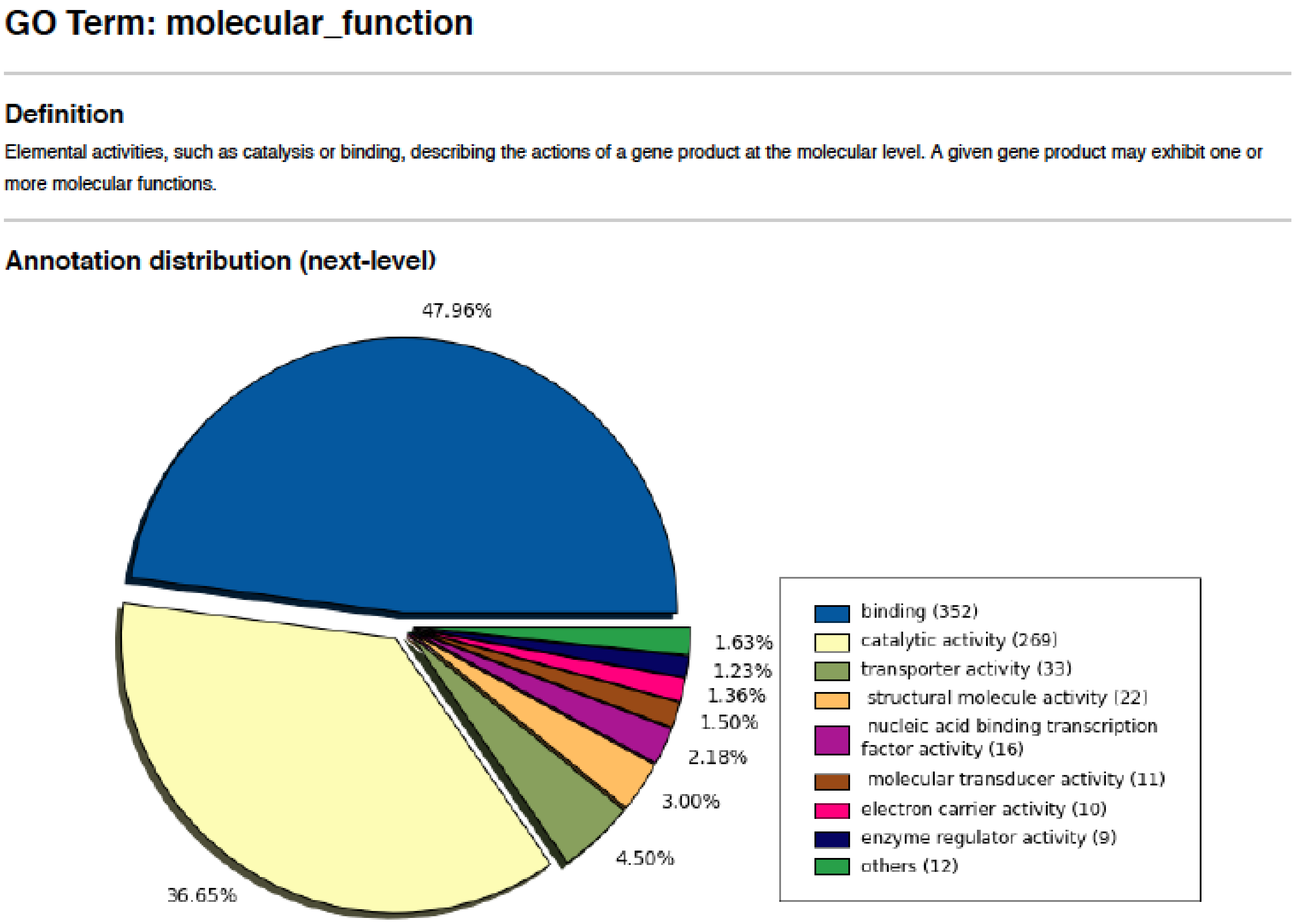
Any process specifically pertinent to the functioning of integrated living units: cells, tissues, organs, and organisms. A process is a collection of molecular events with a defined beginning and end.

Annotation distribution (next-level)

24.79%
24.23%
12.58%
10.44%
8.67%
6.15%
3.91%
3.36%
1.77%
1.30%
1.30%
1.03%
0.47%
metabolic process (266)
cellular process (260)
single-organism process (135)
response to stimulus (112)
biological regulation (93)
establishment of localization (66)
cellular component organization or biogenesis (42)
developmental process (36)
multicellular organismal process (19)
reproductive process (14)
multi-organism process (14)
growth (11)
others (5)

Feature annotation tree

http://bioinformatica.inta.gov.ar/ATGCtrigo/ontology/termtree/GO/0008150/2



GO Term: molecular_function

Definition

Elemental activities, such as catalysis or binding, describing the actions of a gene product at the molecular level. A given gene product may exhibit one or more molecular functions.

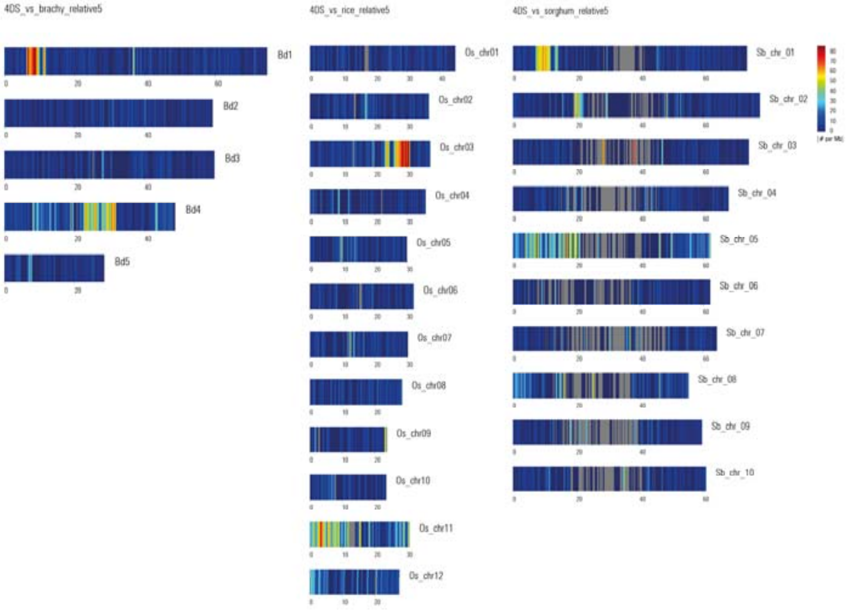
Annotation distribution (next-level)

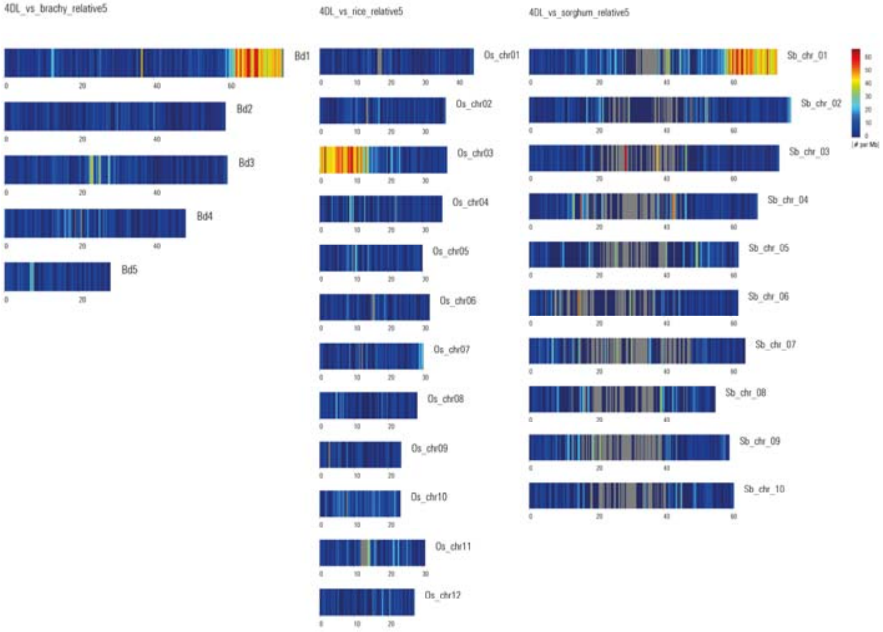
47.96%
36.65%
4.50%
3.00%
2.18%
1.50%
1.36%
1.23%
1.63%
binding (352)
catalytic activity (269)
transporter activity (33)
structural molecule activity (22)
nucleic acid binding transcription factor activity (16)
molecular transducer activity (11)
electron carrier activity (10)
enzyme regulator activity (9)
others (12)

http://bioinformatica.inta.gov.ar/ATGCtrigo/ontology/termtree/GO/0003674/2

GO Term: cellular_component

Assessment of syntenic regions among wheat chromosome 4D and reference genomes (Mihaela Martis, MIPS, Germany)





4DS **4DL**

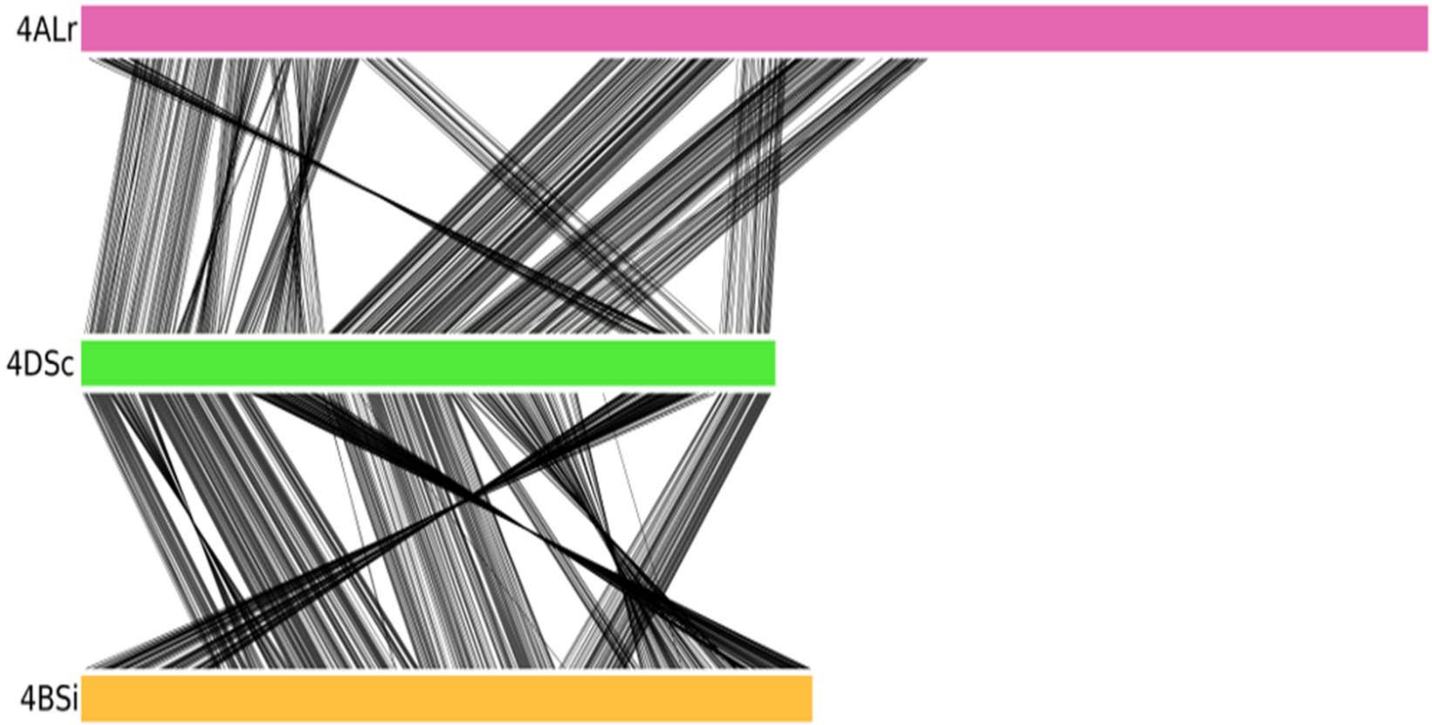
A virtual map (Mihaela Martis, MIPS, Germany)

The “GenomeZipper” was used to structure and order genes identified by wheat 4DS and 4DL contigs on the basis of collinearity to the reference grass genomes

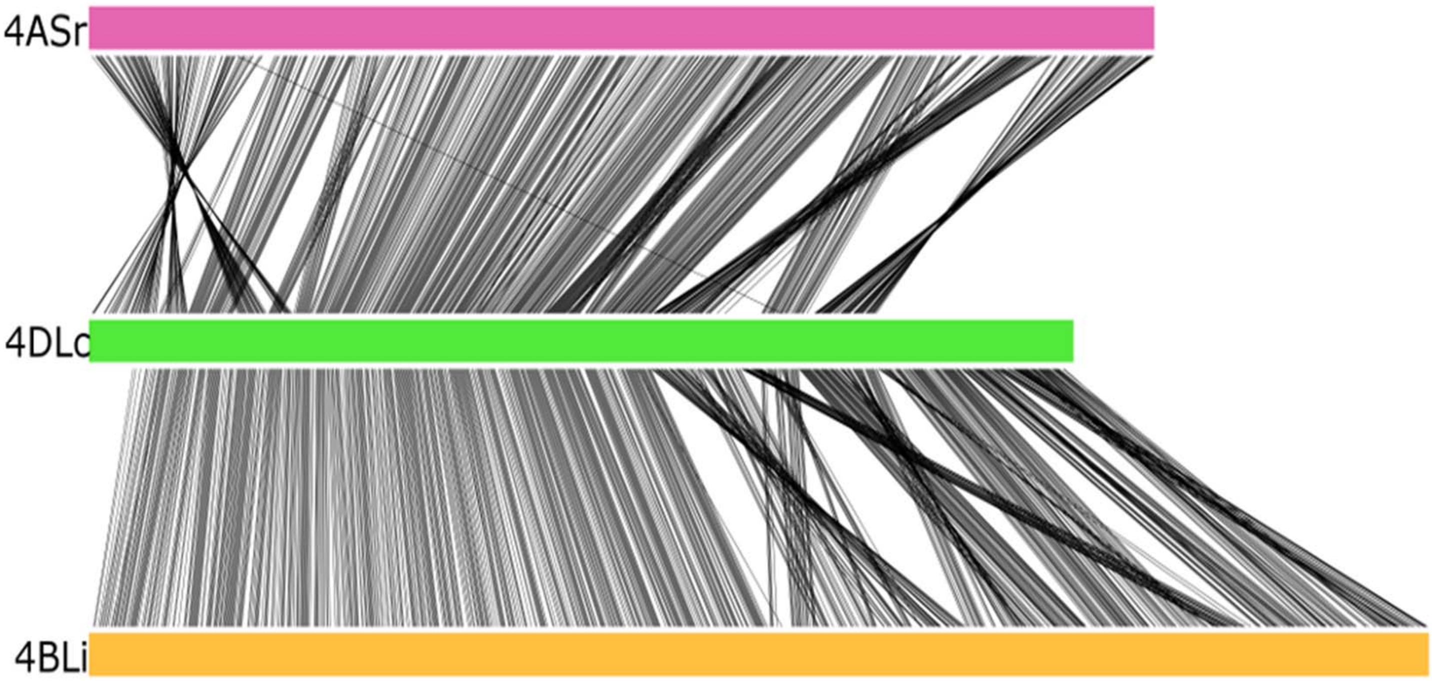
Data sets	4DS	4DL
No. of marker	16	16
No. of wheat flcDNAs	101	174
No. of contigs	1407	1918
No. of matched wheat ESTs	289	409
No. of Brachypodium genes	554	795
No. of rice genes	521	701
No. of sorghum genes	433	759
No. of predicted genes	270	111
No. of gene loci associated with genes from reference genomes	892	1081
Total no. of gene loci	902	1092

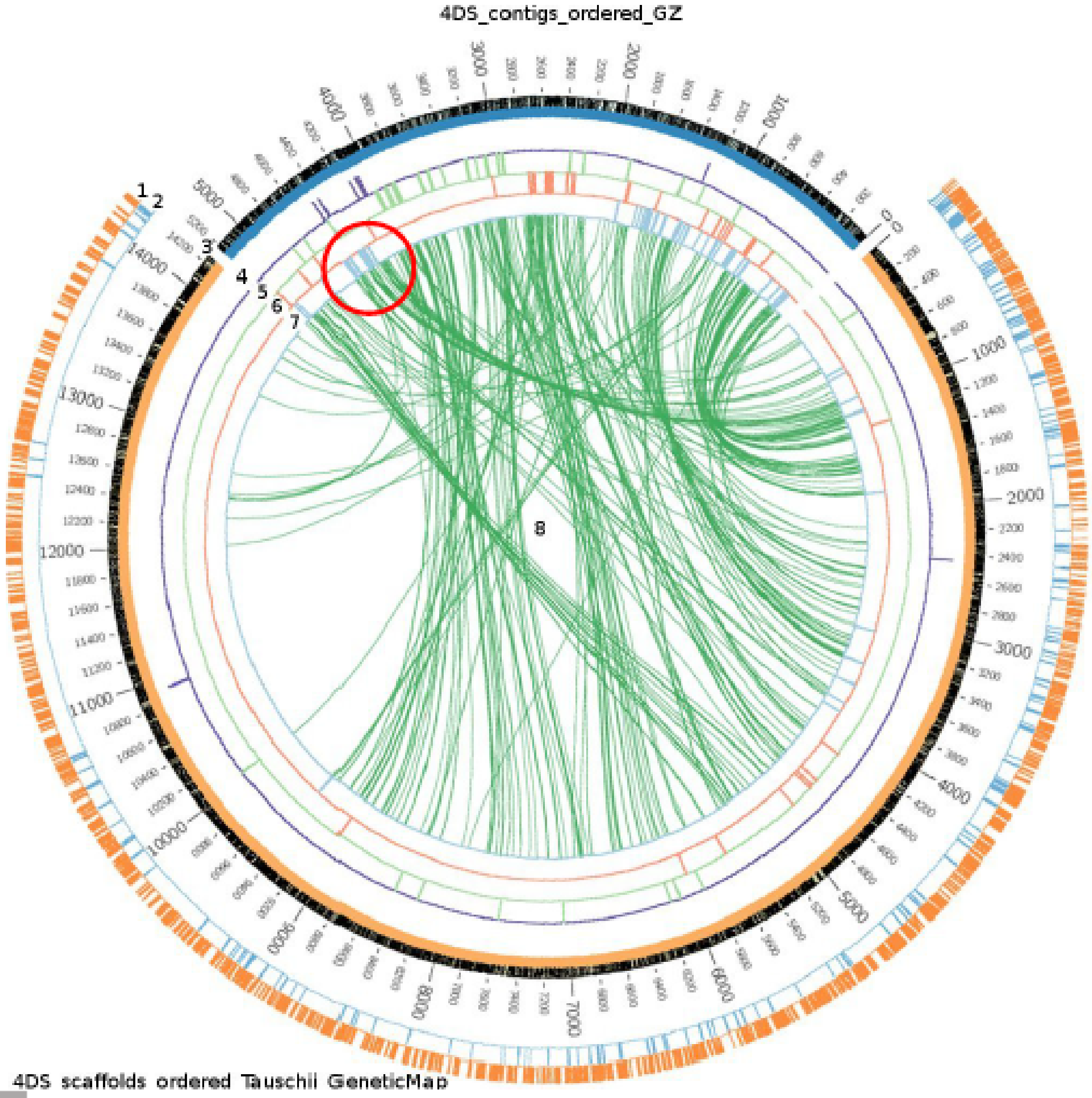
- ✓ 9000 syntenic genes in 4A using 454 reads
- ✓ 2900 genes in 4D Ae tauschii

Virtual gene map of 4DSc vs. virtual gene maps of $4A_r$ and $4B_i$

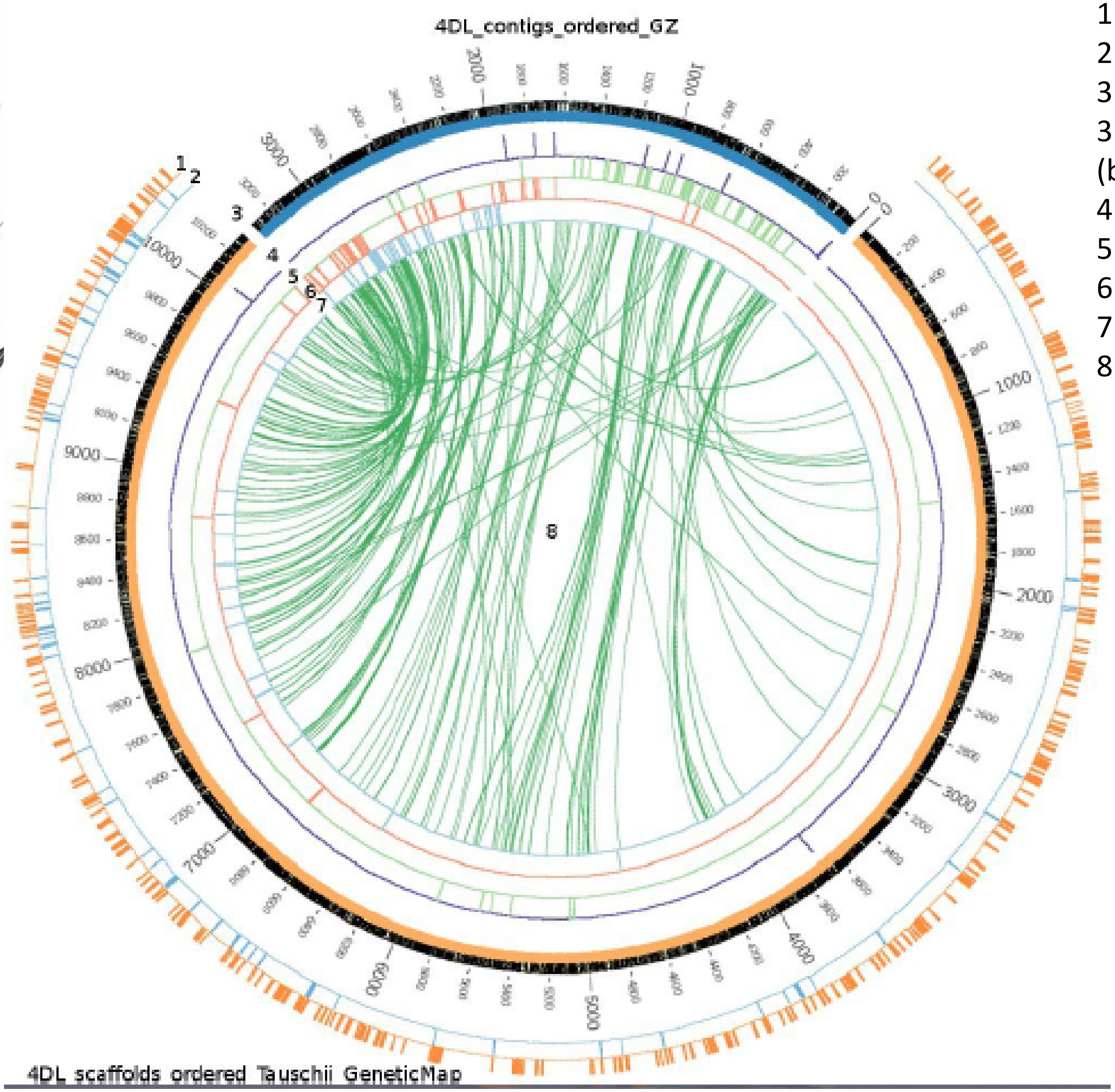


Virtual gene map of 4DLc vs. virtual gene maps of 4A$_r$ and 4B$_i$





1 repeats annotated by TREPplus DB
2 TriAnnot annotated genes.
3 (orange) AeT 4DS scaffolds
3 (blue) 4DS GZ contigs
(black/white line sizes in scale)
4 (blue) ESTs in 4DS1-0.53 del bin
5 (green) ESTs in 4DS1-0.53-0.67
6 (red) ESTs in 4DS3-0.67-0.82
7 (blue) ESTs in 4DS2-0.82-1.00
8 (green lines) matched contigs



1 repeats annotated by TREPplus DB
2 TriAnnot annotated genes.
3 (orange) AeT 4DS scaffolds
3 (blue) 4DL GZ contigs
(black/white line sizes in scale)
4 (blue) ESTs in 4DL9-0.31 del bin
5 (green) ESTs in 4DL9-0.31-0.56
6 (red) ESTs in 4DL13-0.56-0.71
7 (blue) ESTs in 4DL12-0.71-1.00
8 (green lines) matched contigs

thank you!

INTA

Rivarola M, Vanzetti L, Gonzalez S, Tabbita F, Bonafede M, Cativelli M, Tranquilli G, Paniego N, Helguera M

CONICET

Garbus I, Romero JR, Echenique V

Valarik M, Simkova H, Dolezel J

mips
munich information center
for protein sequence

Martis M, Mayer K

INRA

Leroy P, Feuillet C

TGAC
The Genome Analysis Centre™

Clavijo B, Wright J, Cáccamo M

questions: helguera.marcelo@inta.gob.ar, echenque@criba.edu.ar