

Toulouse, France



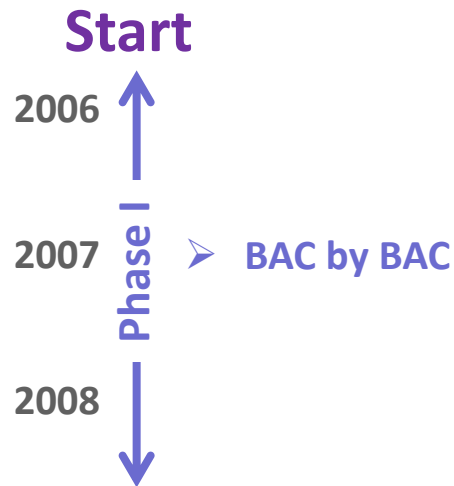
High-quality *de novo* genome assembly of the tomato genome using Long Read sequencing

Mohamed Zouine

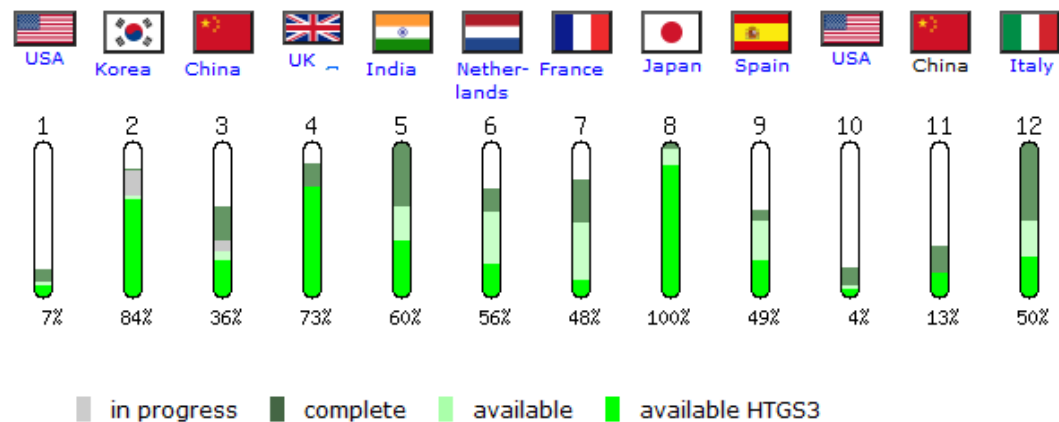
 @GBF_Lab



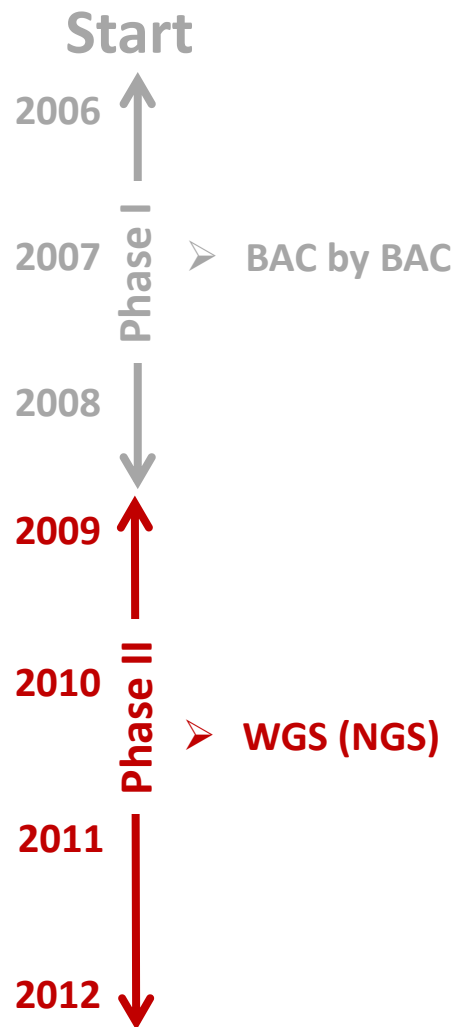
Strategy and history



- ✓ **Genome**
 - 12 chromosomes
 - 950 Mb genome
 - ~220 Mb of contiguous gene-rich euchromatin
- ✓ **Produce finished sequence of the euchromatin**
 - 25% of the genome carries >90% of all genes (~28,000-35,000)
- ✓ **Most feasible strategy at the time**
 - BAC-by-BAC sequencing approach



Moving to WGS/NGS strategy



Combines sequencing technologies

Type	Amount	
454	29 Gb	} NGS
SOLiD	~ 60 Gb	
SBM	3 Gb	
BAC ends	300,000 reads (~150,000 pairs)	
Fosmid ends	150,000 reads (~75,000 pairs)	
BAC contigs	70 Mb, 36% of <u>euchromatin</u>	+
Physical map	10X BAC coverage	

Two hybrid assemblies

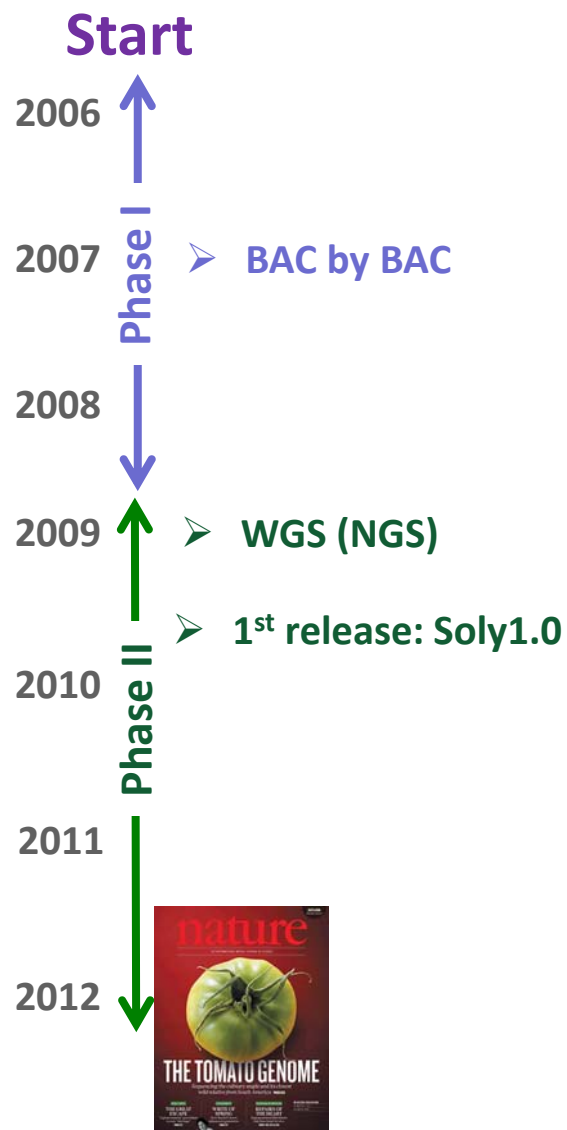
↓

CABOG
(GBF, Fr)

↓

NEWBLER
(NI)

First tomato genome assembly v1.0



Soly1.00

sol genomics network

home | forum | contact | help

search maps sequencing tools

log in | new user

Tomato Genome Shotgun Sequence Prerelease

December 1, 2009

The [International Tomato Sequencing Project](#) is happy to announce the pre-release of the tomato genome, produced by a whole genome shotgun approach using 454 sequencing, Sanger shotgun sequences, and BAC/fosmid end sequences.

The current release comprises scaffold sequences only. The International Tomato Annotation Group (ITAG) is currently working on a comprehensive, high quality annotation that will be released in early 2010.

Before using the sequences, we ask that you read and agree to the data access agreement below.



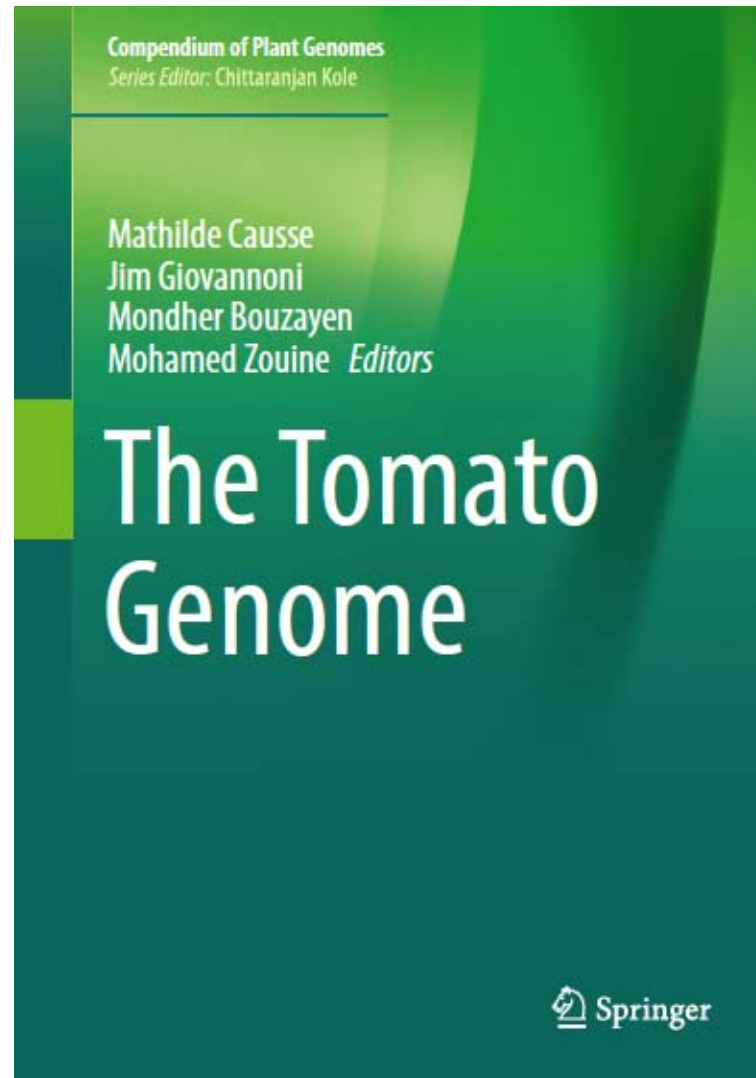
Disclaimers and data access agreement

PLEASE READ BEFORE ACCESSING THE PRE-PUBLICATION TOMATO GENOME SEQUENCE: The International Tomato Genome Sequencing Consortium is pleased to make available a pre-publication draft assembly of the tomato genome for use by public and private research communities as a resource to enable plant biology discovery and improve the human condition through improved agriculture. This assembly (version 1.0) was produced by the Dutch/French assembly team and includes both 454 data and Sanger sequence data (BAC-ends, fosmid-ends and Selected BAC mixture sequences).

We caution you that the current assembly is a "work-in-progress" and as such is subject to modification prior to publication release (anticipated for mid-2010), some of which is likely to be substantial. Therefore we encourage you to carefully and independently validate any conclusions you may draw from this sequence. We will update this resource as improvements in the assembly are made. We welcome any feedback regarding your successes or that may assist us in improving the quality and accuracy of this sequence.

The pre-publication tomato genome sequence is made available with the understanding that users will respect the rights of those who contributed to this effort to describe the tomato genome in a peer-reviewed publication. This description includes whole genome level analyses on genes, gene families, repetitive sequences etc. We encourage you to review the NIH-NHGRI guidelines on distribution and use of pre-publication genome sequence at <http://www.genome.gov/page.cfm?pageID=10506537>. Any use of the tomato genome sequence prior to its publication should credit "The International Tomato Genome Sequencing Consortium". If you are uncertain about how to credit the use of the sequence or its appropriate use please do not hesitate to contact [Joyce Van Eck](#).

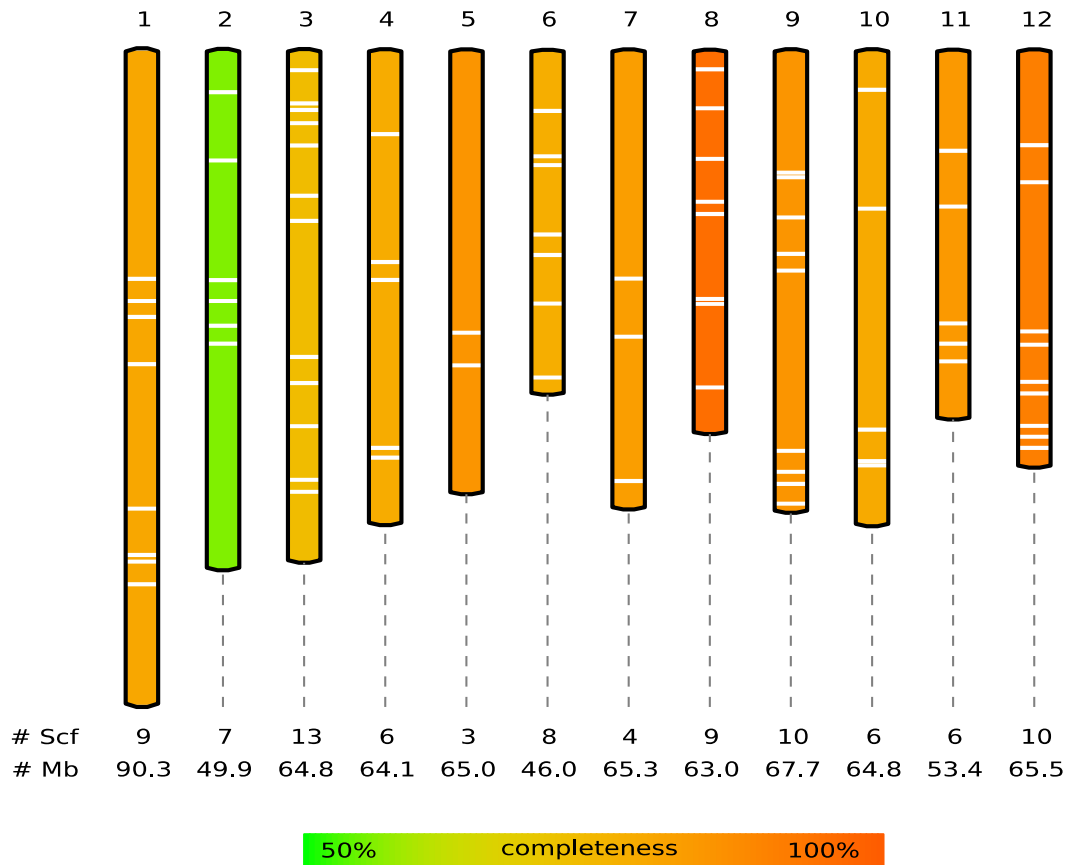
The whole story here:



Current assembly



- SL2.5/iTAG2.4 and now SL3.0/iTAG3.2



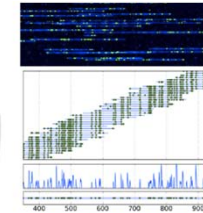
Good assembly but still needs some improvements :

- Increase assembled genome size
110 872 contigs in SL1.0 =>
More than 100 000 gaps
- Decrease Chromosome zero size
- Improve structural annotation

Combining Long read sequencing technologies



PacBio
Long contigs



BioNano
Long Scaffolds

+



10x Genomics

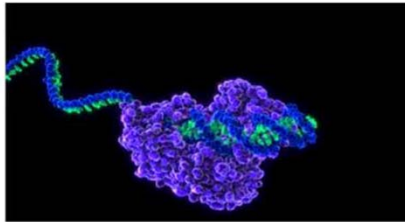
Super-scaffolding & error correction



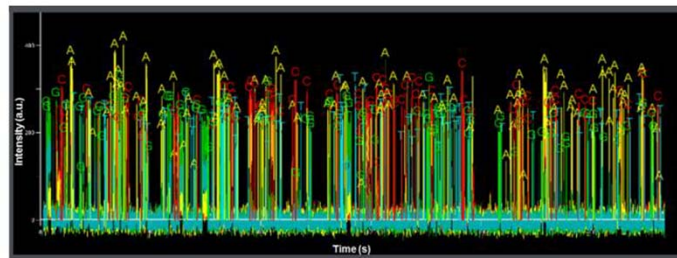
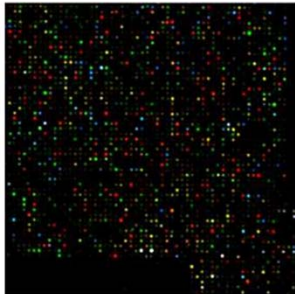
Toward a new and improved tomato genome reference sequence ?

PacBio technology

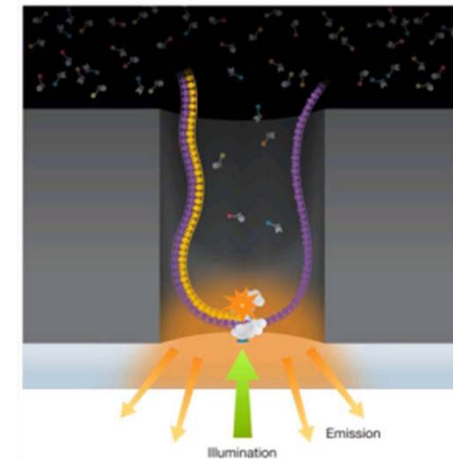
Single Molecule, Real-Time (SMRT) DNA Sequencing



PacBio RS II

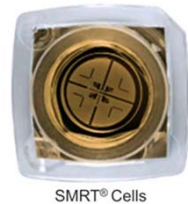


SMRT® Cells

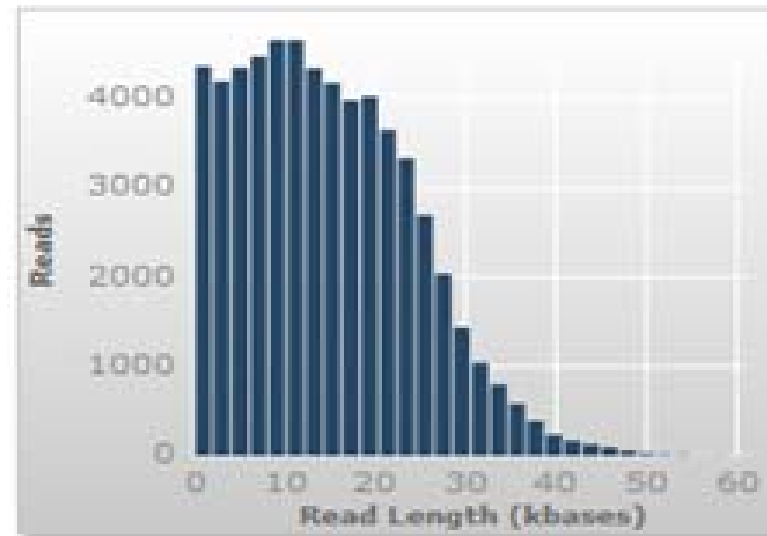


Science, Vol 299, Jan 31 2003, pp682-68
J. Appl. Phys. 103, 034301 (2008)

PacBio sequence Assembly



55 SMARTCells
6.3 Mreads, 67 Mbases, 70X coverage
Max length: 71 kb
Median sequence length: 18 kb



Assembly with CANU:

Number of contigs: **743** (110 872 contigs in SL1.0)

Genome size : **800 Mb**

Longest contig: **13 Mb**

N50 length: **3.4 Mb**

N50 index: **62**

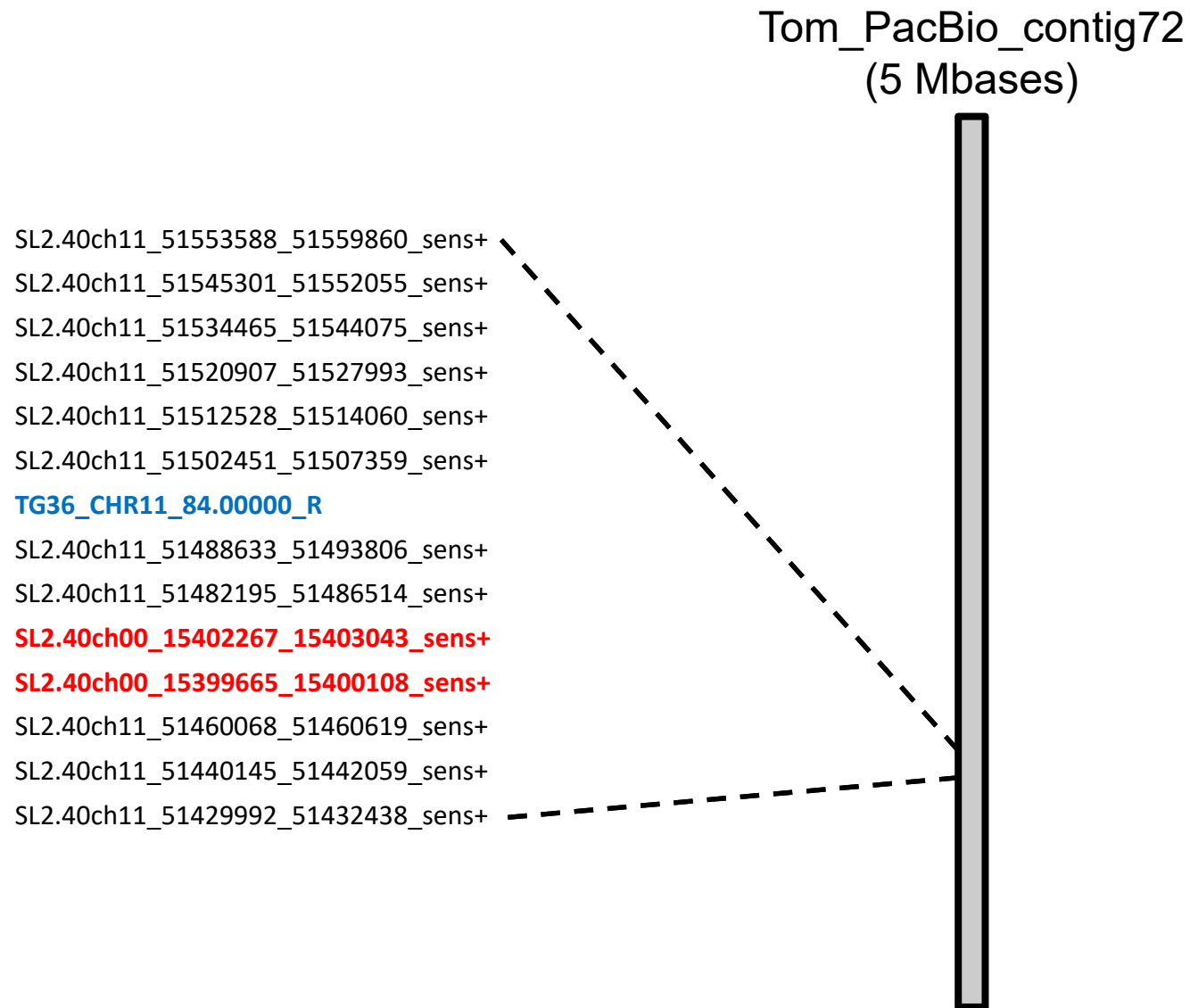
```

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000

```

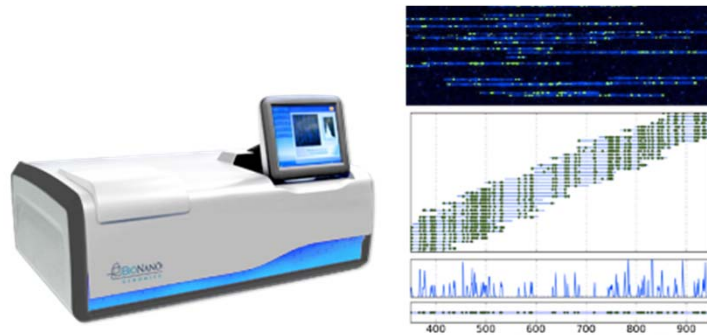
Example of region identified by QTL mapping in the **SL_PacBio** assembly

iTAG2.5 genes & genetic markers mapping



Scaffolding

Optical Mapping
BioNano

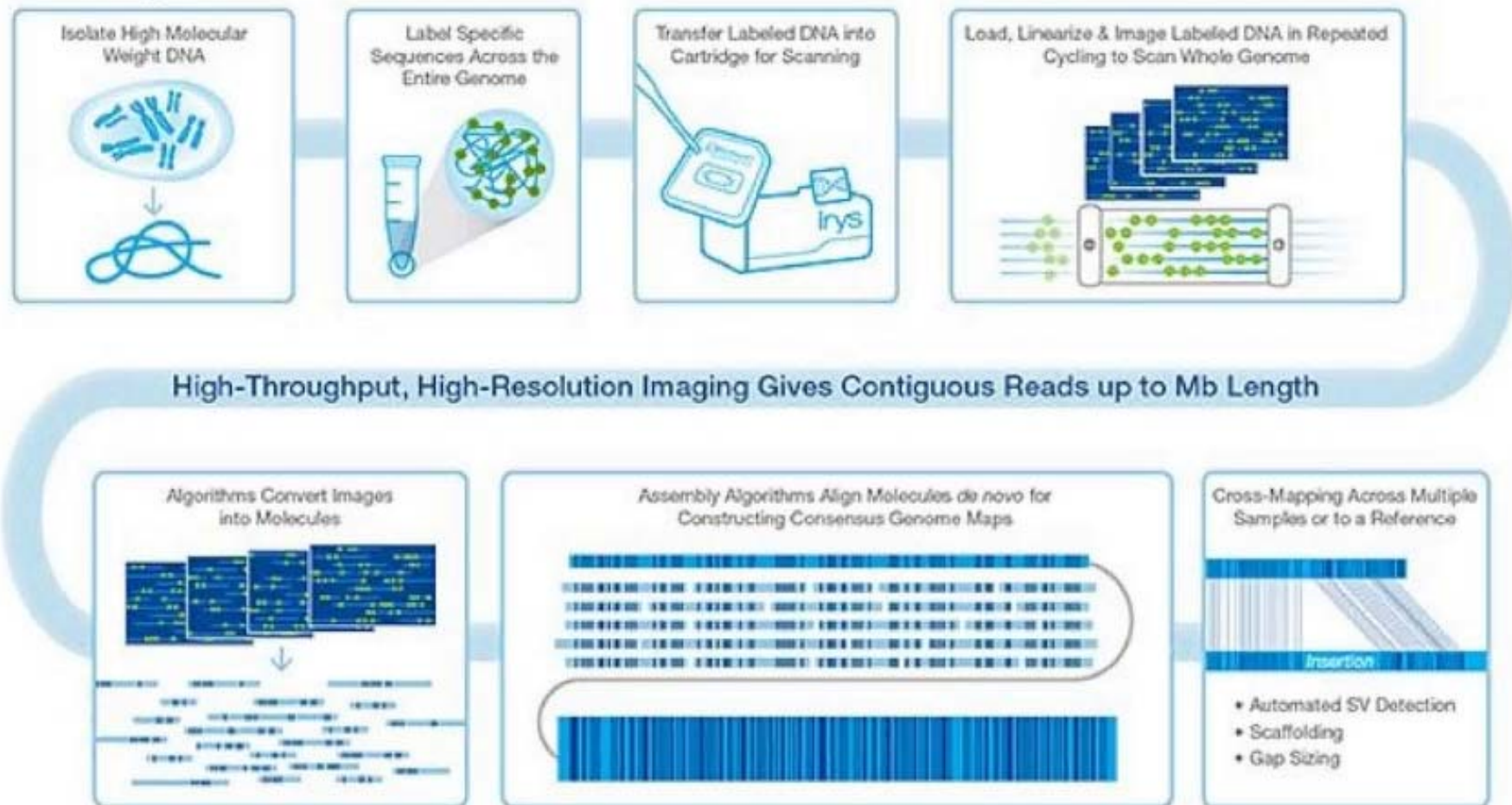


+

Long read
sequencing
Chromium 10x



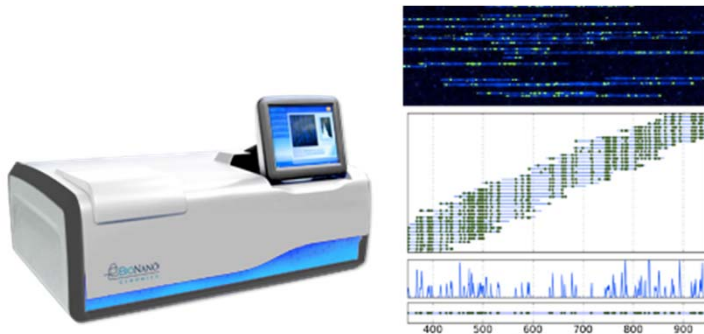
Optical Mapping using BioNano



www.youtube.com/watch?v=XwBI13Q4ilo

Scaffolding

Optical Mapping
BioNano



Two enzymes, 200X coverage

N50: 34 Mb
99% of the Genome in
40 scaffolds

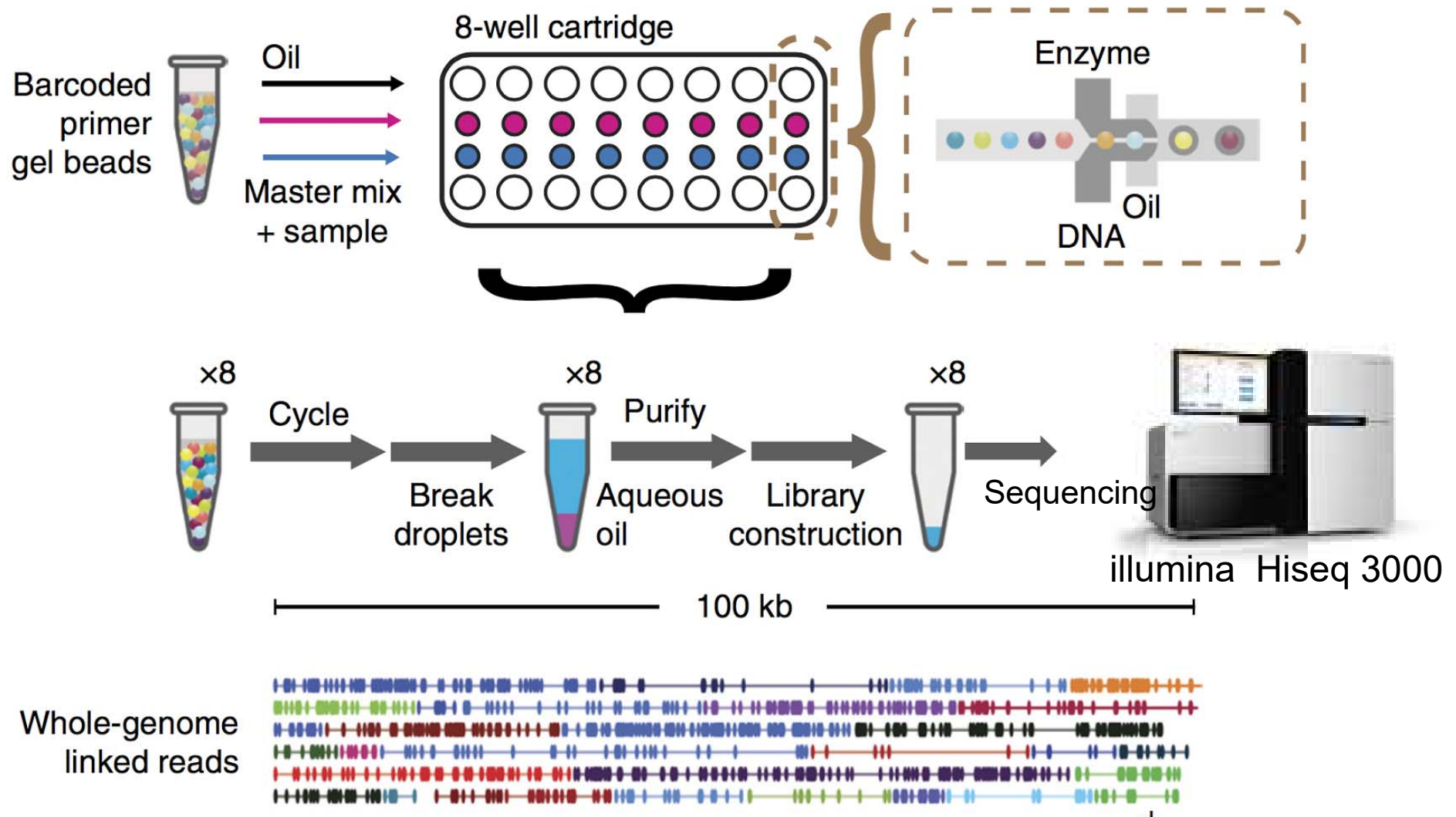
Long read
sequencing
Chromium 10X



+

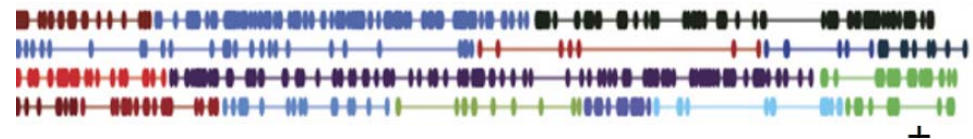
?

Long read sequencing by Chromium



De novo Assembly of 10x reads

Tool: Supernova



Number of contigs: **24 K**
Genome size : **800 Mb**
Longest contig: **11.3 Mb**
N50 length: **1.8 Mb**
N50 index: **125**

Scaffolding with 10x using ARCS

bioRxiv preprint first posted online Jan. 17, 2017; doi: <http://dx.doi.org/10.1101/100750>. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a [CC-BY 4.0 International license](#).

ARCS: Assembly Roundup by Chromium Scaffolding

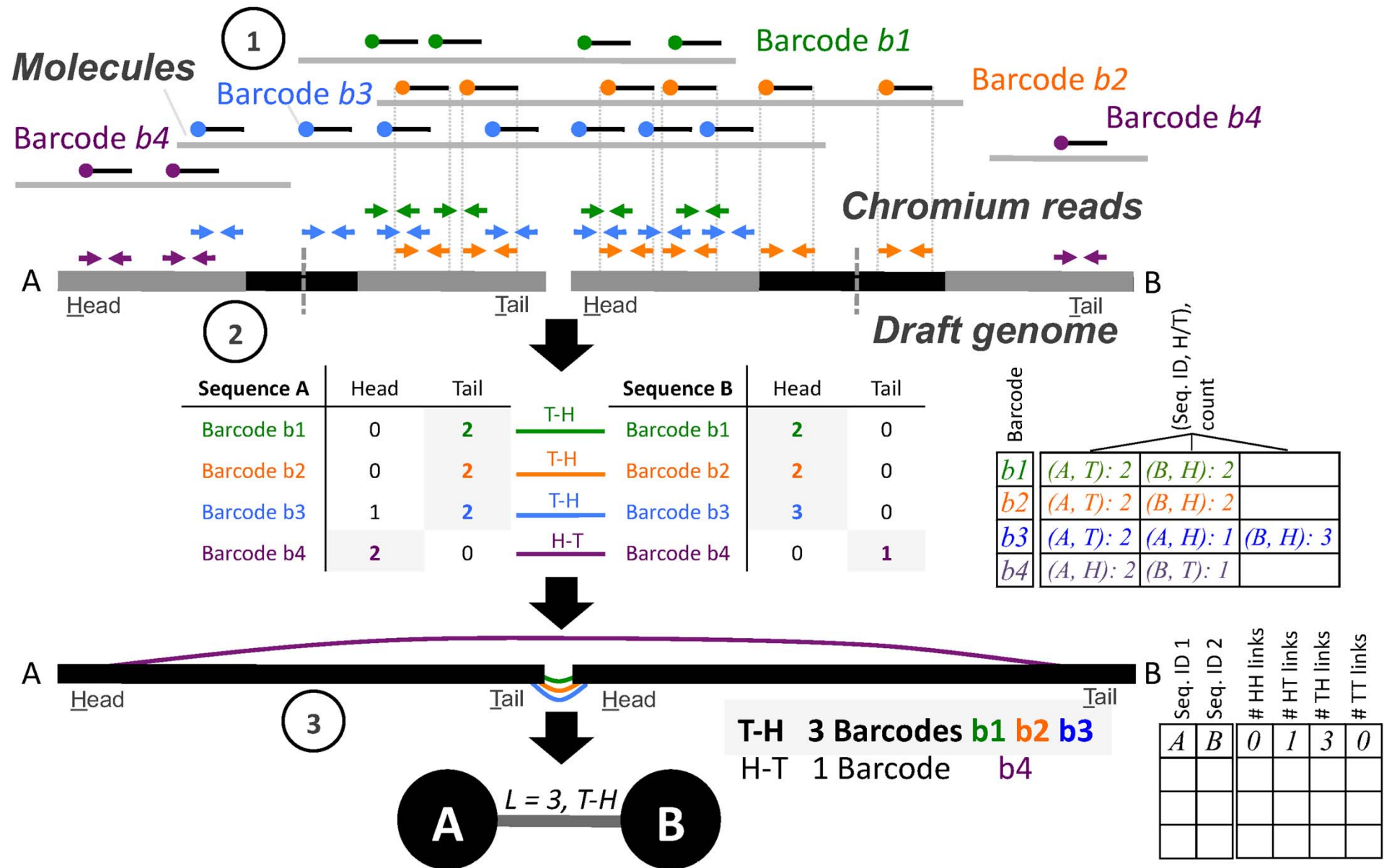
Sarah Yeo¹, Lauren Coombe¹, Justin Chu, René L Warren^{1,2}, and Inanç Birol

BC Cancer Agency, Genome Sciences Centre, Vancouver, BC V5Z 4S6, Canada

¹Authors contributed equally

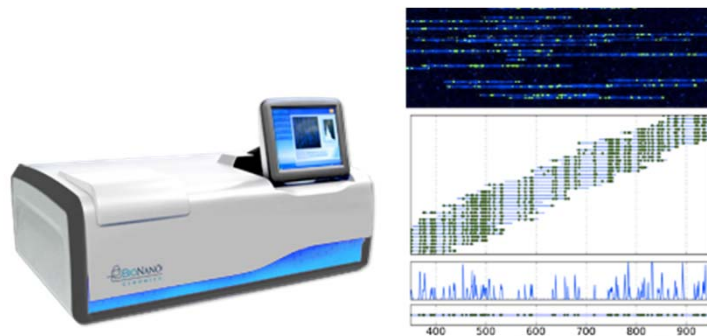
²corresponding author: rwarren@bcgsc.ca

Scaffolding with 10x using ARCS



Scaffolding

Optical Mapping by BioNano



Two enzymes, 300X coverage

N50: 34 Mb

Genome in 40 scaffolds

Genome size 826 Mb

+

Linked read sequencing
By Chromium 10X



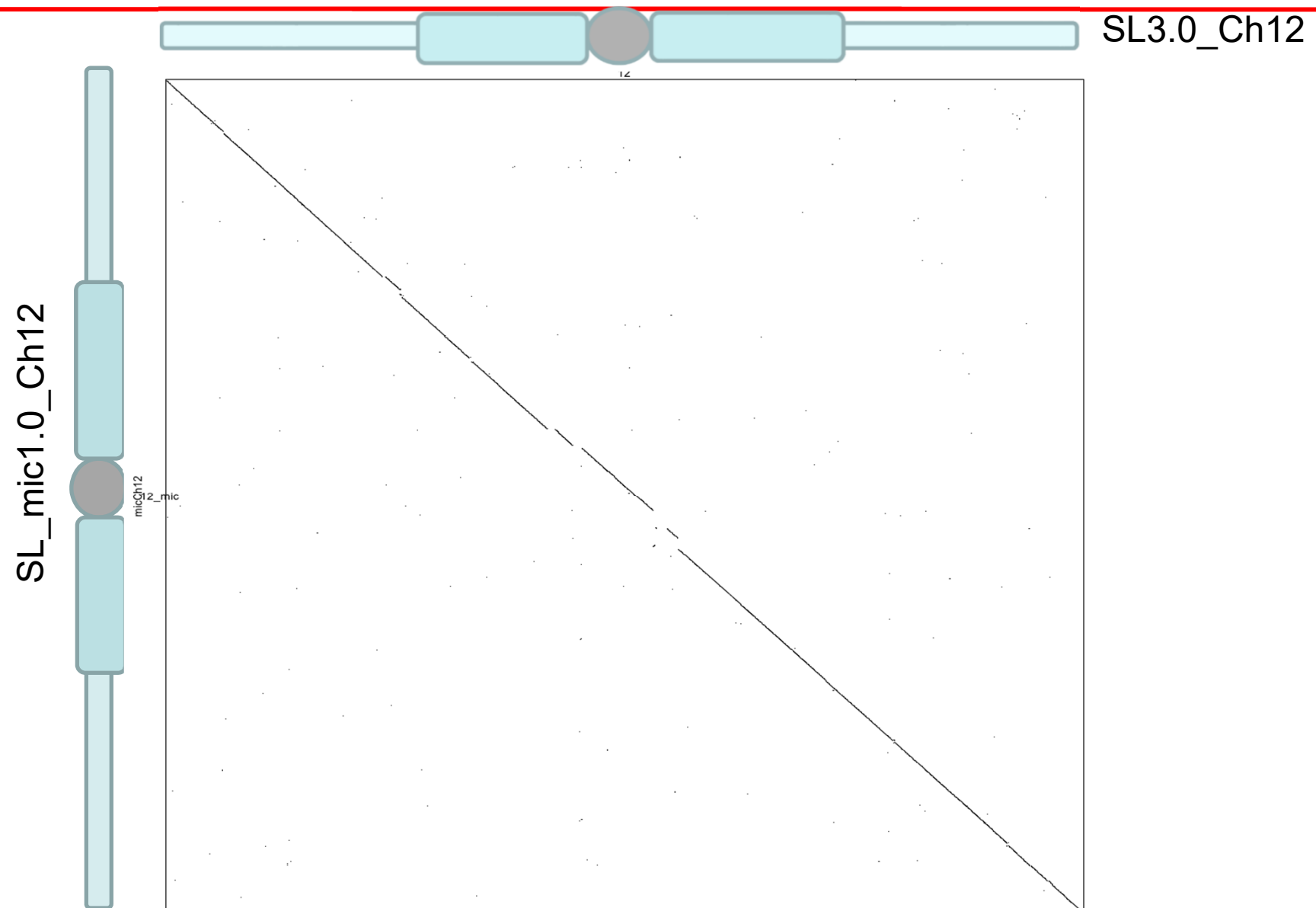
100X illumina coverage

N50: 45 Mb

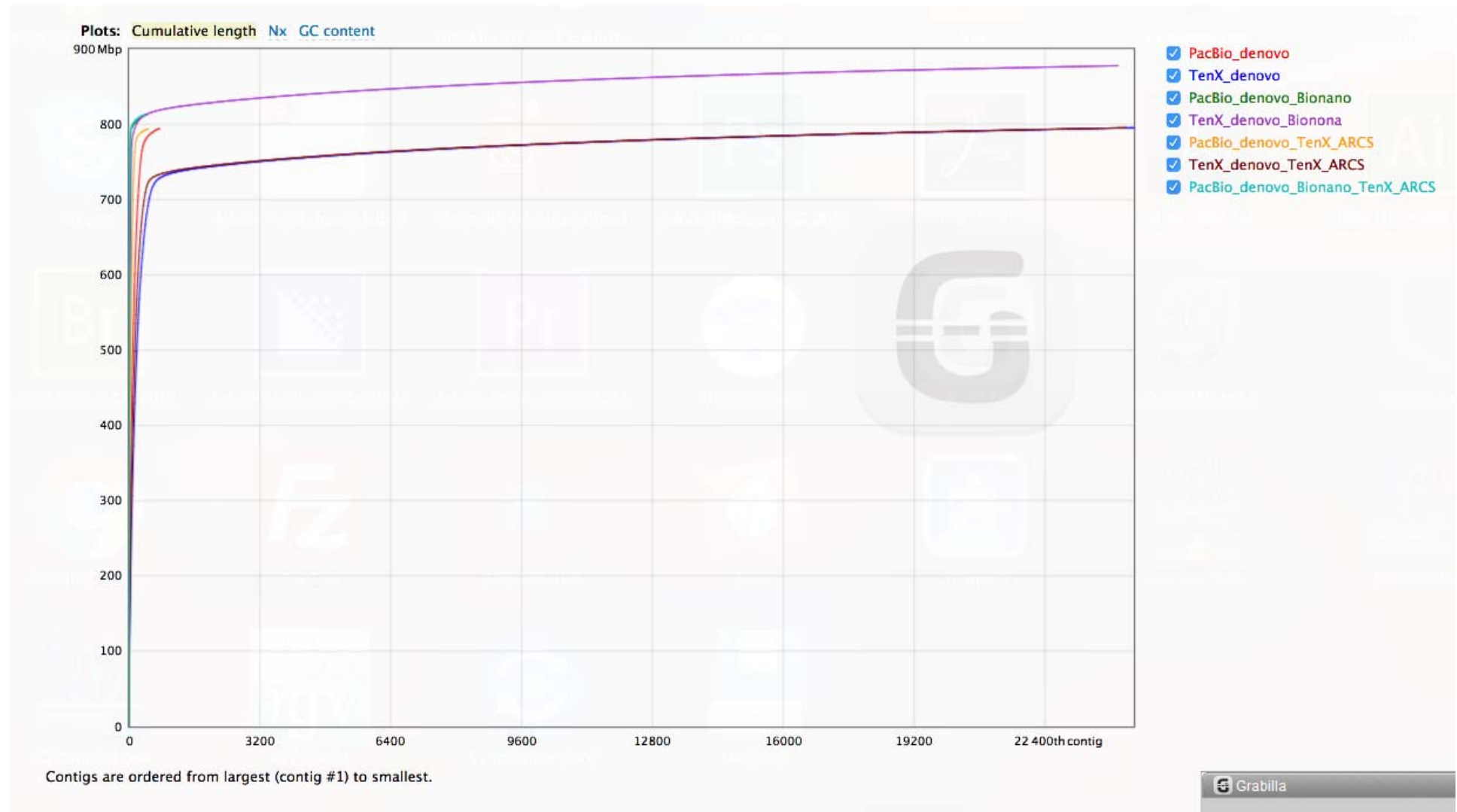
99% Ggenome in

30 scaffolds

Chromosome 12 is in one scaffold



Assemblies metrics comparison



Assemblies metrics comparison:

What is the best combination / low price?

	PacBio	TenX	PacBio_ Bionano	TenX_ Bionano	PacBio TenX_AR CS	TenX TenX_AR CS	PacBio Bionano TenX_AR CS
Contigs number	743	24579	447	24183	475	24381	392
Contigs (≥ 10000 bp)	735	1321	436	925	467	1123	381
Contigs (≥ 50000 bp)	510	684	205	287	265	529	167
Total length	793 Mb	795 Mb	812 Mb	877 Mb	793 Mb	795 Mb	812 Mb
Largest contig	13628425	11337354	59878527	51292213	29789711	11337354	69682238
N50	3394188	1800595	34211668	15645230	7979471	2320671	44368898
N75	1789934	813655	16105568	6758801	3979370	1064715	19357061
L50	63	125	10	17	29	93	8
L75	141	288	18	40	64	218	14
N's per 100 kbp	0.00	6071.07	2325.86	14885.68	0.34	6071.31	2325.92
Cost	€€€	€	€€€€	€€	€€€	€	€€€€€

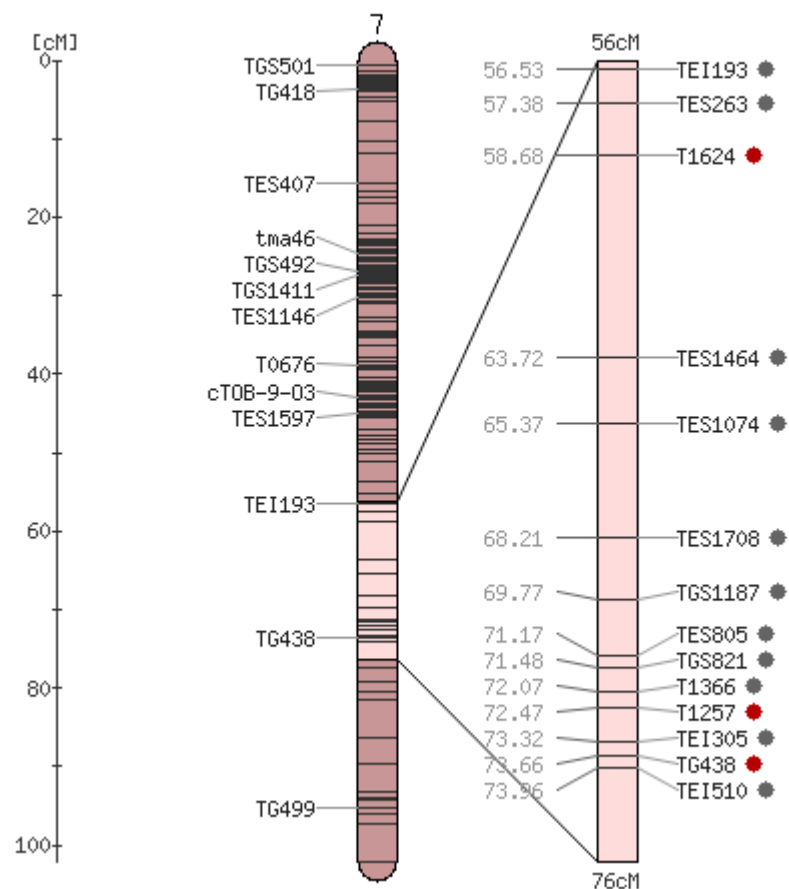
Anchoring scaffolds on chromosomes:

Pseudomolecules construction

- Mapping the 2000 markers on the scaffolds
- Ordering the scaffolds pseudomolecules

Viewing chr 7 of map [Kazusa F2-2000 genetic map](#) [Help]

	markers	Marker collections	
Chromosome 1	250	COS	195
Chromosome 2	182	COSII	111
Chromosome 3	191		
Chromosome 4	171	Total:	306
Chromosome 5	160		
Chromosome 6	156		
Chromosome 7	148	Protocols	
Chromosome 8	149	SNP	138
Chromosome 9	179	SSR	1250
Chromosome 10	158	unknown	682
Chromosome 11	156		
Chromosome 12	170	Total:	2070
Total:	2070		

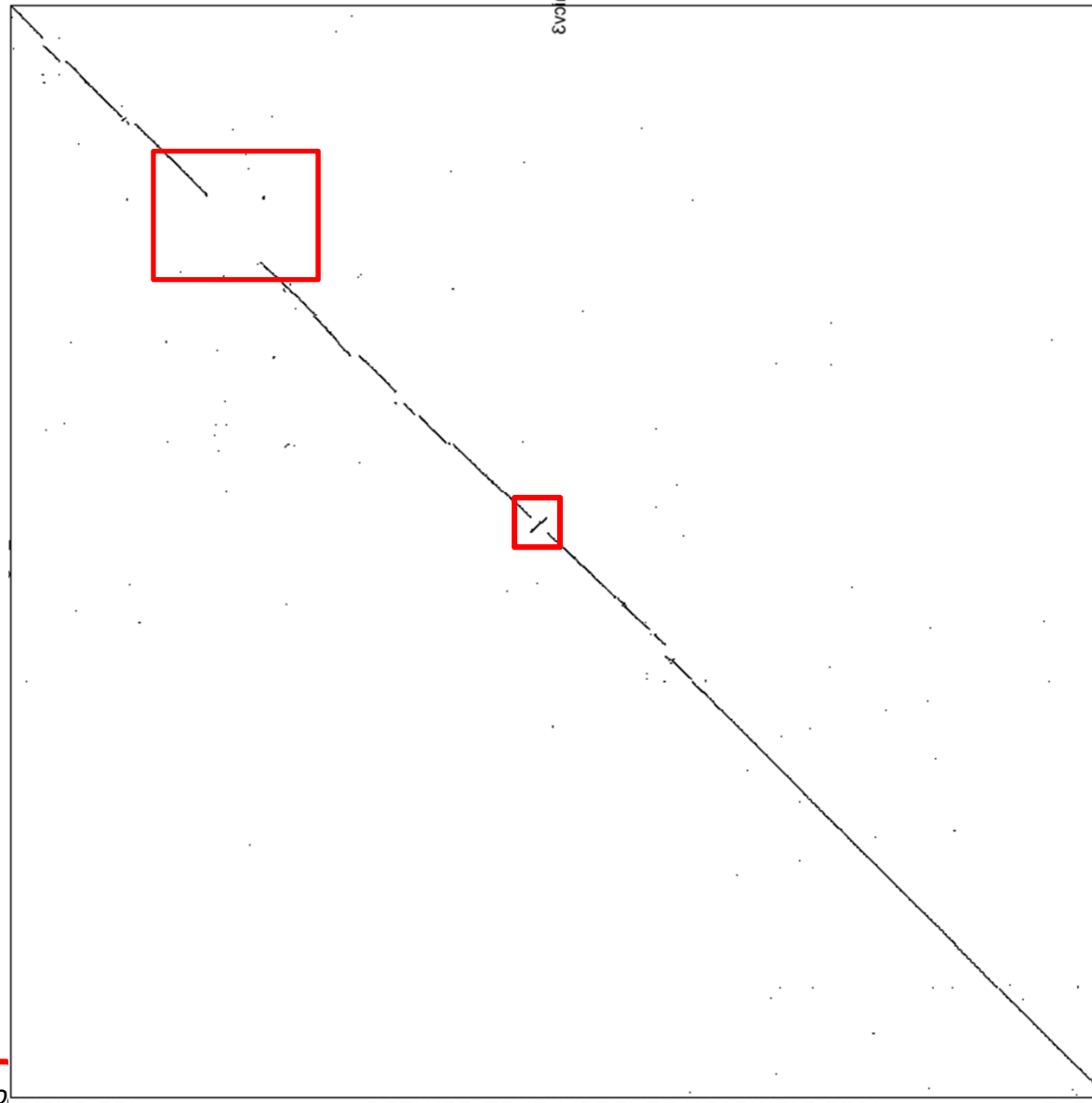


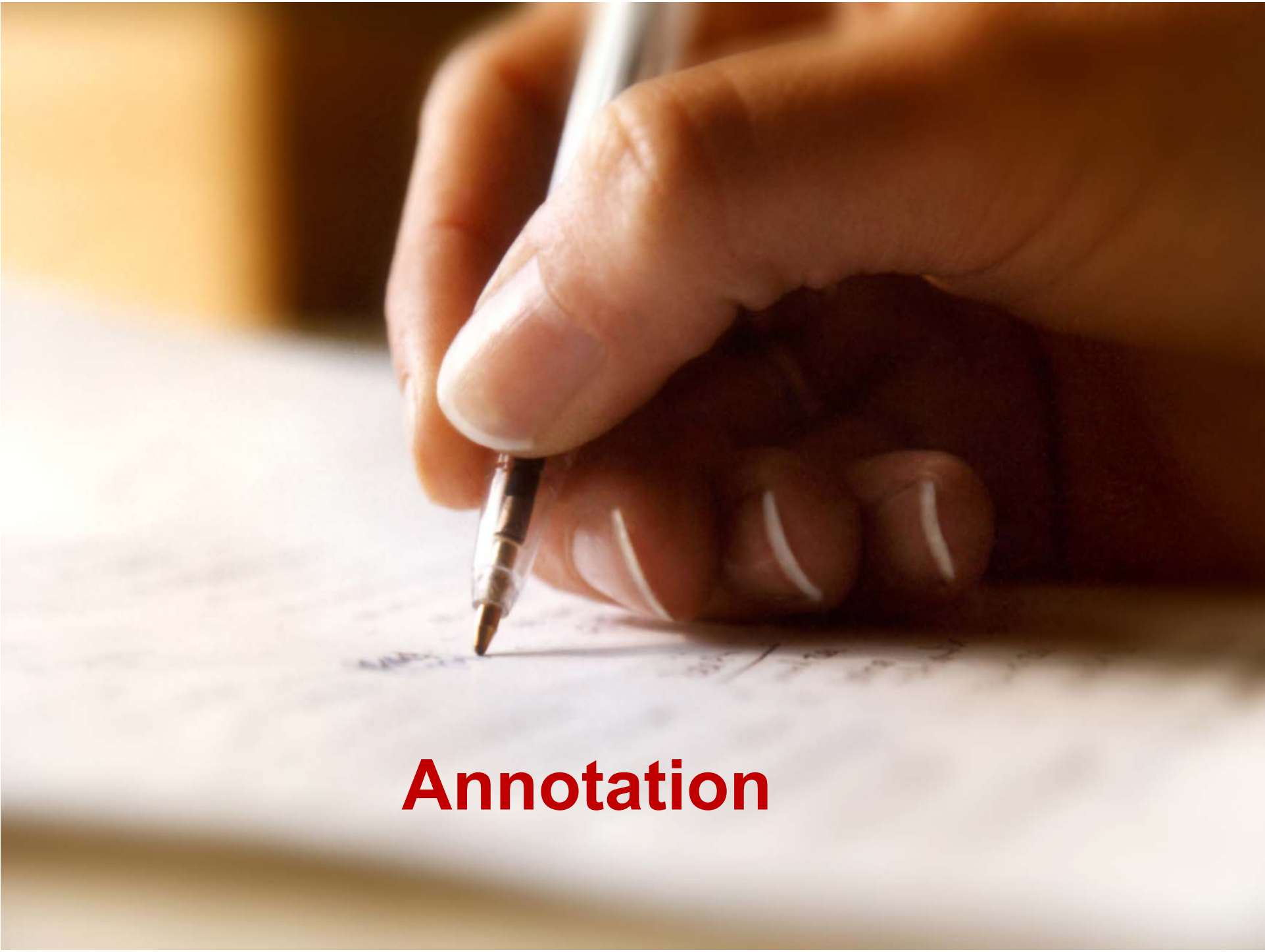
Chromosome 2

SL3.0_Ch2

SL_mic1.0_Ch2

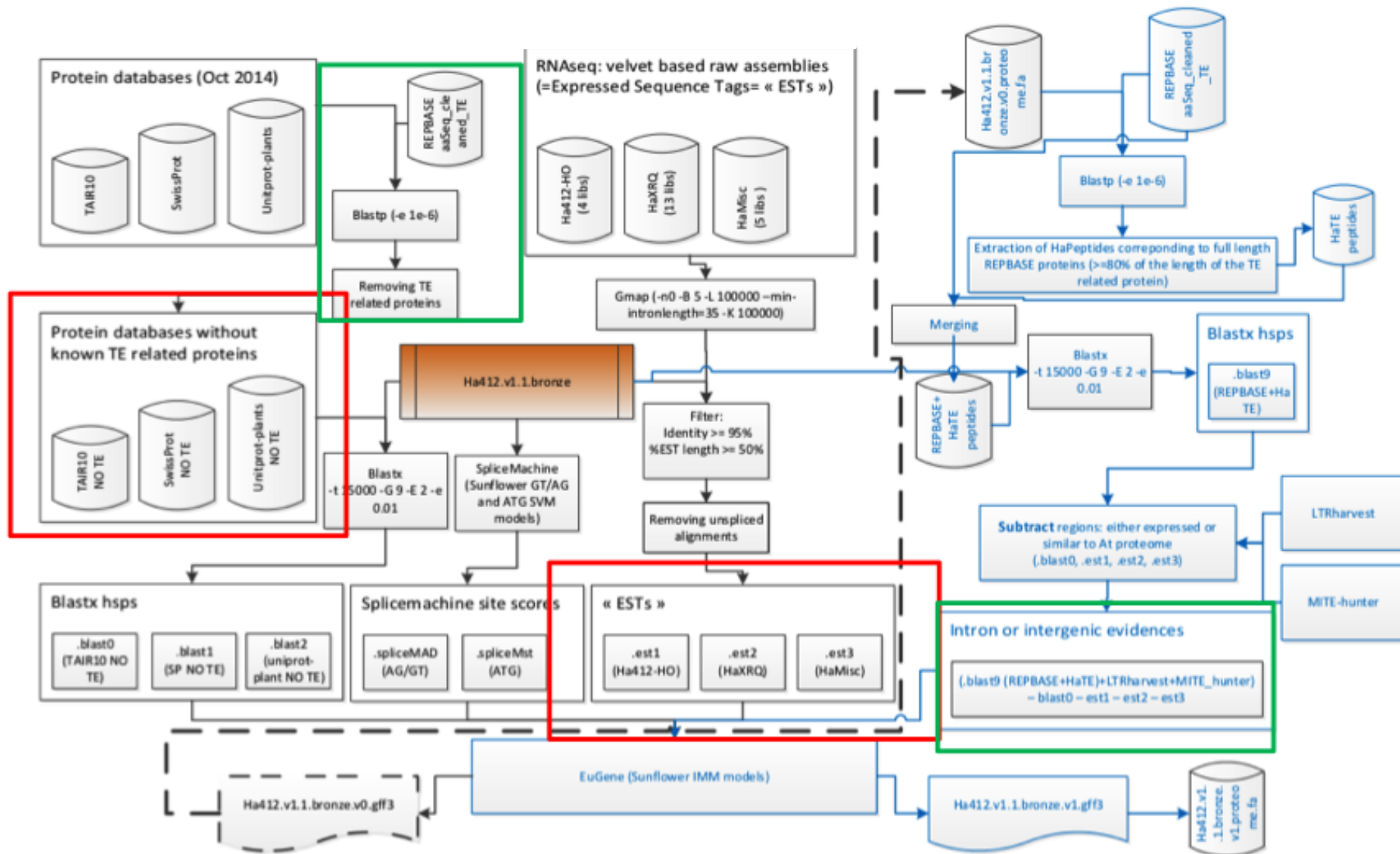
evolm.





Annotation

Tomato EugenEP pipeline



Annotation by transcripts evidence using TomExpress data



TomExpress, a unified tomato RNA-Seq platform for visualization of expression data, clustering and correlation networks

Zouine et al., 2017, the Plant Journal (online)

29 projects

222 conditions

Different cultivars

800 samples (*.fastq sequence files*)

Plant and fruit Development

Pollen

Trichomes

Tomato mutants

Hormones treatments

Biotic interactions

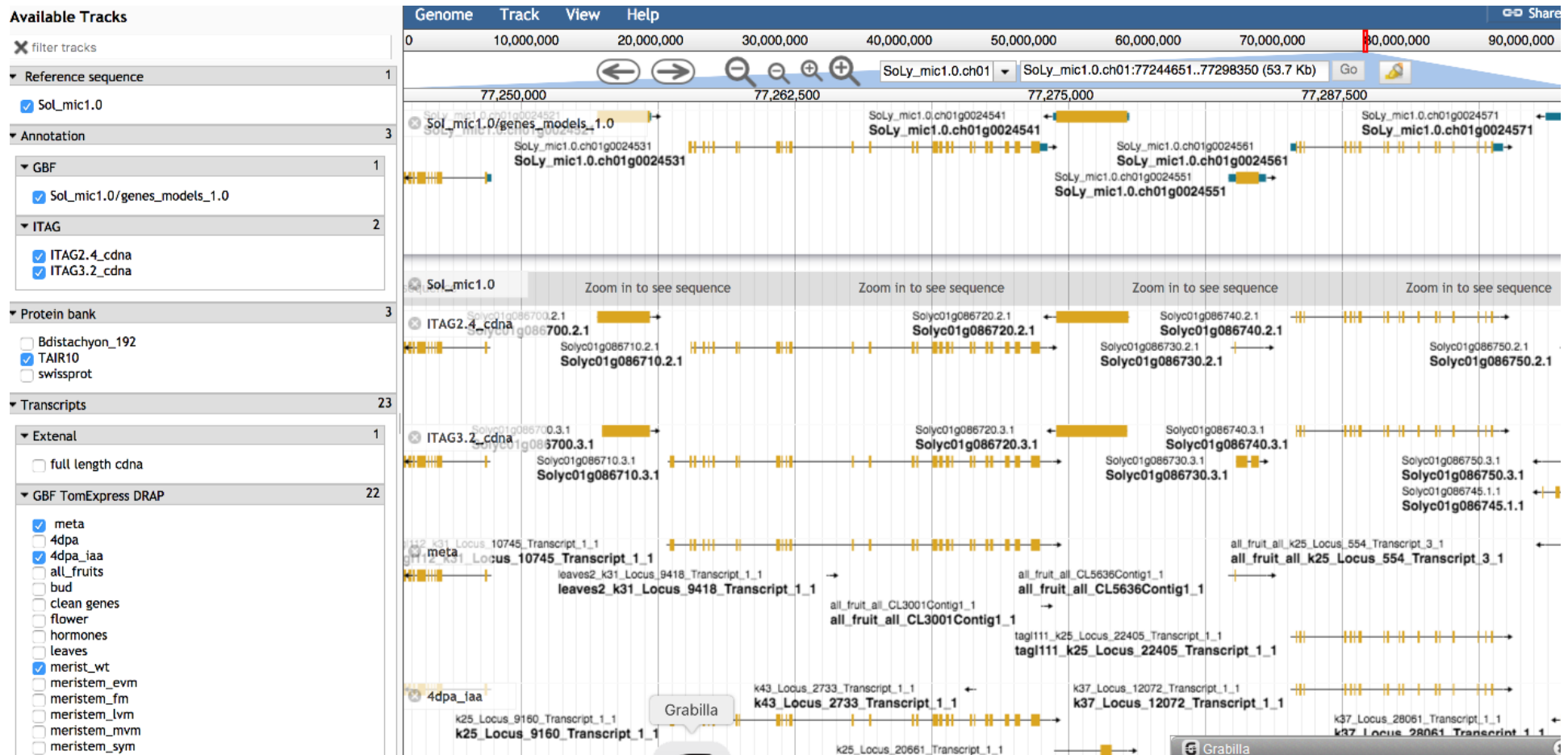
Laser capture

...



De novo assembly and curation of Rna-Seq data

Jbrowse: Exploring annotation



Gene models annotation summary

	SL2.5_ITAG2.4	SL3.0_ITAG3.2	SL_mic1.0/annot1.0
Total number of genes			39488
Number of protein coding genes	34725	35768	36381
% genes with 5' UTR	34%	49%	70%
% genes with 3' UTR	41%	51%	71%
Number of non coding gene			3107

Annotation QC using BUSCO

	SL2.5_ITAG2.4	SL3.0_ITAG3.2	SI_mic1.0/annot1.0	TAIR10
Complete BUSCOs	95,5%	94,5%	96,1%	99,5%
	1376	1361	1383	1433
Complete and single-copy BUSCOs	94,0%	92,5%	94,4%	75,9%
	1354	1332	1359	1093
Complete and duplicated BUSCOs	1,5%	2,0%	1,7%	23,6%
	22	29	24	340
Fragmented BUSCOs	2,8%	2,8%	1,6%	0,2%
	40	40	23	3
Missing BUSCOs	1,7%	2,7%	2,3%	0,3%
	24	39	34	4

Conclusion and perspectives

- New and high quality *de novo* genome assembly using pacBio, Bionano and 10x technologies
- One full chromosome have been obtained
- New GM annotation with high busco score
- Functional annotation is ongoing
- Need of manual curation of gene models and names
- Performing Hi-C sequencing (ongoing)
- *De novo* genome assembly of wild relatives is ongoing using 10x - ON - Bionano

Tomato genome *de novo* assembly of wild relatives is ongoing



S. lycopersicum



S. galapagense



S. cheesmaniae



S. pimpinellifolium



S. neorickii



S. chmielewskii



S. chilense



S. arcanum



S. corneliomulleri



S. huaylasense



S. peruvianum



S. habrochaites



S. pennellii



S. lycopersicoides



S. sitiens



S. ochranthum



S. juglandifolium

International project.

Acknowledgments

Genotoul-Bioinfo

Klopp Christophe

GetPlage

BOUCHEZ Olivier
Lopez-Roques Céline
Roulet Alain
Claire Kuchly
Eché Camille
Donnadieu Cécile

GBF Lab

BOUZAYEN Mondher
FRASSE Pierre
MAZA Elie
DJARI Anis
Clément Folgoa
Margot Zahm
Zouine Mohamed

LIPM

Gouzy Jérôme
Sallet Erika

CNRGV

Bergès Hélène
Marande William
Arribat Sandrine



