# Physical map of the wheat chromosome arm 3DS

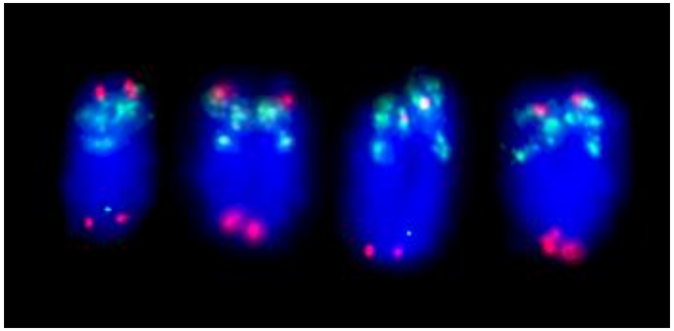
**Jan Bartoš**

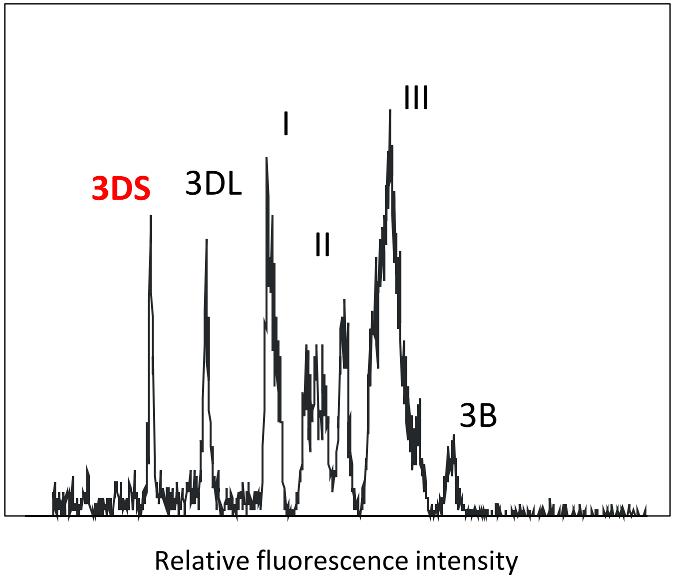
Centre of  Region Haná for Biotechnological and Agricultural Research
Institute of Experimental Botany
Šlechtitelů 31
783 71 Olomouc - Holice

# Wheat chromosome arm 3DS

**Chromosome arm 3DS characteristics**

- Estimated size 321 Mbp
- Less than 2% of wheat genome
- Low level of polymorphism in D genome
- Important genes
  - Ph2 locus (pairing homologs)
  - Yr49 (yellow rust resistance)

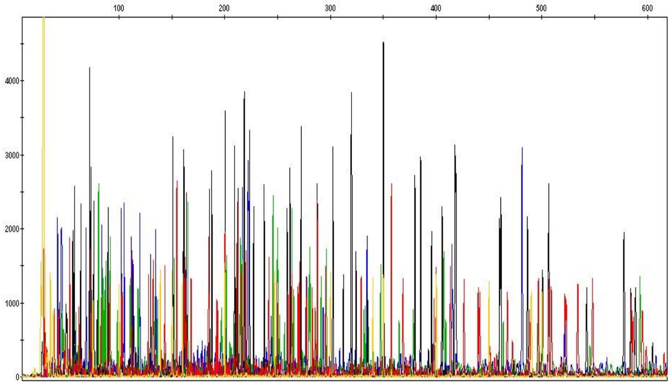




# 3DS physical map

**BAC library and fingerprinting**

- 36,864 clones
- 11x chromosome coverage
- 27,880 useful fingerprints



**Automated assembly**

- FPC based
- Following IWGSC rules
- Cut-off: 1e-75 => 1e-45

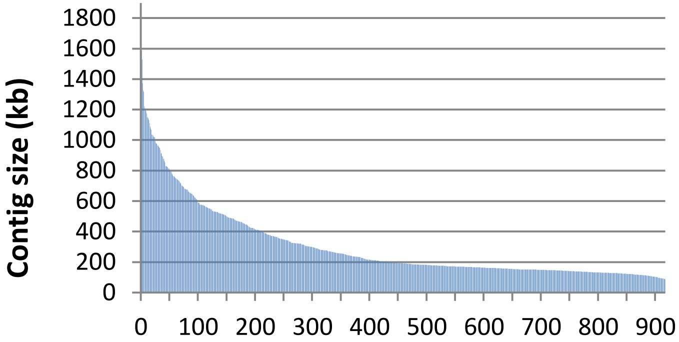
| | Automated assembly |
|---|---|
| Cut-off | 1e-45 |
| Contigs | 1,360 |
| Q-clones | 282 |
| Assembly length (Mb) | 310 (97%) |
| Longest contig (kb) | 1,092 |
| N50 contig length (kb) | 244 |
| MTP (clones) | 3,823 |

# 3DS physical map

**Manual assembly**

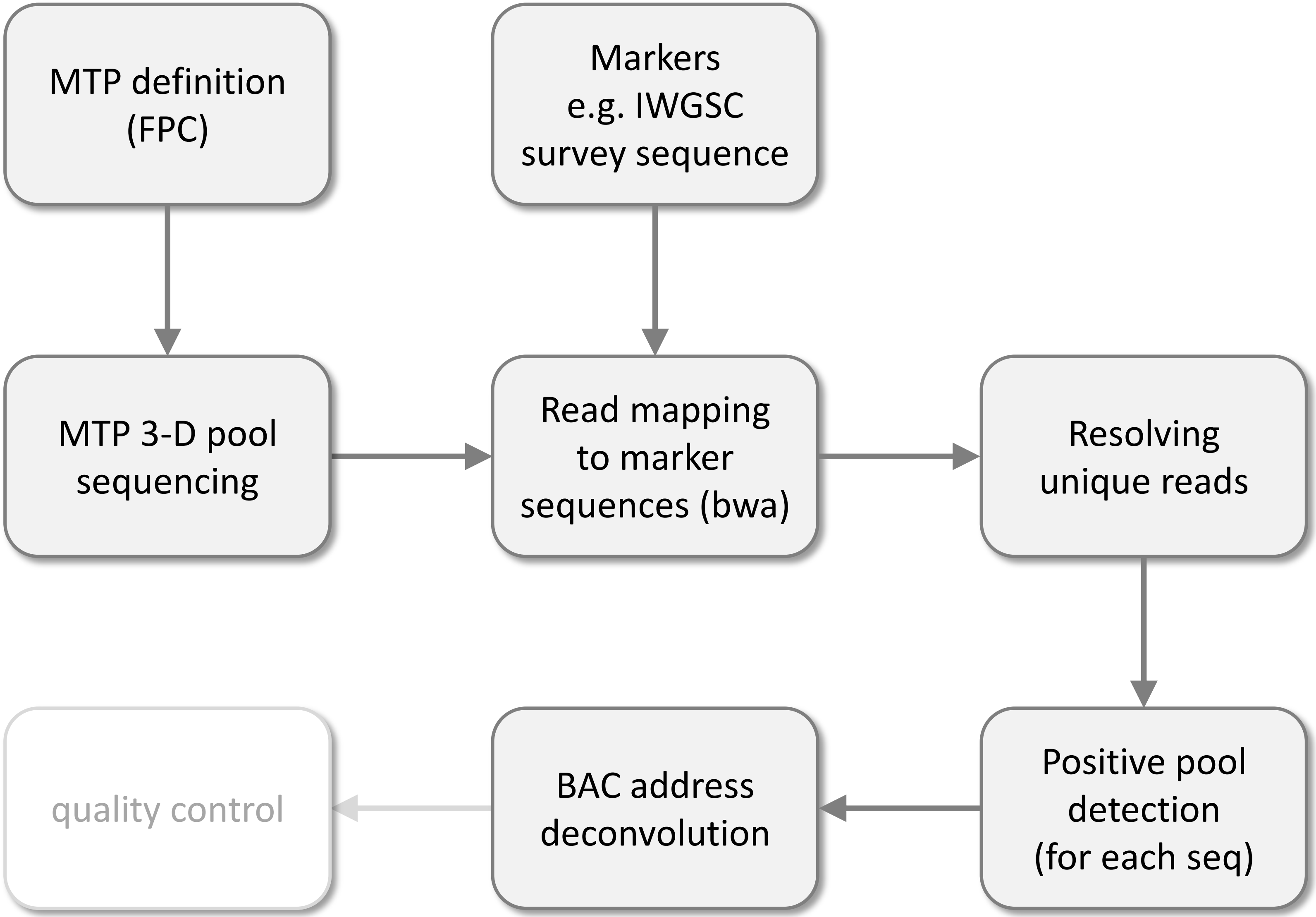
- FPC based
- Cut-off: 1e-45 => 1e-15
- Correction using LTC

**Distribution of contig sizes**



| | Automated assembly | Manual assembly |
|---|---|---|
| Cutoff | 1e-45 | 1e-15 |
| Contigs | 1,360 | 918 |
| Q-clones | 282 | 499 |
| Assembly length (Mb) | 310 (97%) | 278 (87%) |
| Longest contig (kb) | 1,092 | 1,870 |
| N50 contig length (kb) | 244 | 412 |
| MTP (clones) | 3,823 | --- |

# *In silico* anchoring workflow



MTP definition (FPC)

Markers e.g. IWGSC survey sequence

MTP 3-D pool sequencing

Read mapping to marker sequences (bwa)

Resolving unique reads

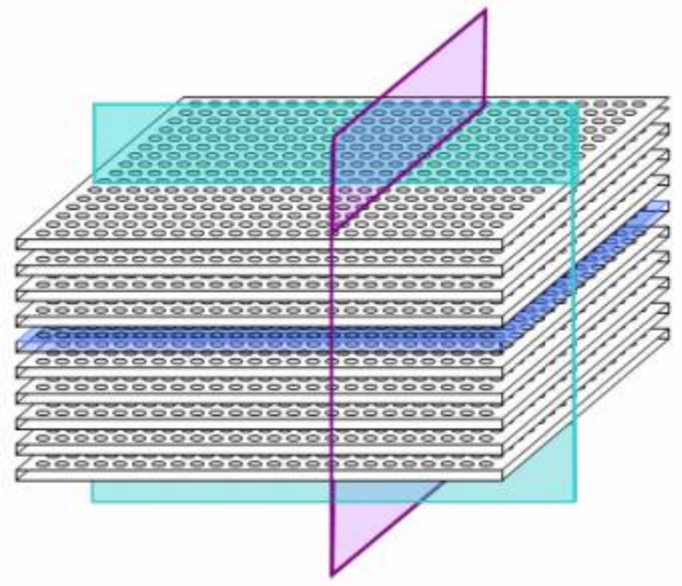
Positive pool detection (for each seq)

BAC address deconvolution

quality control

# MTP pool sequencing

- MTP 3,823 clones
- Fifty 3D MTP pools (10 plates, 16 rows, 24 columns)
- Pools of each dimensions sequenced as indexed libraries on Illumina HiSEQ
- 367,907,030 reads (2 x 100 bp)
- Unequal pool coverage
- 6 – 166x (mean 35x; median 23.5x)



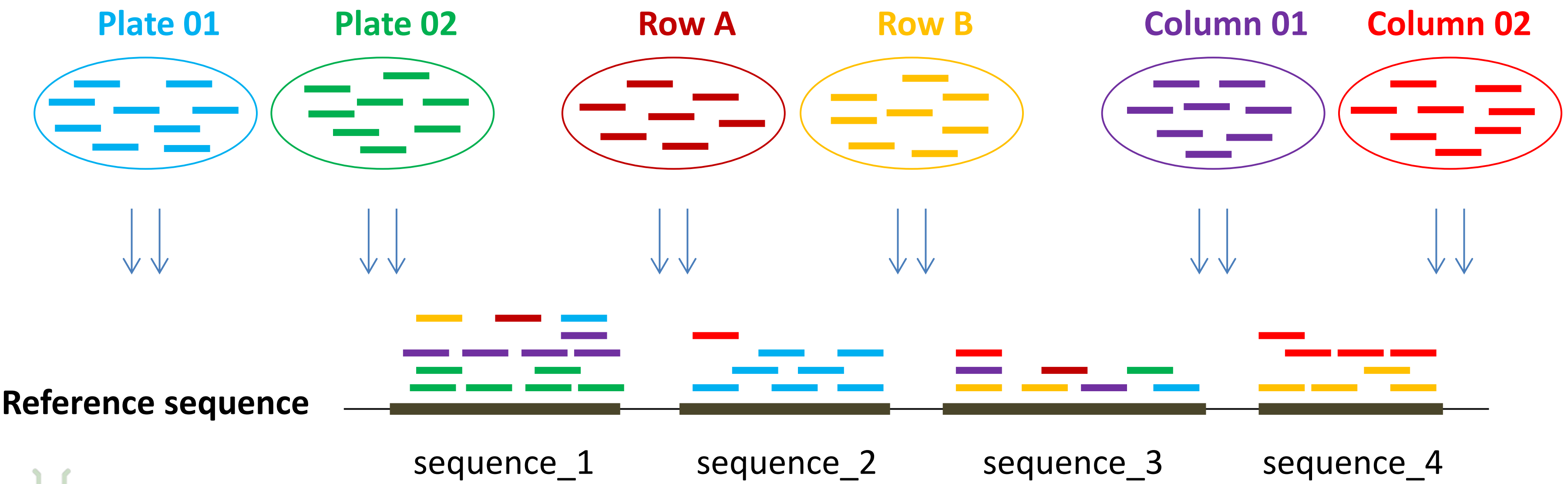
# Read mapping to marker sequences

**Reference sequence**

- IWGSC 3DS survey sequence
- 314,944 sequences
- Total length 145,374,274 bp (45% of chromosome arm)

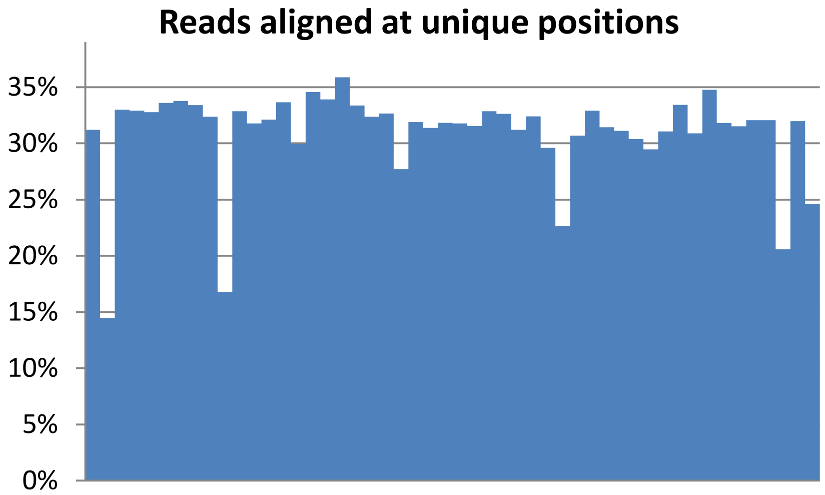
**Read alignment**

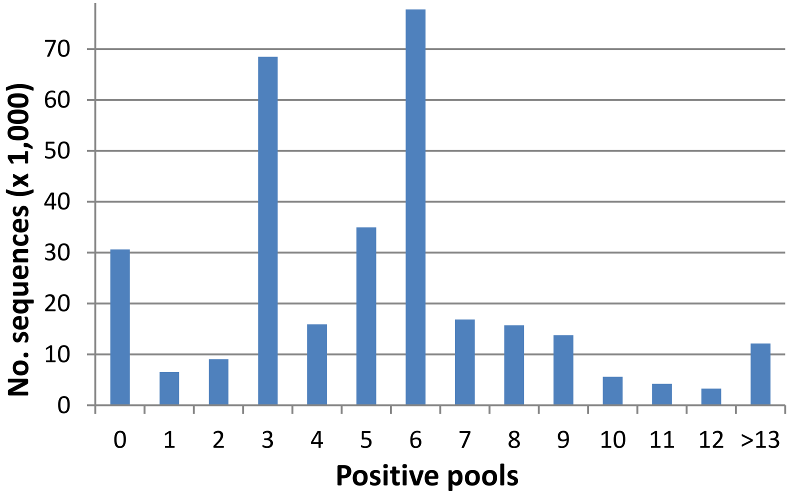
- Using Burrows-Wheeler aligner
- Reads of each pool renamed to track their origin
- Maximal coverage 30x/pool



# Positive pool identification

- Only reads mapped to unique position with no mismatch used
- **Positive pools identified individually for each sequence**
- Aligned reads counted for each pool
- Number of aligned reads normalized by pool coverage
- Pool positive if normalized read number ≥ 20% of average for pools with at least one aligned read
- **At least 1 plate, 1 row and 1 column pool for 258,146 seqs**



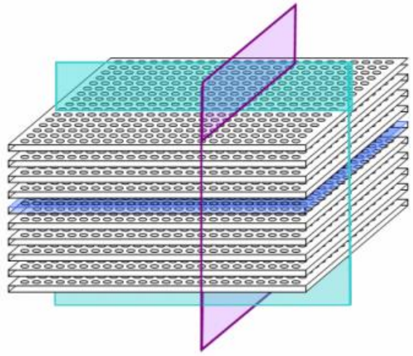


# BAC address deconvolution

**1) One positive pool in each dimension (1 – 1 – 1)**

- Direct BAC clone identification

  *Plate07 – RowC – Column18 --> TaaCsp3DShA_0055B07*

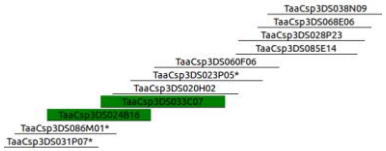


**2) Multiple positive pools in at least one dimension (e.g. 2 – 2 – 2)**

- Identification of all candidate BAC clones

a) Check contig information for all clones

b) Check possible overlap in case of end clones

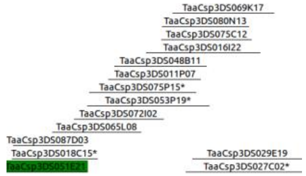


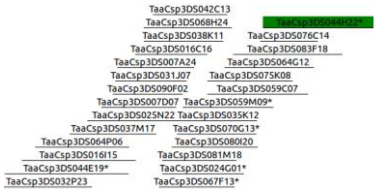
**3) Sequence not anchored if:**

a) Positive pool is missing for plate, row or column.

b) Five or more positive pools in at least one dimension.

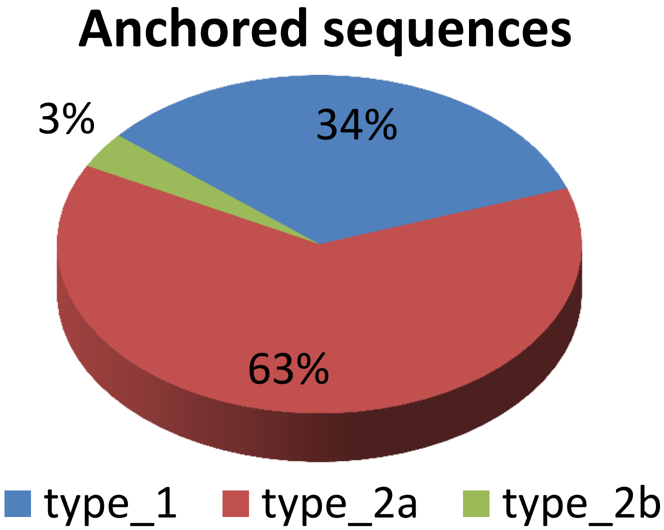
c) No positive clone was found in step 2).

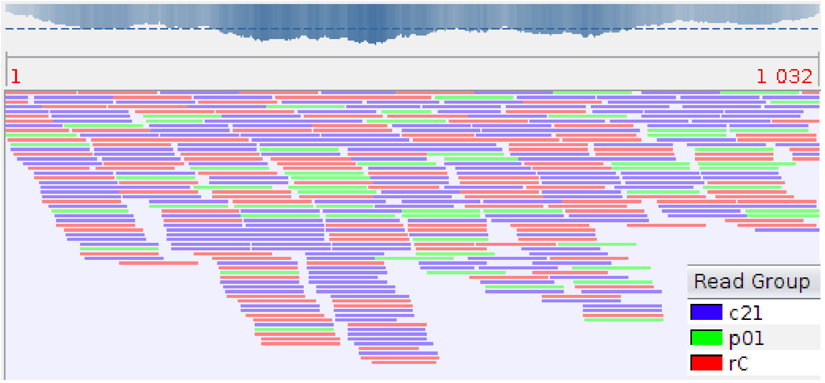




# Anchoring results

- **Anchored 184,880 sequences**
  - 58.7% survey sequences
- **96,784,747 bp anchored**
  - 66.6% of survey sequence length
  - 30.2% estimated arm length
- 878 contigs with at least one sequence
- 1 – 2,514 sequences per contig


**Anchored sequences**



# Analysis of anchored sequences

**184,880 anchored sequences**

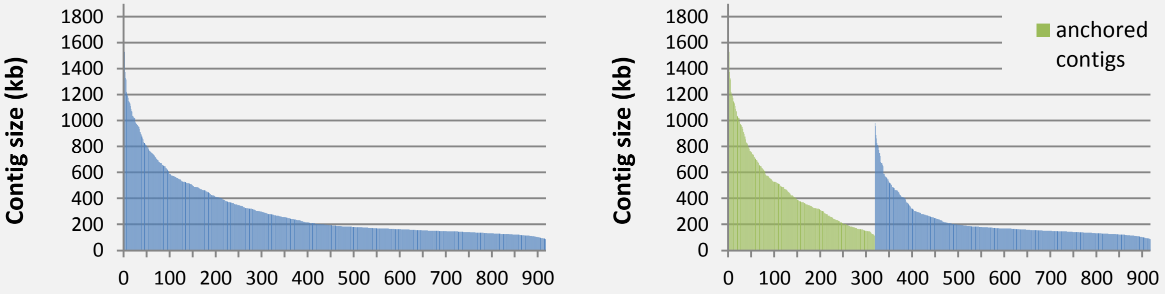
**DArT**

- 194 DArTs identified in 182 sequences
- 125 contigs anchored by DArT markers (1 – 6 markers/contig)

**Gene fragments**

- 1,906 gene models/fragments identified in 3DS survey sequences
- 1,408 (73.9%) genes/fragment anchored (by 1,372 sequences)
- 377 contigs contain at least one gene (1 – 24/contig)
- 793 organized using GenomeZipper approach
- 291 contigs anchored by GenomeZipper (1-17 gene fragments/contig)

## 319 contigs anchored - 53.4% of physical map length

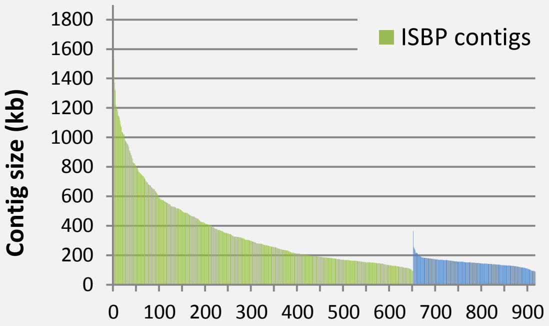


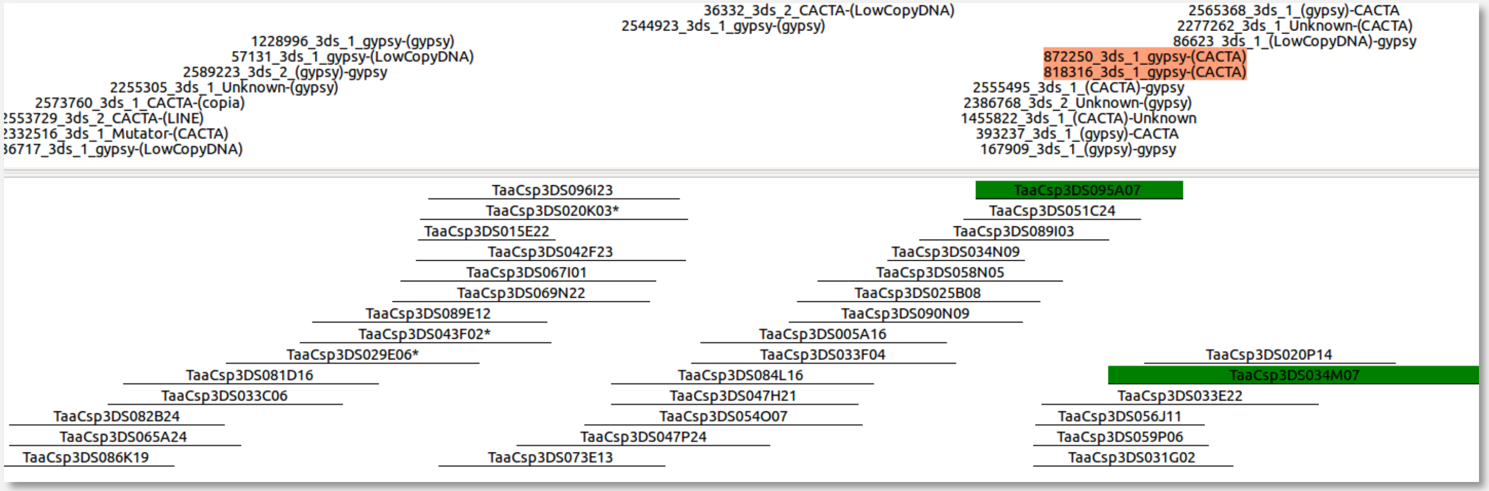
# Analysis of anchored sequences

### Repeat junctions

- IsbpFinder used to identify repeat junctions (potential ISBP markers)
- 24,517 TE insertions with preserved ends were found in 3DS survey sequences
- 17,684 (72.1%) ISBPs anchored to contigs (in 13,870 sequences)
- Up to 232 ISBPs in one contig
- 652 contigs (85.6% of physical map length) have at least one insertion site

## 184,880 anchored sequences




36332_3ds_2_CACTA-(LowCopyDNA)
2544923_3ds_1_gypsy-(gypsy)
2565368_3ds_1_(gypsy)-CACTA
2277262_3ds_1_Unknown-(CACTA)
86623_3ds_1_(LowCopyDNA)-gypsy
1228996_3ds_1_gypsy-(gypsy)
57131_3ds_1_gypsy-(LowCopyDNA)
2589223_3ds_2_(gypsy)-gypsy
2255305_3ds_1_Unknown-(gypsy)
2573760_3ds_1_CACTA-(copia)
2553729_3ds_2_CACTA-(LINE)
2332516_3ds_1_Mutator-(CACTA)
36717_3ds_1_gypsy-(LowCopyDNA)
872250_3ds_1_gypsy-(CACTA)
818316_3ds_1_gypsy-(CACTA)
2555495_3ds_1_(CACTA)-gypsy
2386768_3ds_2_Unknown-(gypsy)
1455822_3ds_1_(CACTA)-Unknown
393237_3ds_1_(gypsy)-CACTA
167909_3ds_1_(gypsy)-gypsy
TaaCsp3DS096I23
TaaCsp3DS095A07
TaaCsp3DS020K03*
TaaCsp3DS051C24
TaaCsp3DS015E22
TaaCsp3DS089I03
TaaCsp3DS042F23
TaaCsp3DS034N09
TaaCsp3DS067I01
TaaCsp3DS058N05
TaaCsp3DS069N22
TaaCsp3DS025B08
TaaCsp3DS089E12
TaaCsp3DS090N09
TaaCsp3DS043F02*
TaaCsp3DS005A16
TaaCsp3DS029E06*
TaaCsp3DS033F04
TaaCsp3DS020P14
TaaCsp3DS081D16
TaaCsp3DS084L16
TaaCsp3DS034M07
TaaCsp3DS033C06
TaaCsp3DS047H21
TaaCsp3DS033E22
TaaCsp3DS082B24
TaaCsp3DS054O07
TaaCsp3DS056J11
TaaCsp3DS065A24
TaaCsp3DS047P24
TaaCsp3DS059P06
TaaCsp3DS086K19
TaaCsp3DS073E13
TaaCsp3DS031G02

# Quality control

**DArT**

- 40 contigs with more than one DArT
- 74% same or close position on DArT map

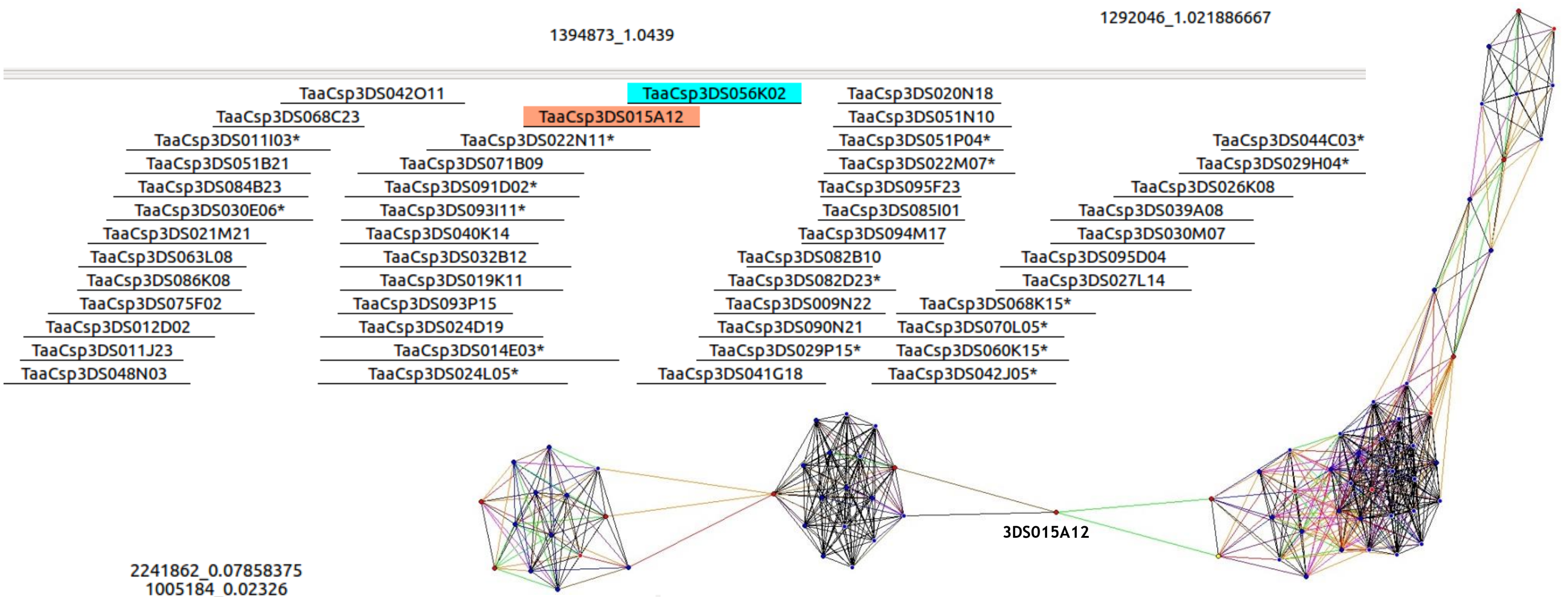
**GenomeZipper**

- 192 contigs with more than one gene fragment
- 70% neighbour positions on GenomeZipper

**Hmmm… Anchoring error rate is overestimated**

Additional sources of error

- BAC contig miss-assembly
- Genetic mapping of DArT markers
- Incorrect position of  gene fragment in GenomeZipper



1394873_1.0439

1292046_1.021886667

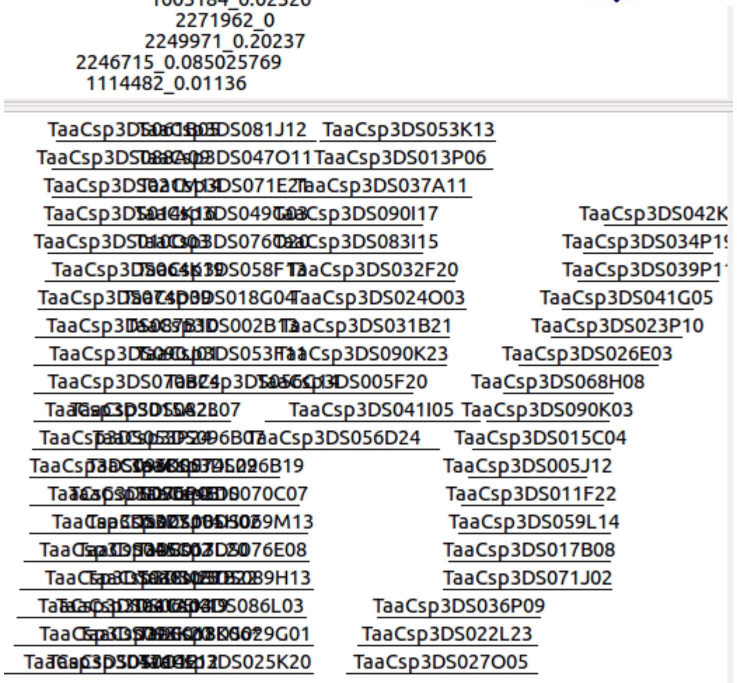
TaaCsp3DS042O11
TaaCsp3DS068C23
TaaCsp3DS011I03*
TaaCsp3DS051B21
TaaCsp3DS084B23
TaaCsp3DS030E06*
TaaCsp3DS021M21
TaaCsp3DS063L08
TaaCsp3DS086K08
TaaCsp3DS075F02
TaaCsp3DS012D02
TaaCsp3DS011J23
TaaCsp3DS048N03

TaaCsp3DS015A12
TaaCsp3DS022N11*
TaaCsp3DS071B09
TaaCsp3DS091D02*
TaaCsp3DS093I11*
TaaCsp3DS040K14
TaaCsp3DS032B12
TaaCsp3DS019K11
TaaCsp3DS093P15
TaaCsp3DS024D19
TaaCsp3DS014E03*
TaaCsp3DS024L05*

TaaCsp3DS056K02
TaaCsp3DS020N18
TaaCsp3DS051N10
TaaCsp3DS051P04*
TaaCsp3DS022M07*
TaaCsp3DS095F23
TaaCsp3DS085I01
TaaCsp3DS094M17
TaaCsp3DS082B10
TaaCsp3DS082D23*
TaaCsp3DS009N22
TaaCsp3DS090N21
TaaCsp3DS029P15*
TaaCsp3DS041G18

TaaCsp3DS068K15*
TaaCsp3DS070L05*
TaaCsp3DS060K15*
TaaCsp3DS042J05*

TaaCsp3DS044C03*
TaaCsp3DS029H04*
TaaCsp3DS026K08
TaaCsp3DS039A08
TaaCsp3DS030M07
TaaCsp3DS095D04
TaaCsp3DS027L14

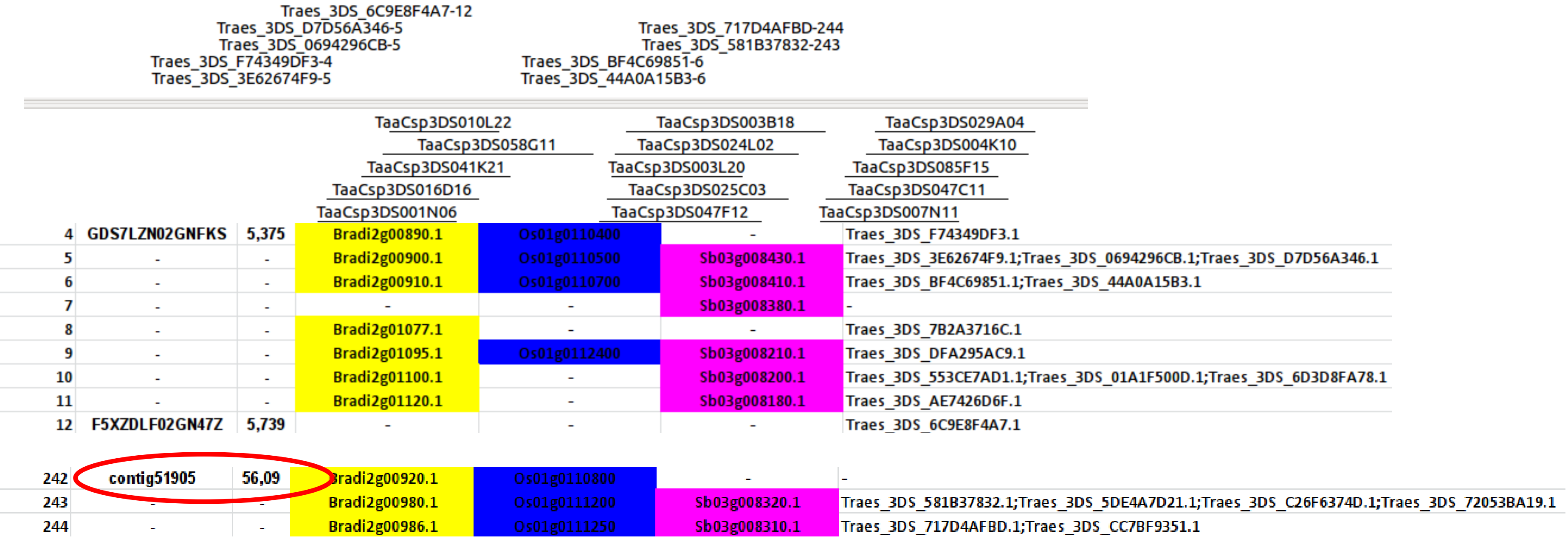


2241862_0.07858375
1005184_0.02326
2271962_0
2249971_0.20237
2246715_0.085025769
1114482_0.01136

TaaCsp3D[illegible]DS081J12 TaaCsp3DS053K13
TaaCsp3D[illegible]DS047O11 TaaCsp3DS013P06
TaaCsp3D[illegible]DS071E2[illegible]aaCsp3DS037A11
TaaCsp3D[illegible]DS049[illegible]aCsp3DS090I17
TaaCsp3D[illegible]DS076[illegible]aCsp3DS083I15
TaaCsp3D[illegible]DS058F1[illegible]aaCsp3DS032F20
TaaCsp3D[illegible]DS018G04[illegible]aaCsp3DS024O03
TaaCsp3D[illegible]DS002B1[illegible]aaCsp3DS031B21
TaaCsp3D[illegible]DS053[illegible]aCsp3DS090K23
TaaCsp3DS07[illegible]DS005F20
Taa[illegible]2B07 TaaCsp3DS041I05
TaaCs[illegible]96B0[illegible]aaCsp3DS056D24
TaaCs[illegible]26B19
Taa[illegible]070C07
Taa[illegible]9M13
Taa[illegible]076E08
Taa[illegible]9H13
Ta[illegible]S086L03
Taa[illegible]29G01
Ta[illegible]DS025K20

TaaCsp3DS042K
TaaCsp3DS034P1
TaaCsp3DS039P1
TaaCsp3DS041G05
TaaCsp3DS023P10
TaaCsp3DS026E03
TaaCsp3DS068H08
TaaCsp3DS090K03
TaaCsp3DS015C04
TaaCsp3DS005J12
TaaCsp3DS011F22
TaaCsp3DS059L14
TaaCsp3DS017B08
TaaCsp3DS071J02
TaaCsp3DS036P09
TaaCsp3DS022L23
TaaCsp3DS027O05

**Contig miss-assembly is significant source of error**



Traes_3DS_6C9E8F4A7-12
Traes_3DS_D7D56A346-5
Traes_3DS_0694296CB-5
Traes_3DS_F74349DF3-4
Traes_3DS_3E62674F9-5

Traes_3DS_717D4AFBD-244
Traes_3DS_581B37832-243
Traes_3DS_BF4C69851-6
Traes_3DS_44A0A15B3-6

TaaCsp3DS010L22
TaaCsp3DS058G11
TaaCsp3DS041K21
TaaCsp3DS016D16
TaaCsp3DS001N06

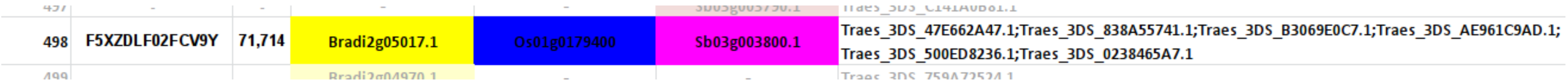
TaaCsp3DS003B18
TaaCsp3DS024L02
TaaCsp3DS003L20
TaaCsp3DS025C03
TaaCsp3DS047F12

TaaCsp3DS029A04
TaaCsp3DS004K10
TaaCsp3DS085F15
TaaCsp3DS047C11
TaaCsp3DS007N11

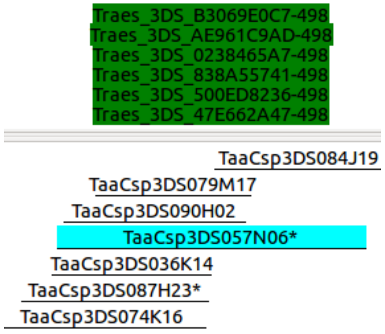
| | | | | | | |
|---|---|---|---|---|---|---|
| 4 | GDS7LZN02GNFKS | 5,375 | Bradi2g00890.1 | Os01g0110400 | - | Traes_3DS_F74349DF3.1 |
| 5 | - | - | Bradi2g00900.1 | Os01g0110500 | Sb03g008430.1 | Traes_3DS_3E62674F9.1;Traes_3DS_0694296CB.1;Traes_3DS_D7D56A346.1 |
| 6 | - | - | Bradi2g00910.1 | Os01g0110700 | Sb03g008410.1 | Traes_3DS_BF4C69851.1;Traes_3DS_44A0A15B3.1 |
| 7 | - | - | - | - | Sb03g008380.1 | - |
| 8 | - | - | Bradi2g01077.1 | - | - | Traes_3DS_7B2A3716C.1 |
| 9 | - | - | Bradi2g01095.1 | Os01g0112400 | Sb03g008210.1 | Traes_3DS_DFA295AC9.1 |
| 10 | - | - | Bradi2g01100.1 | - | Sb03g008200.1 | Traes_3DS_553CE7AD1.1;Traes_3DS_01A1F500D.1;Traes_3DS_6D3D8FA78.1 |
| 11 | - | - | Bradi2g01120.1 | - | Sb03g008180.1 | Traes_3DS_AE7426D6F.1 |
| 12 | F5XZDLF02GN47Z | 5,739 | - | - | - | Traes_3DS_6C9E8F4A7.1 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 242 | contig51905 | 56,09 | Bradi2g00920.1 | Os01g0110800 | - | - |
| 243 | - | - | Bradi2g00980.1 | Os01g0111200 | Sb03g008320.1 | Traes_3DS_581B37832.1;Traes_3DS_5DE4A7D21.1;Traes_3DS_C26F6374D.1;Traes_3DS_72053BA19.1 |
| 244 | - | - | Bradi2g00986.1 | Os01g0111250 | Sb03g008310.1 | Traes_3DS_717D4AFBD.1;Traes_3DS_CC7BF9351.1 |

# Physical localization of gene fragments at identical GenomeZipper position



| | | | | | | |
|---|---|---|---|---|---|---|
| 497 | - | - | - | - | Sb03g003790.1 | Traes_3DS_C141A0B81.1 |
| 498 | F5XZDLF02FCV9Y | 71,714 | Bradi2g05017.1 | Os01g0179400 | Sb03g003800.1 | Traes_3DS_47E662A47.1;Traes_3DS_838A55741.1;Traes_3DS_B3069E0C7.1;Traes_3DS_AE961C9AD.1;<br>Traes_3DS_500ED8236.1;Traes_3DS_0238465A7.1 |
| 499 | | | Bradi2g04970.1 | - | - | Traes_3DS_759A72524.1 |



Traes_3DS_B3069E0C7-498
Traes_3DS_AE961C9AD-498
Traes_3DS_0238465A7-498
Traes_3DS_838A55741-498
Traes_3DS_500ED8236-498
Traes_3DS_47E662A47-498

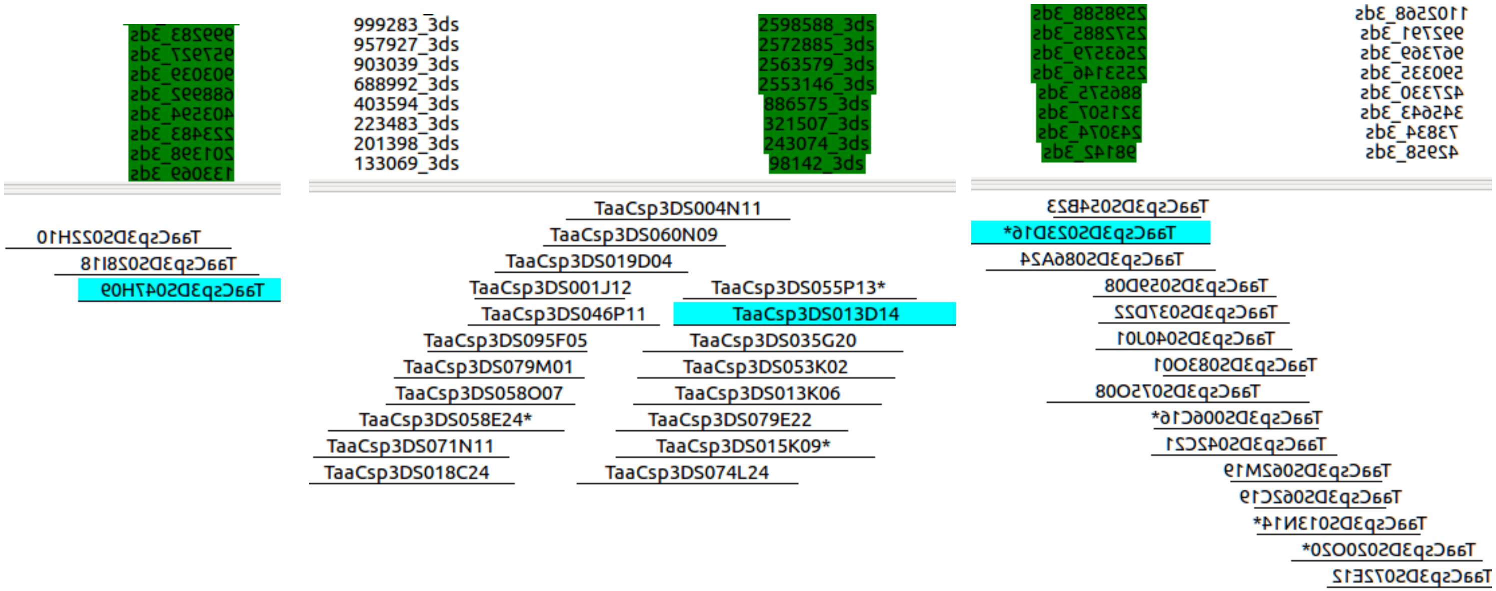
TaaCsp3DS084J19
TaaCsp3DS079M17
TaaCsp3DS090H02
TaaCsp3DS057N06*
TaaCsp3DS036K14
TaaCsp3DS087H23*
TaaCsp3DS074K16

- **178 GenomeZipper positions with multiple gene fragments**
- **For 161 (90.5%) fragments have identical position**

# Additional assembly improvement

6,362 sequences of anchoring type 2b) could be used to merge contigs

- Sequences anchored to clones in different contigs
- Match of the clones at e-10



# Conclusion

- We developed protocol for high-throughput contig anchoring
- 66% of survey sequence (97 Mbp) anchored to physical map
- 74% genes identified in survey sequences localized in BAC clones
- 53% of the physical map organized through anchoring to DArT genetic map and 3DS GenomeZipper

# Future perspective

- Additional validation of results (including wet lab)
- Cleaning and integration of ISBP markers, polymorphism identification within CS x Renan population
- Sequencing of 3,823 clones of MTP

# Acknowledgement

Jaroslav Doležel
Kateřina Cvikova
Jan Šafář
Hana Šimková

Diversity Arrays Technology
IT'S A BULLSEYE WITH DArT

Andrzej Killian

IGA

Federica Cattonaro

IWGSC

URGI

Michael Alaux

GAČR
CZECH SCIENCE FOUNDATION

MŠMT

MINISTRY OF EDUCATION
YOUTH AND SPORTS