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

Mohamed Zouine

@GBF_Lab

Strategy and history

Start

2006

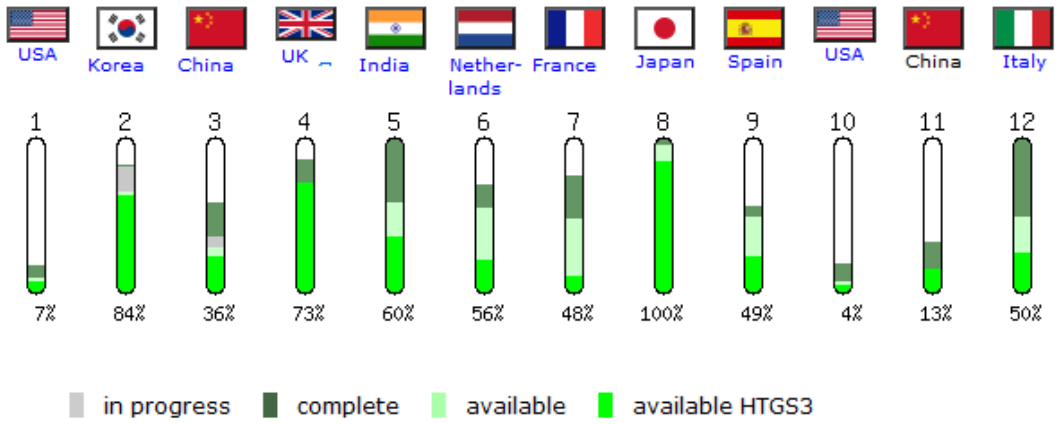
2007 — Phase I ➢ **BAC by BAC**

2008

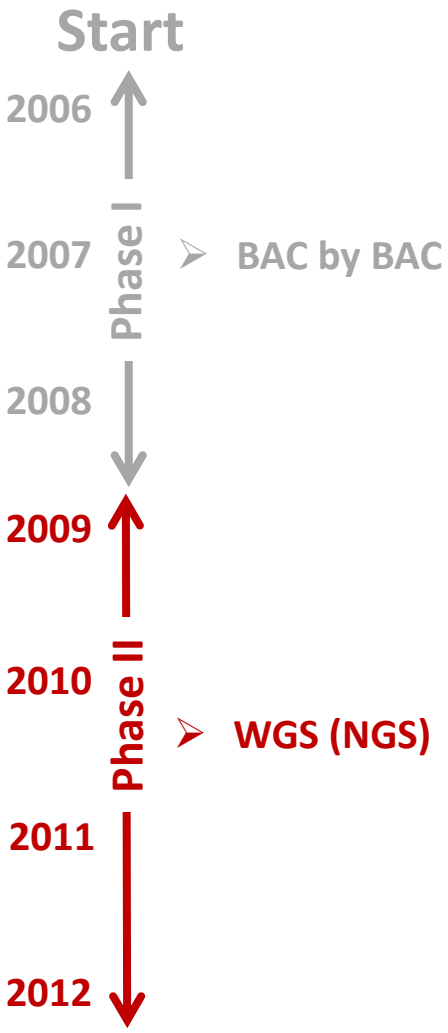
✓Genome
- 12 chromosomes
- 950 Mb genome
- ~220 Mb of contiguous gene-rich euchromatin

✓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
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 euchromatin
Physical map	10X BAC coverage

NGS

+

Two hybrid assemblies

CABOG
(GBF, Fr)

NEWBLER
(NI)

First tomato genome assembly v1.0

Start

2006

2007 — Phase I

- BAC by BAC

2008

2009 — Phase II

- WGS (NGS)
- 1st release: Soly1.00

2010

2011

2012

Soly1.00

sol genomics network

home | forum | contact | help

search | maps | sequencing | tools | sol search

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.

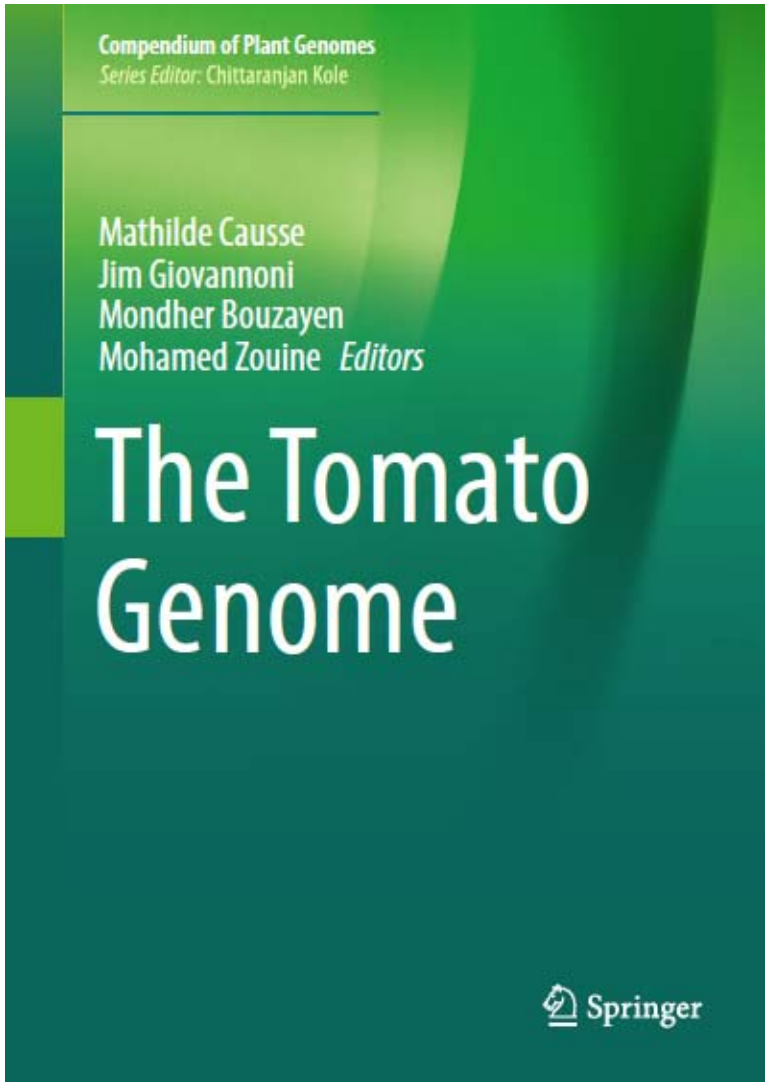
Disclaimers and data access agreement

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.

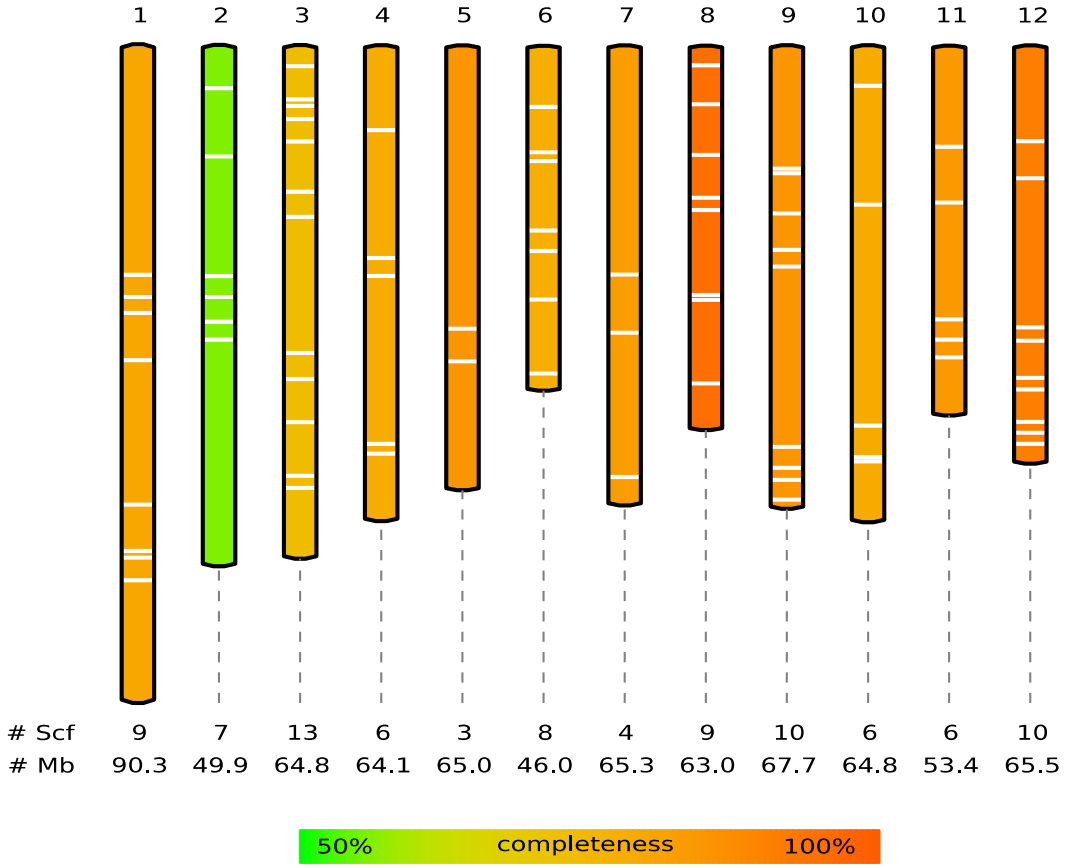
The pre-publication tomato genome sequence is made available with the understanding that users will respect the rights of those who contributed to this effort to describe the tomato genome in a peer-reviewed publication. This description includes whole genome level analyses on genes, gene families, repetitive 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

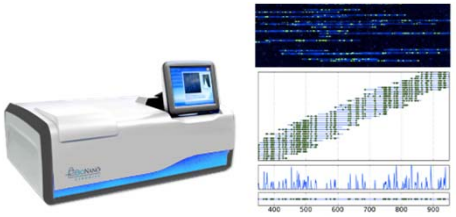
-Decrease Chromosome zero size
-Improve structural annotation

Combining Long read sequencing technologies

PacBio

Long contigs

+



BioNano

Long Scaffolds

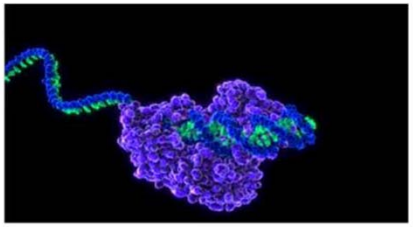
10x Genomics

Super-scaffolding & error correction

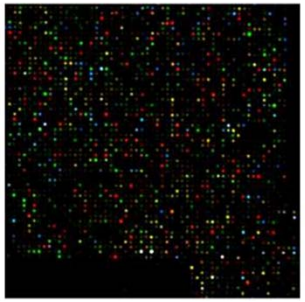
Toward a new and improved tomato genome reference sequence ?

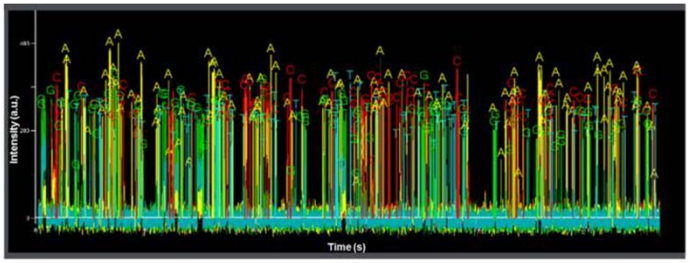
PacBio technology

Single Molecule, Real-Time (SMRT) DNA Sequencing

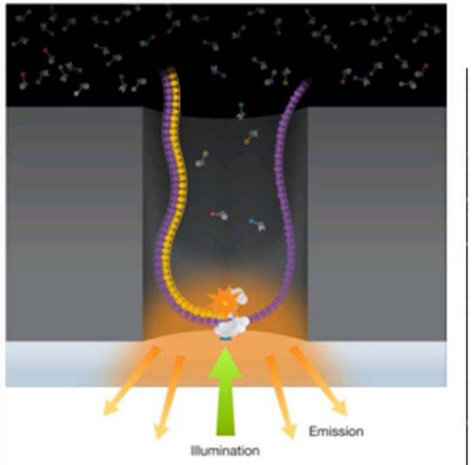


PacBio *RS* II





SMRT® Cells

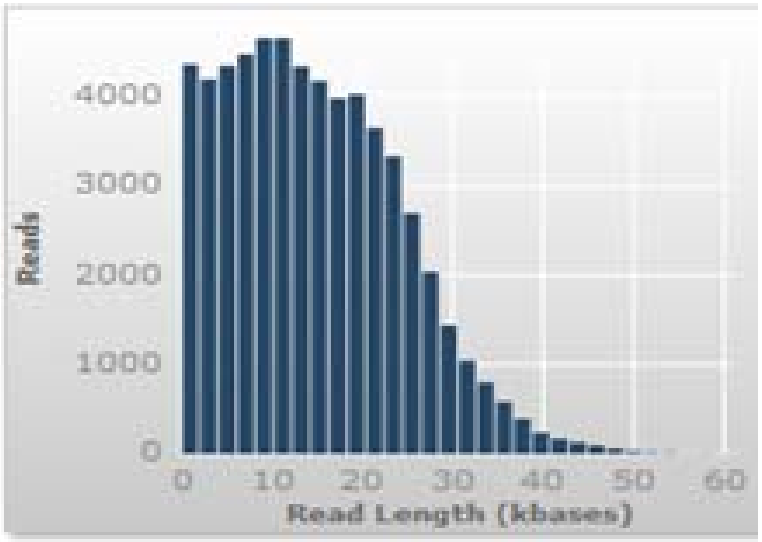


Science, Vol 299, Jan 31 2003, pp682-68

J. Appl. Phys. 103, 034301 (2008)

PacBio sequence Assembly

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

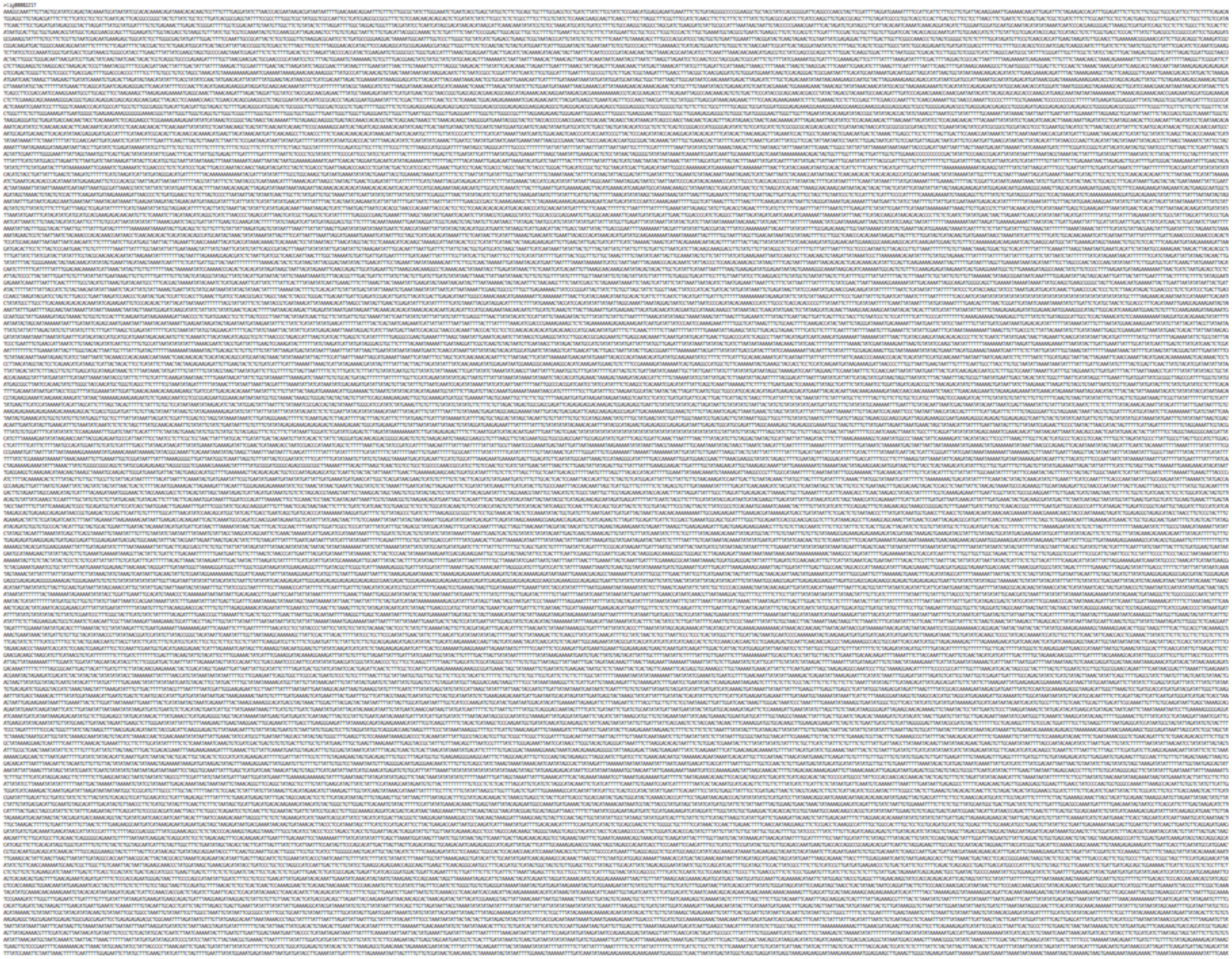


Assembly with CANU:

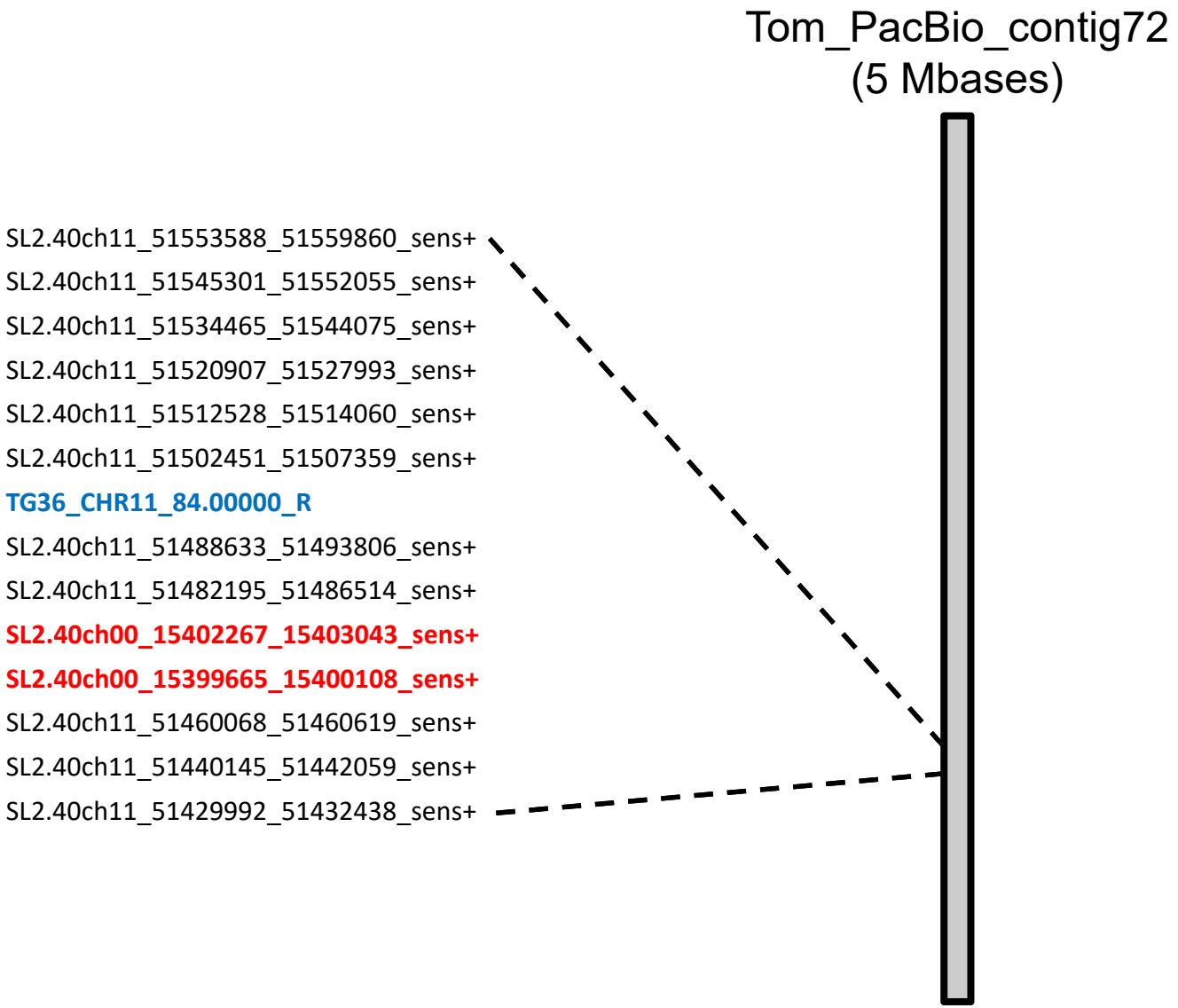
Number of contigs: **743** **(110 872 contigs in SL1.0)**
Genome size : **800 Mb**
Longest contig: **13 Mb**
N50 length: **3.4 Mb**
N50 index: **62**

Example of region identified by QTL mapping
In the SL2.50 assembly

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

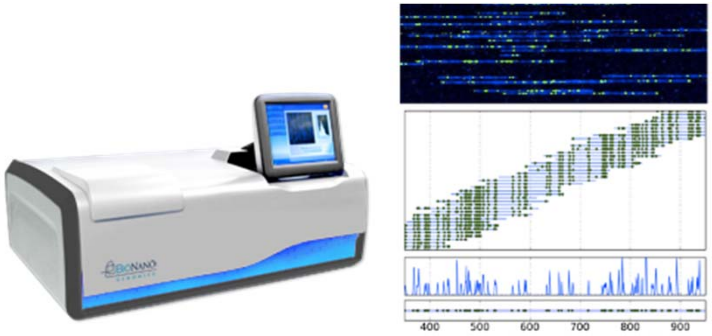


iTAG2.5 genes & genetic markers mapping



Scaffolding

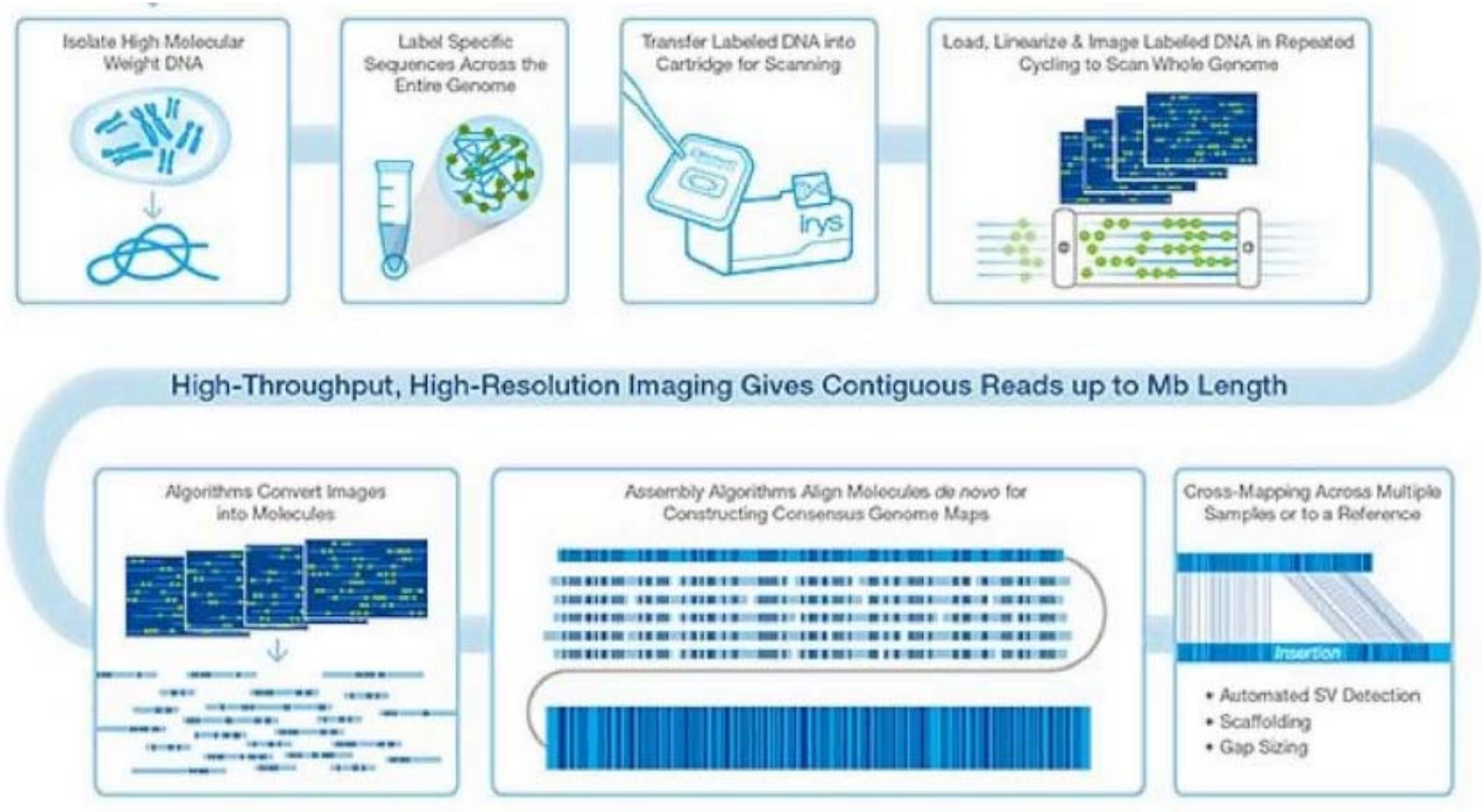
Optical Mapping
BioNano



+

Long read
sequencing
Chromium 10x

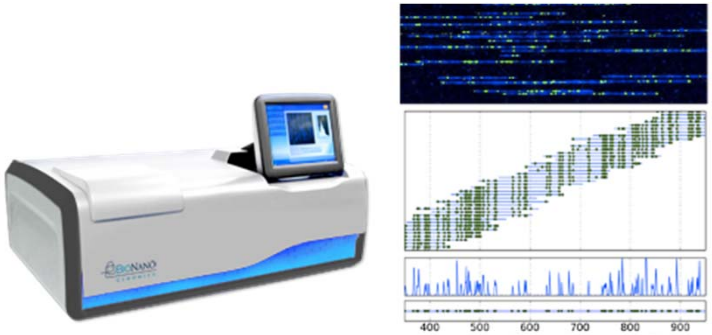
Optical Mapping using BioNano



www.youtube.com/watch?v=XwBI13Q4ilo

Scaffolding

Optical Mapping
BioNano



Long read
sequencing
Chromium 10X

Two enzymes, 200X coverage

N50: 34 Mb

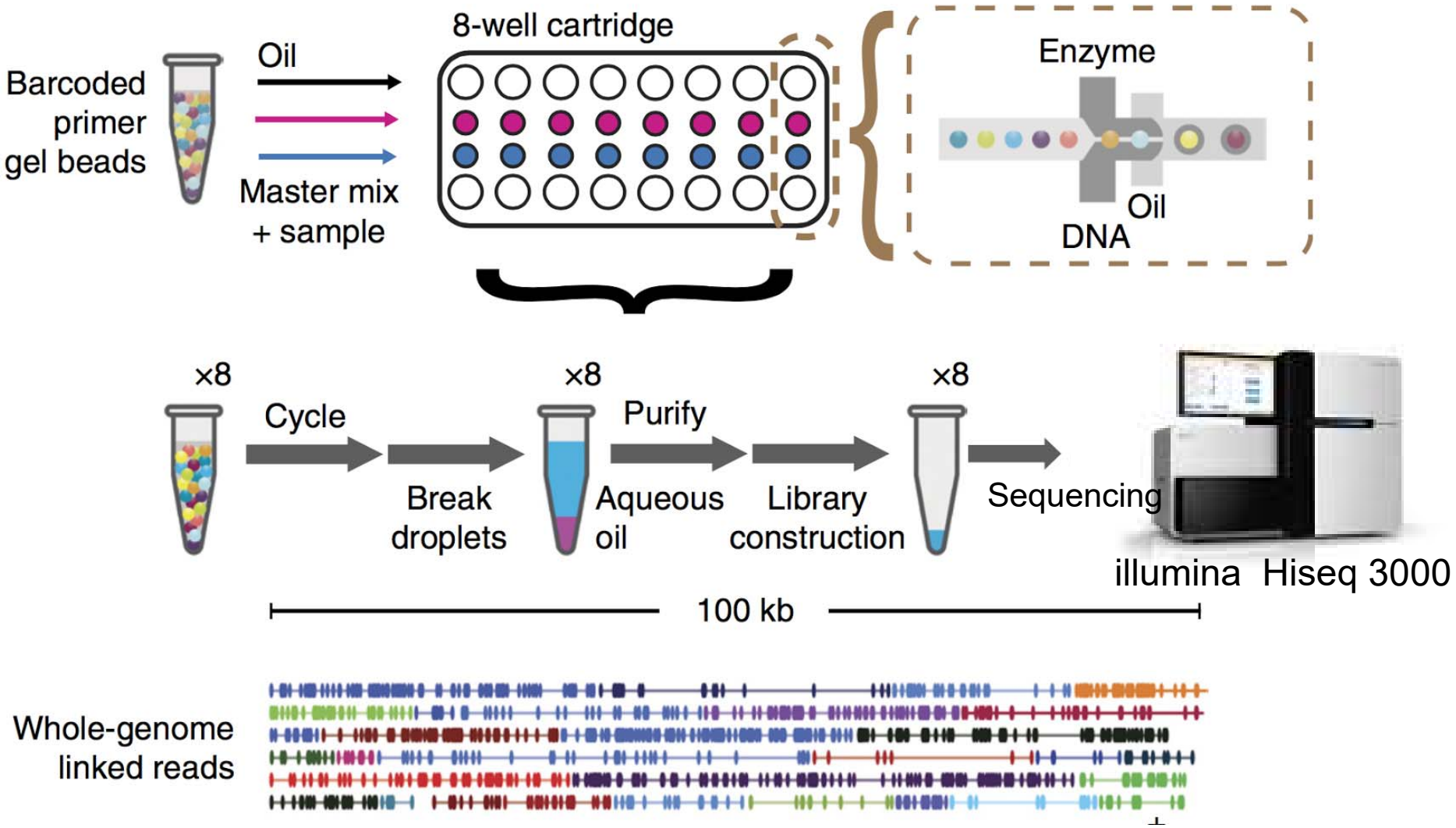
99% of the Genome in

40 scaffolds

+

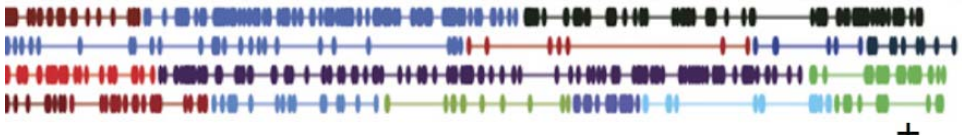
?

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**

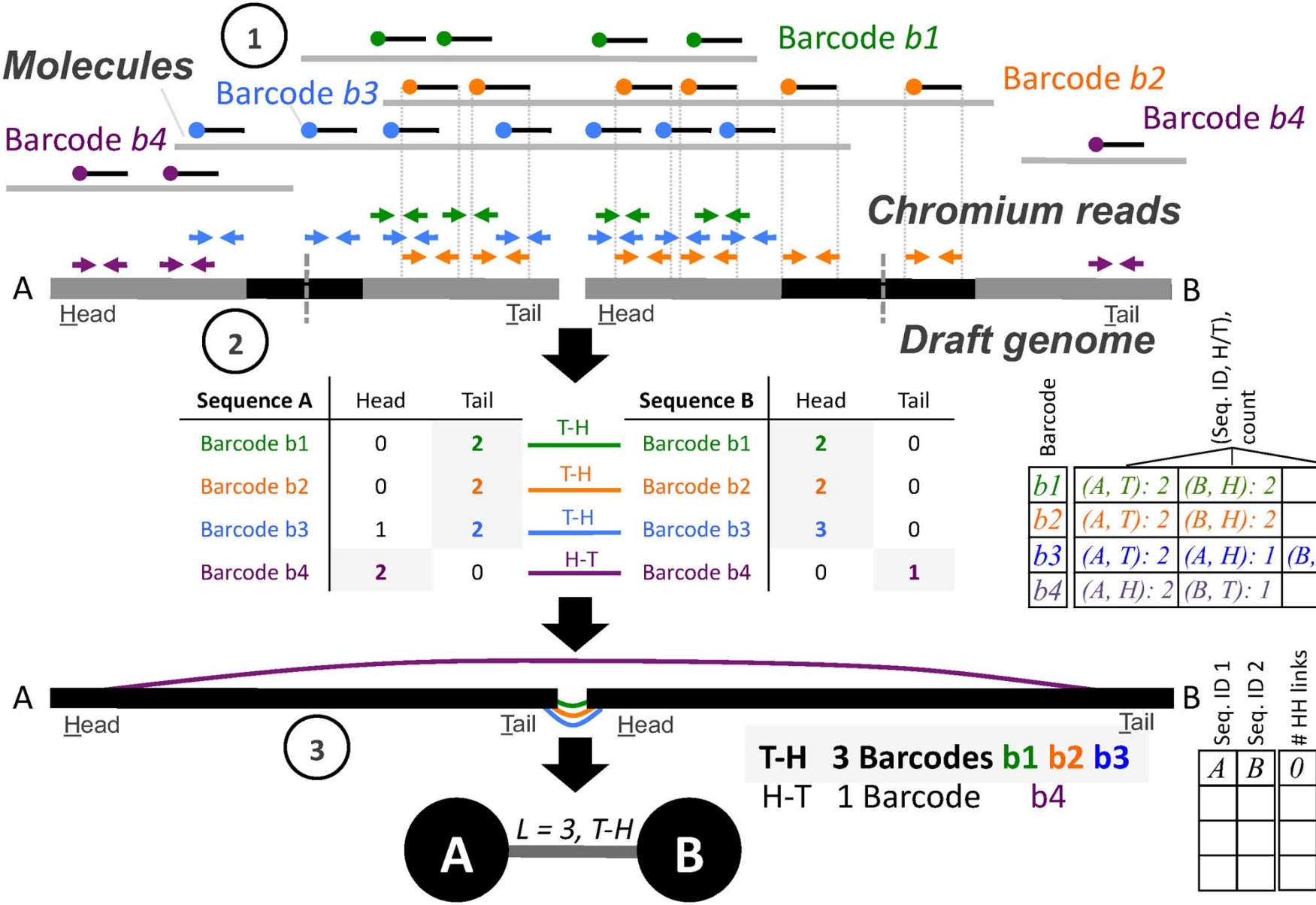
Scaffolding with 10x using ARCS

ARCS: Assembly Roundup by Chromium Scaffolding

Sarah Yeo[1], Lauren Coombe[1], Justin Chu, René L Warren[1,2], and Inanç Birol

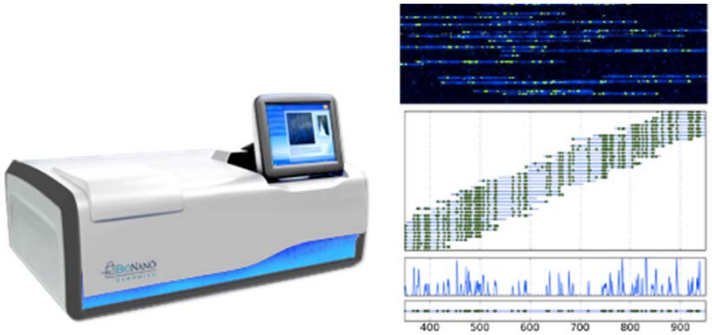
BC Cancer Agency, Genome Sciences Centre, Vancouver, BC V5Z 4S6, Canada
[1]Authors contributed equally
[2]corresponding author: rwarren@bcgsc.ca

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

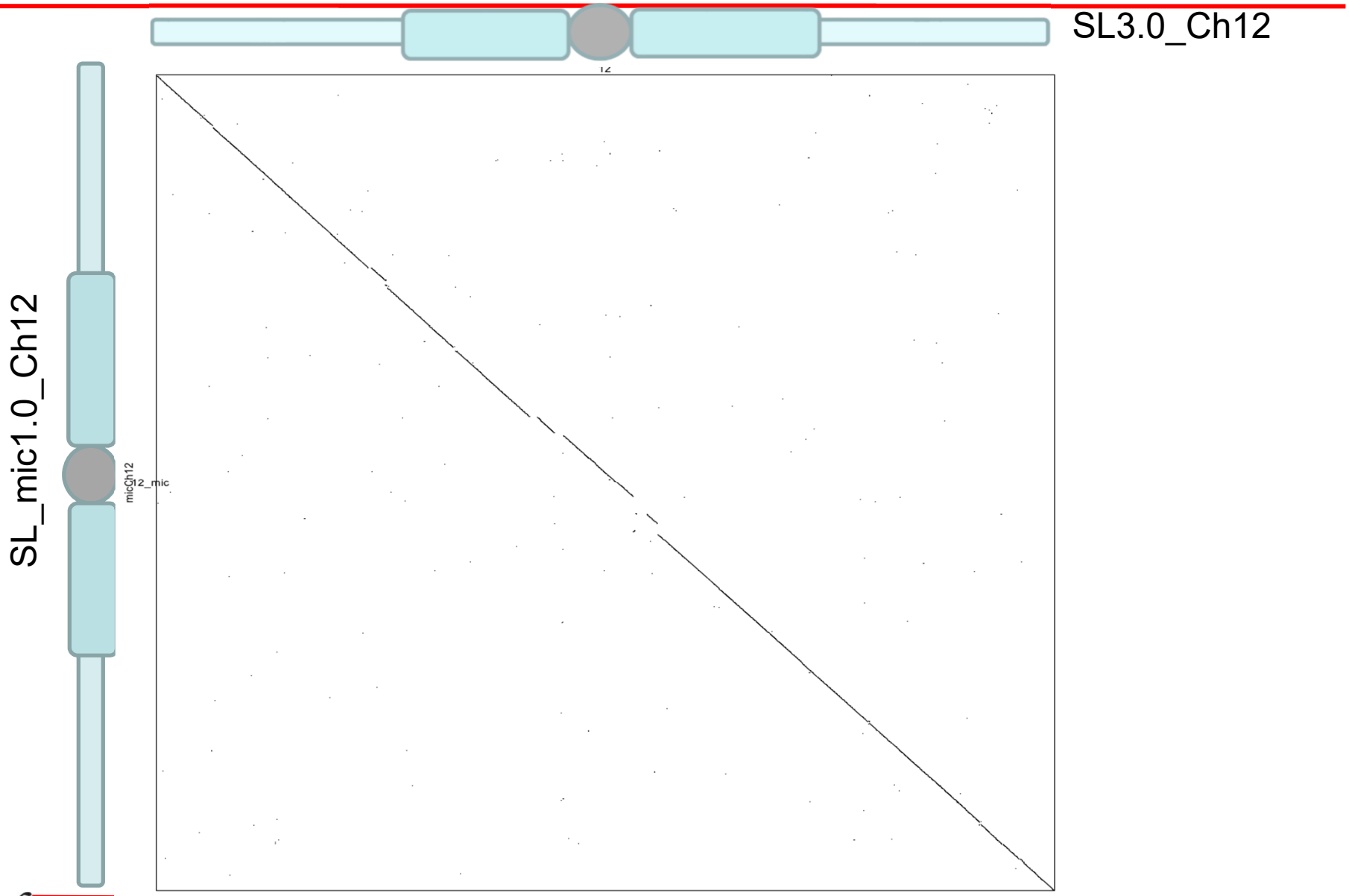
100X illumina coverage

N50: 45 Mb

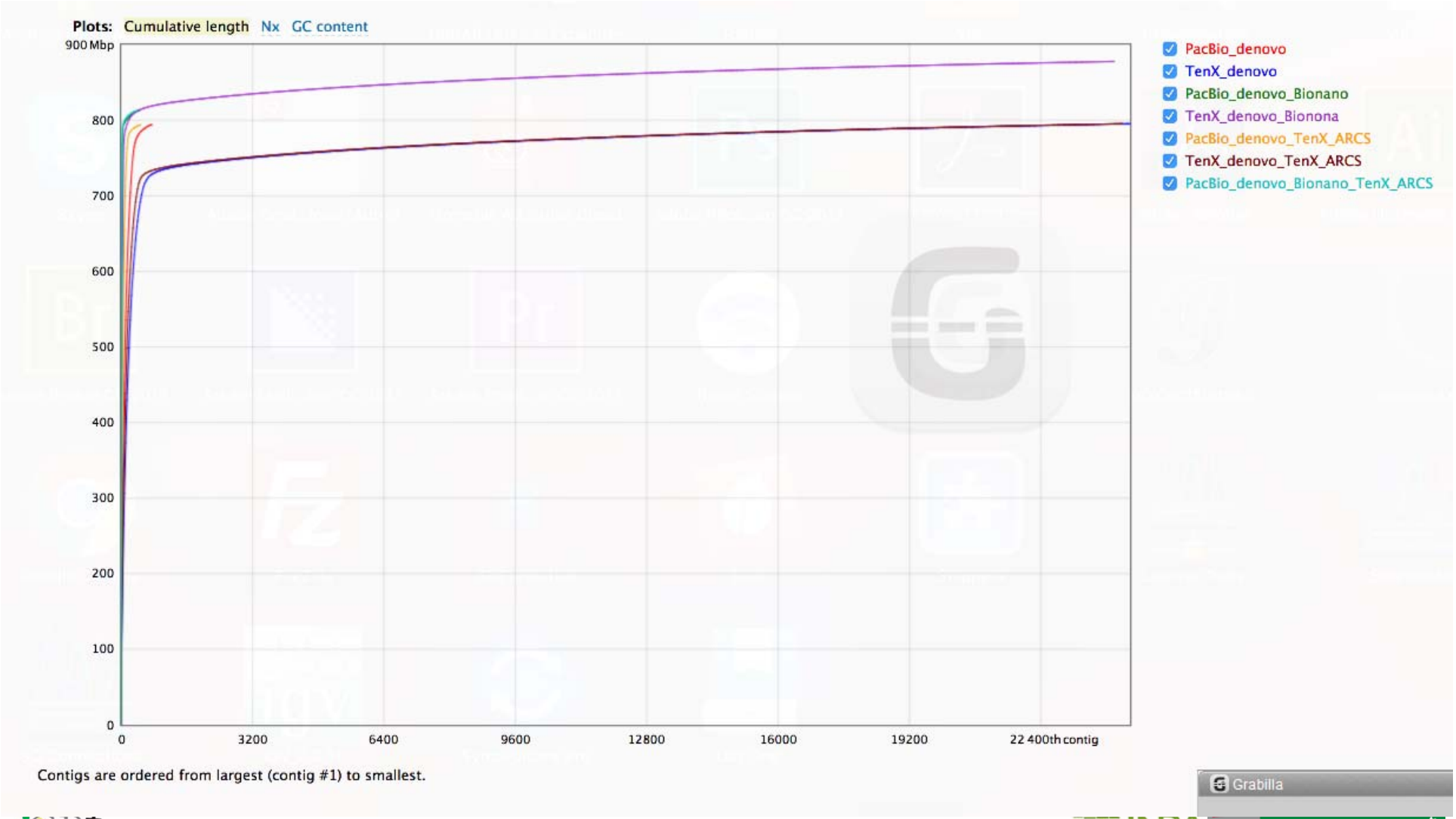
99% Ggenome in

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_ Bionona	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	€€€	€	€€€€	€€	€€€	€	€€€€€

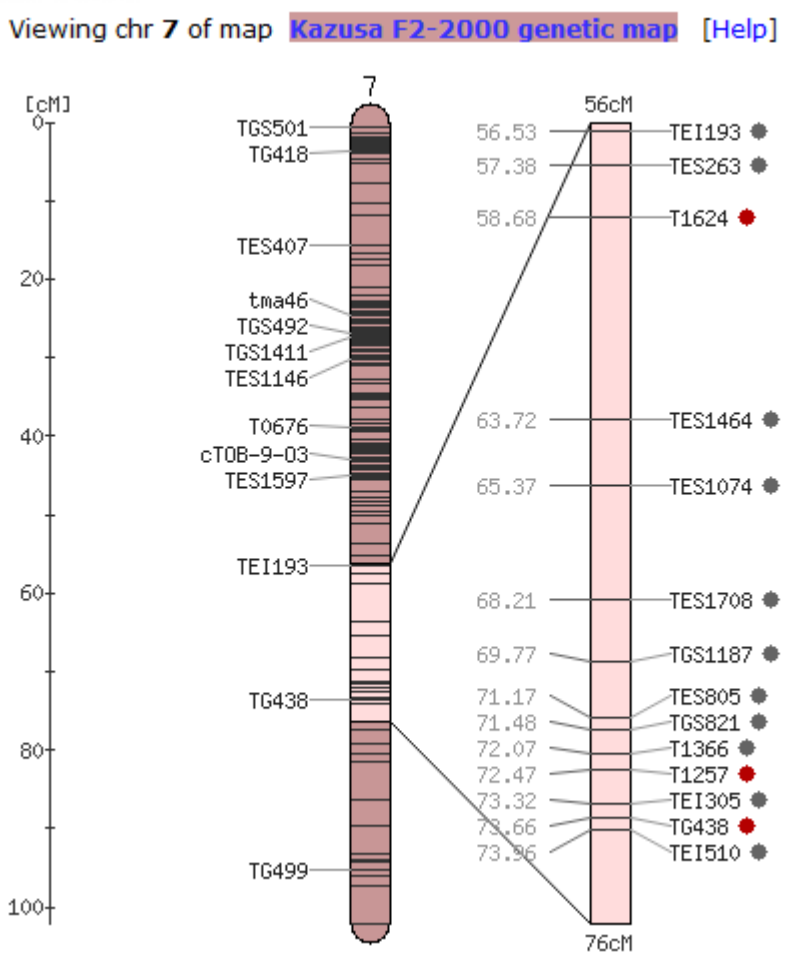
Anchoring scaffolds on chromosomes:
Pseudomolecules construction

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

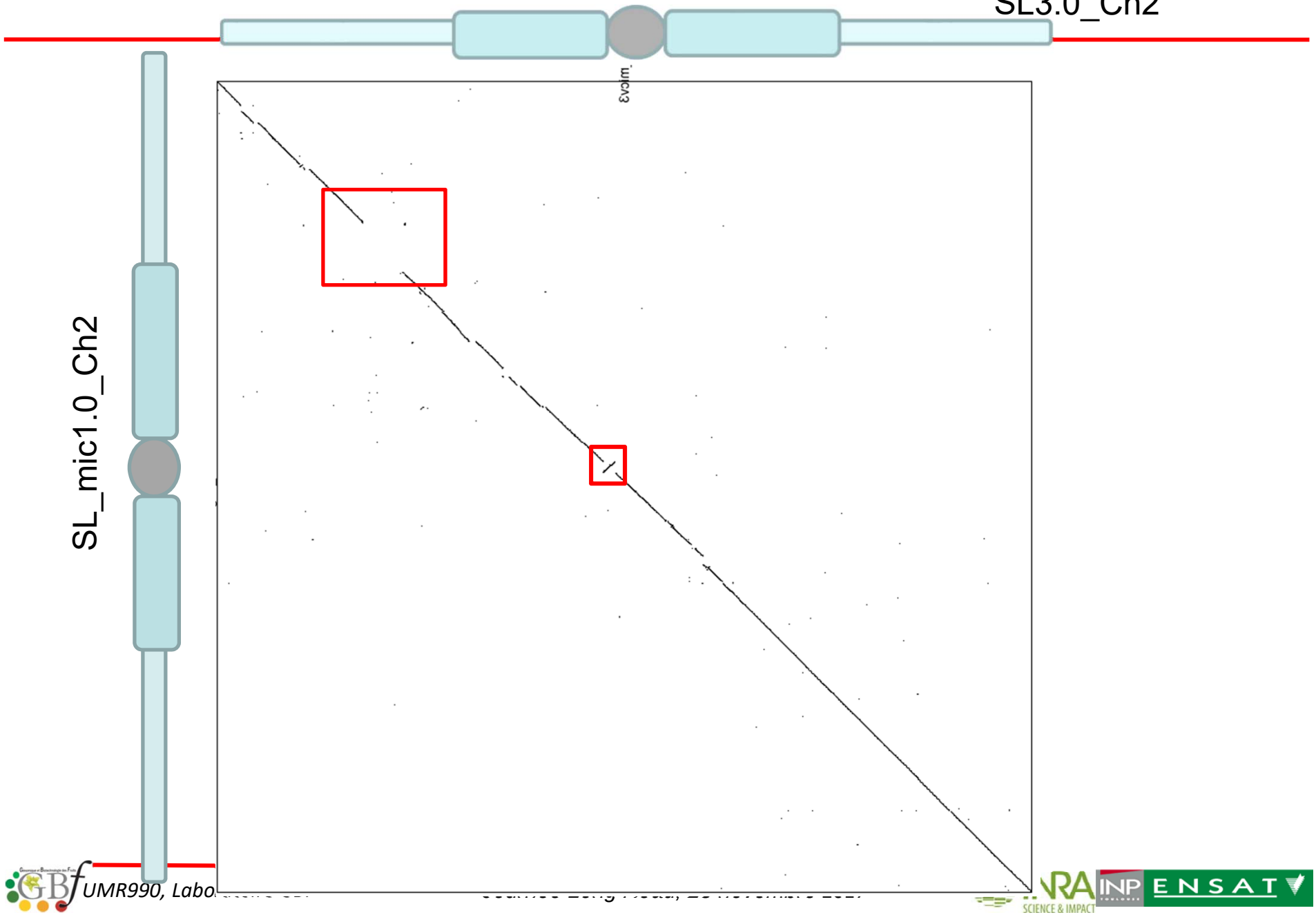
	markers
Chromosome 1	250
Chromosome 2	182
Chromosome 3	191
Chromosome 4	171
Chromosome 5	160
Chromosome 6	156
Chromosome 7	148
Chromosome 8	149
Chromosome 9	179
Chromosome 10	158
Chromosome 11	156
Chromosome 12	170
Total:	2070

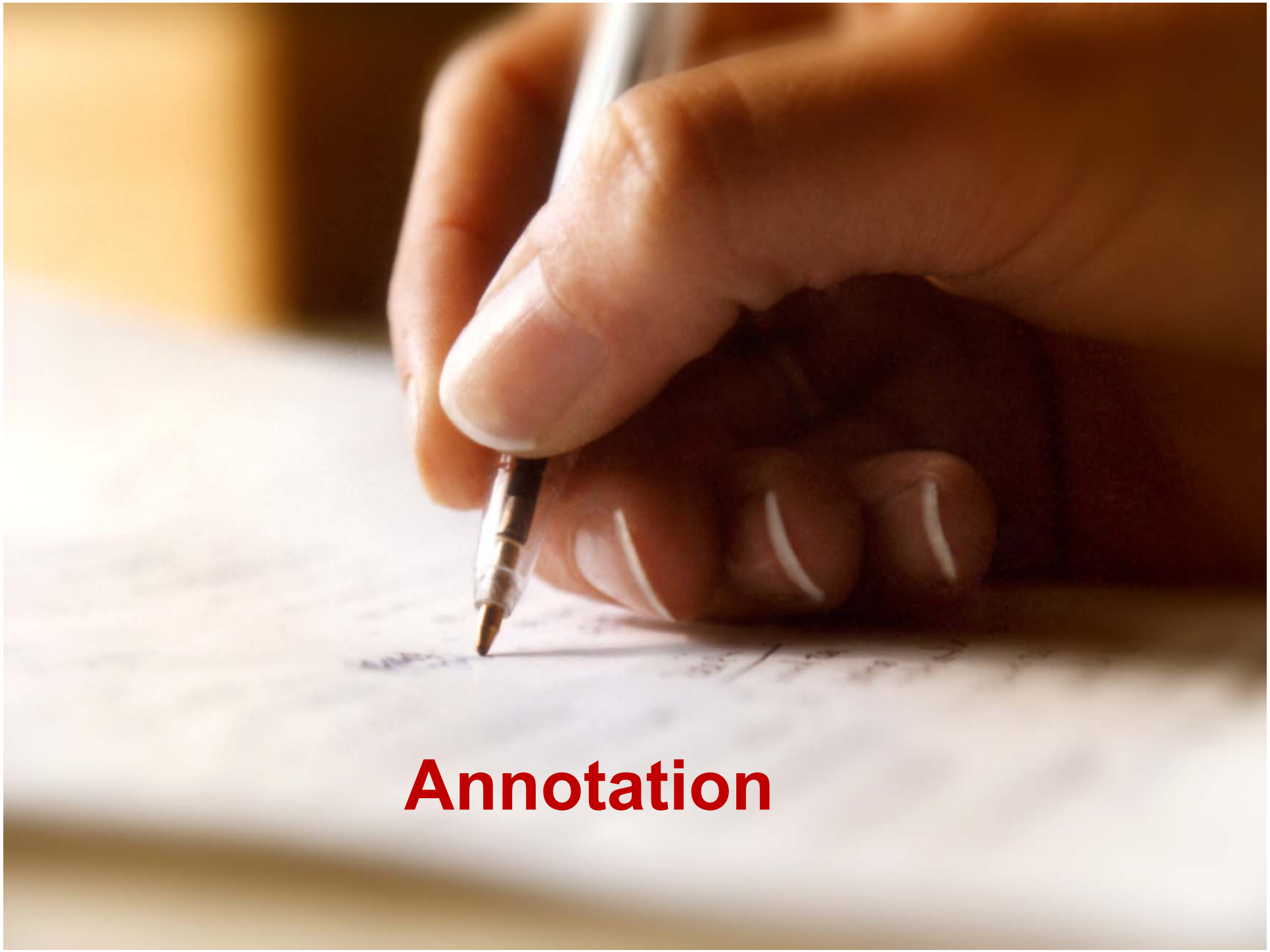
Marker collections	
COS	195
COSII	111
Total:	**306**

Protocols	
SNP	138
SSR	1250
unknown	682
Total:	**2070**

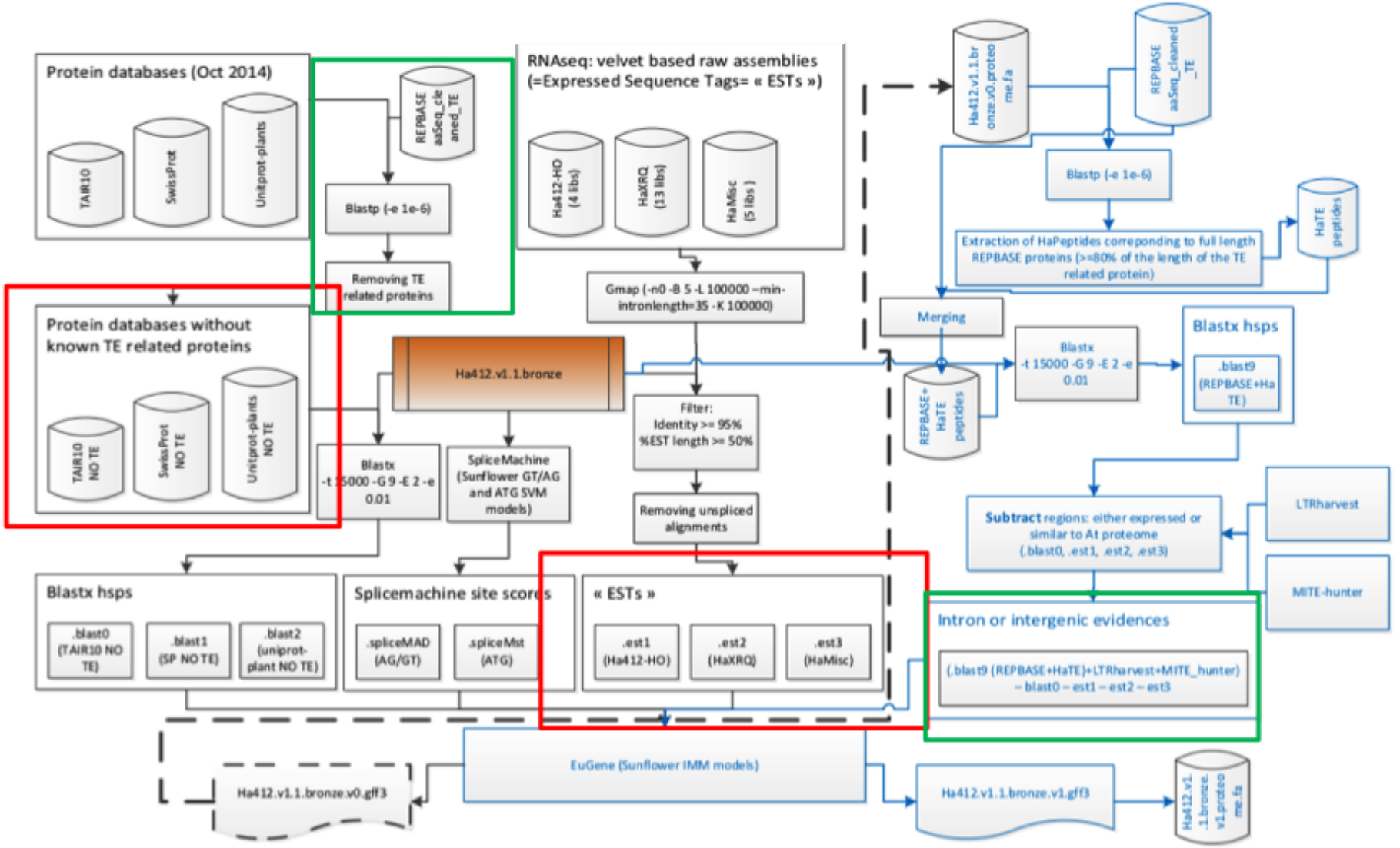


Chromosome 2




Annotation

Tomato EugenEP pipeline



Annotation by transcripts evidence using TomExpress data

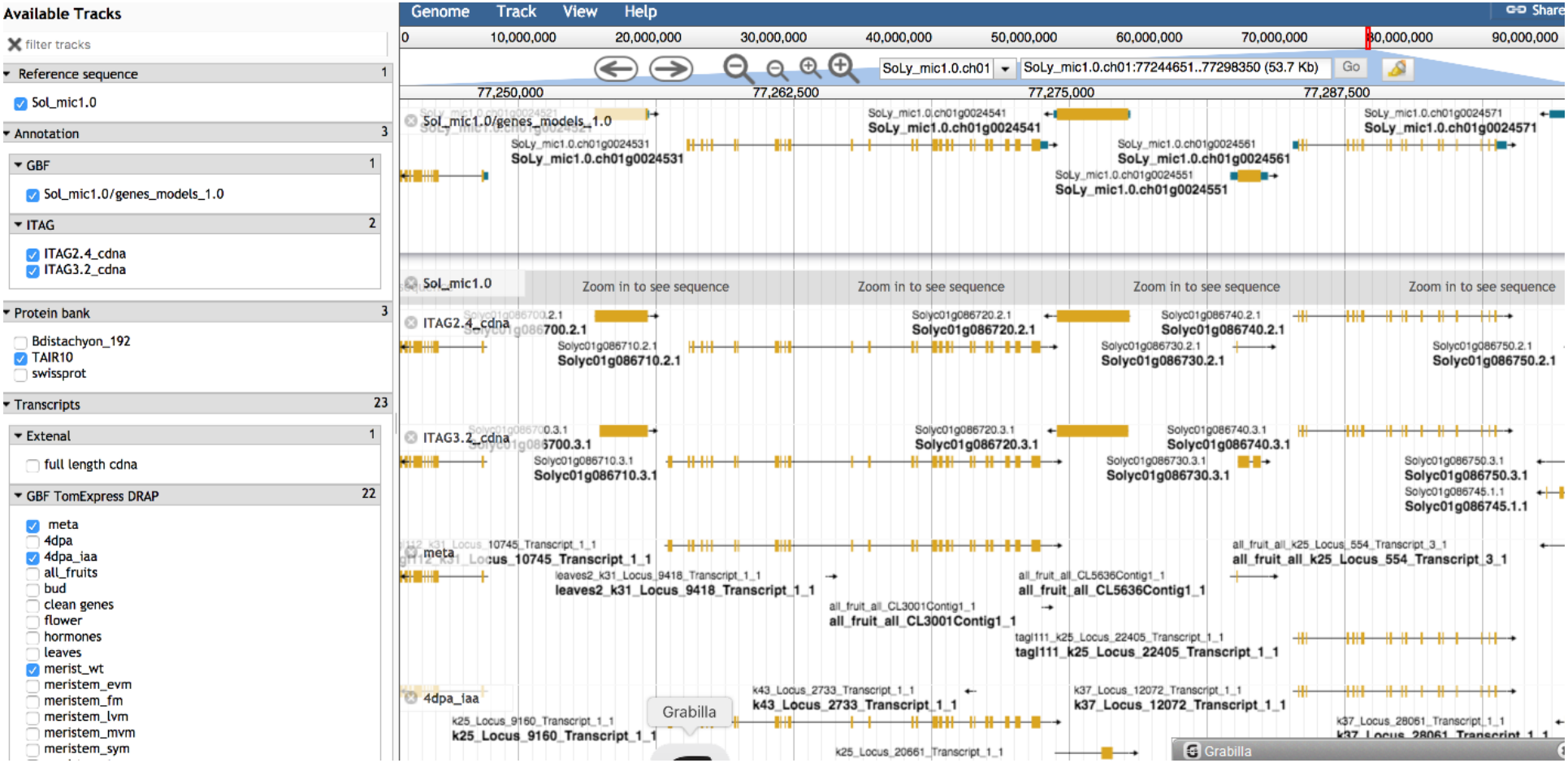
TomExpress, a unified tomato RNA-Seq platform for visualization of expression data, clustering and correlation networks
Zouine et al., 2017, the Plant Journal (online)

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

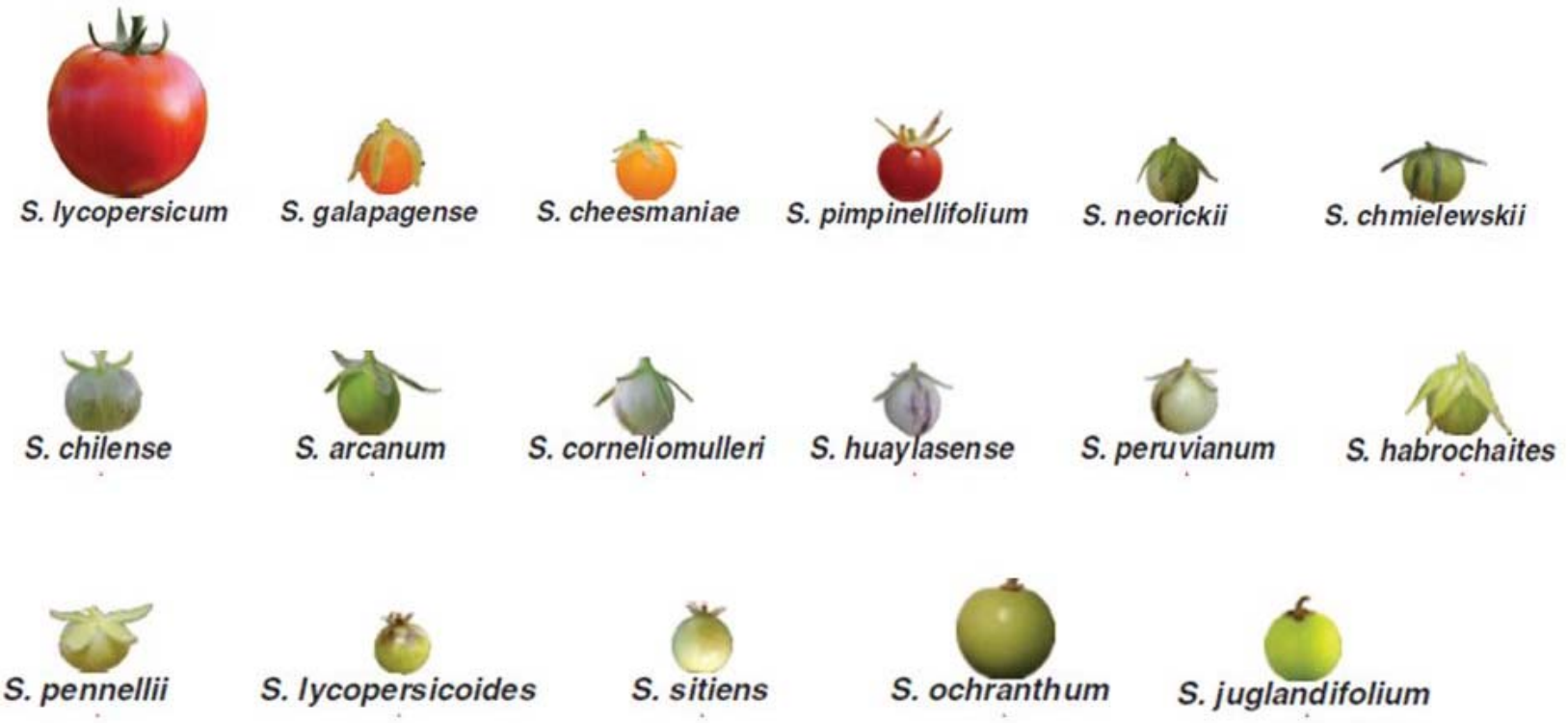
	SL2.5_ITAG2.4	SL3.0_ITAG3.2	Sl_mic1.0/annot1.0	TAIR10
Complete BUSCOs	95,5%	94,5%	96,1%	99,5%
	1376	1361	1383	1433
Complete and single-copy BUSCOs	94,0%	92,5%	94,4%	75,9%
	1354	1332	1359	1093
Complete and duplicated BUSCOs	1,5%	2,0%	1,7%	23,6%
	22	29	24	340
Fragmented BUSCOs	2,8%	2,8%	1,6%	0,2%
	40	40	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



International project.

Acknowledgments

Genotoul-Bioinfo

Klopp Christophe

GBF Lab

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

GetPlage

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

LIPM

Gouzy Jérôme
Sallet Erika

CNRGV

Bergès Hélène
Marande William
Arribat Sandrine

LABEX TULIP

HORIZON 2020

TomGEM

Agence Nationale de la Recherche ANR

TomEpiSet