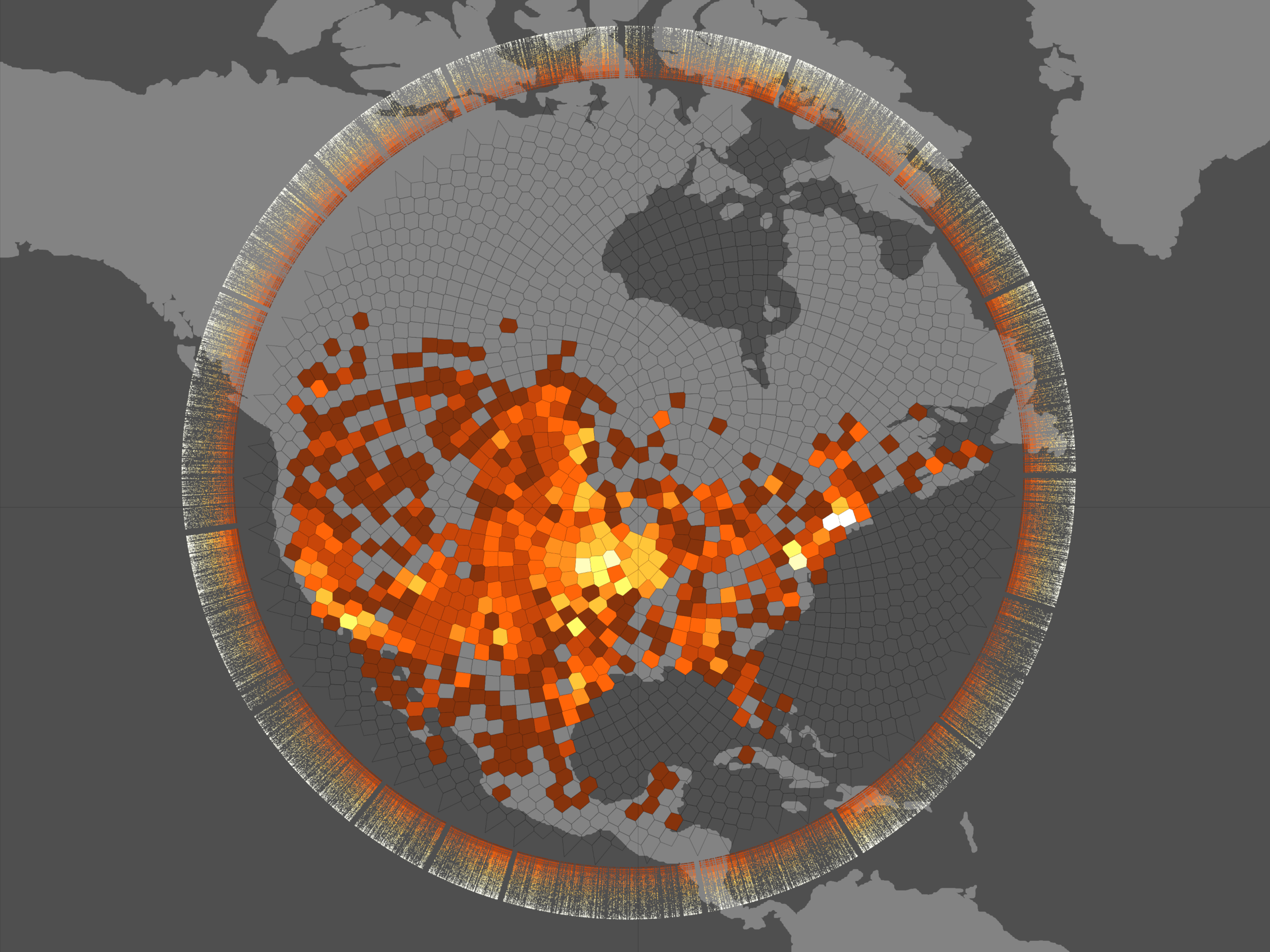
Sunflower Genome

2015-01-14

Chris Grassa

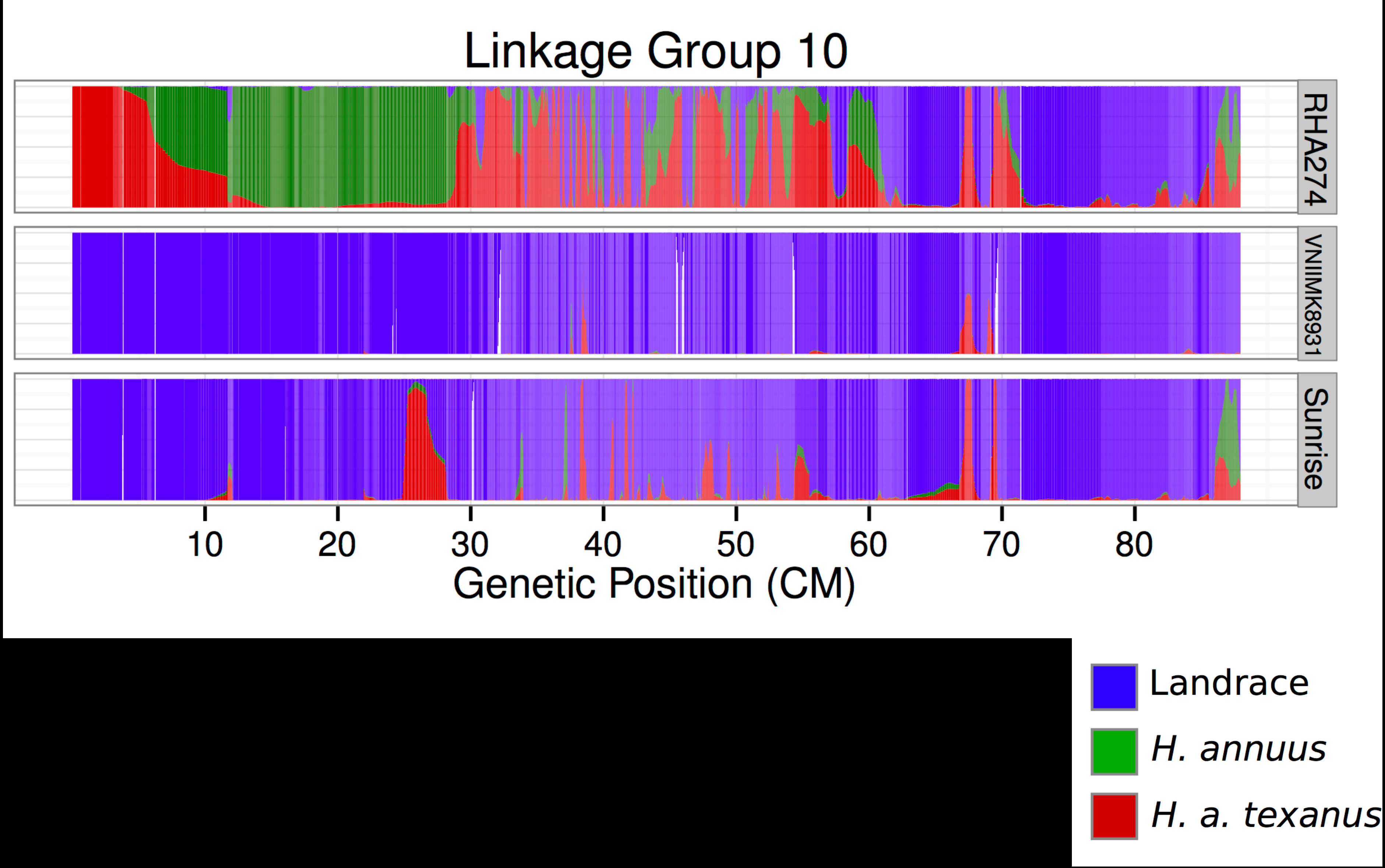


Closest relative of domesticated sunflower

- Woody
- Downy-mildew resistance
- Bi-modal flowering time

Helianthus argophyllus (coastal)
Brook Moyers

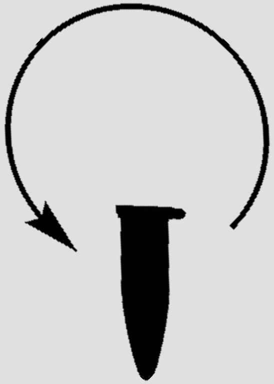
Domestication & Improvement



Linkage Group 10
RHA274
VNIIMK8931
Sunrise
10
20
30
40
50
60
70
80
Genetic Position (CM)
Landrace
H. annuus
H. a. texanus

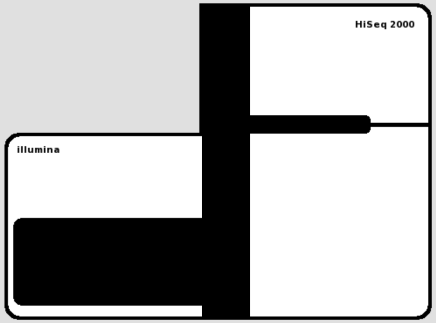
Mitochondrial Genome Assembly

9-day-old
seedlings

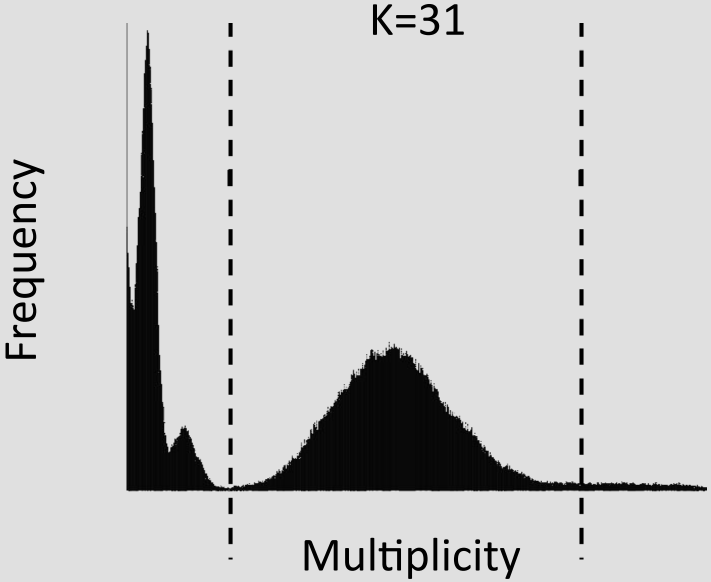


mitochondria
enriched

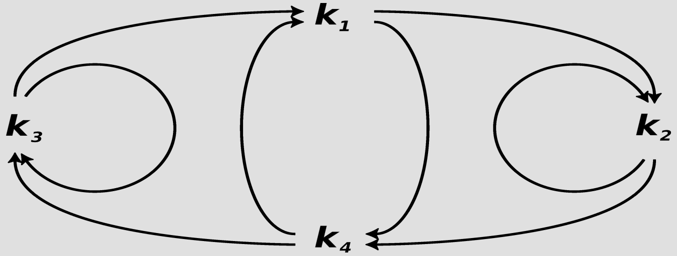
Whole-Genome-Shotgun
DNA library



1/48th
lane



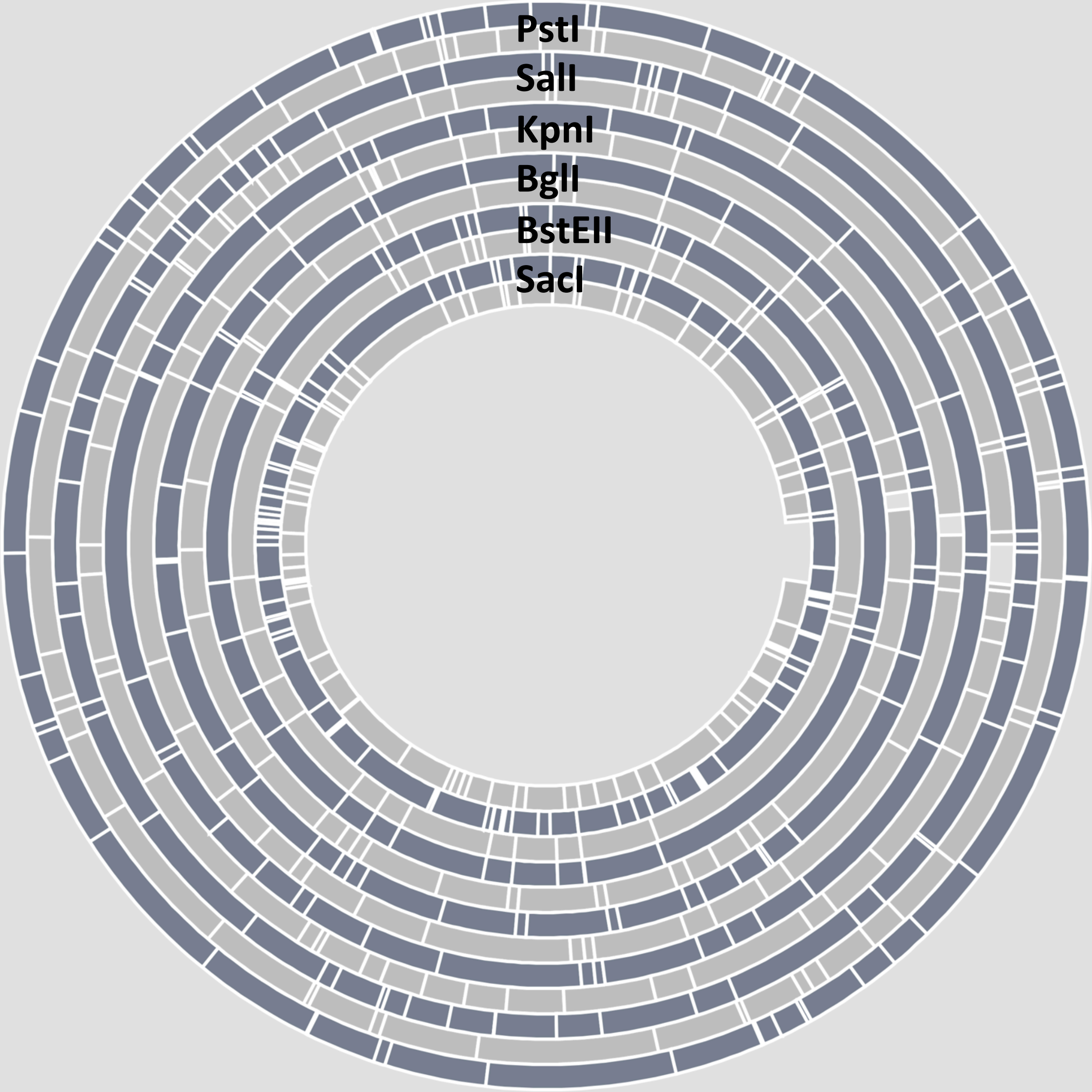
K-mer frequency partitioning

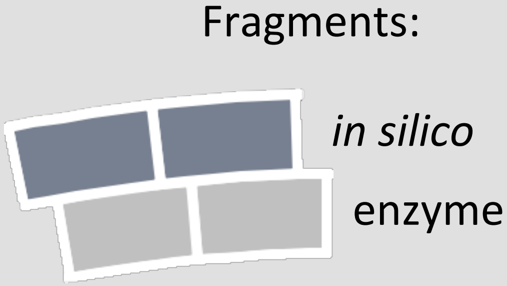


De Bruijn graph assembly

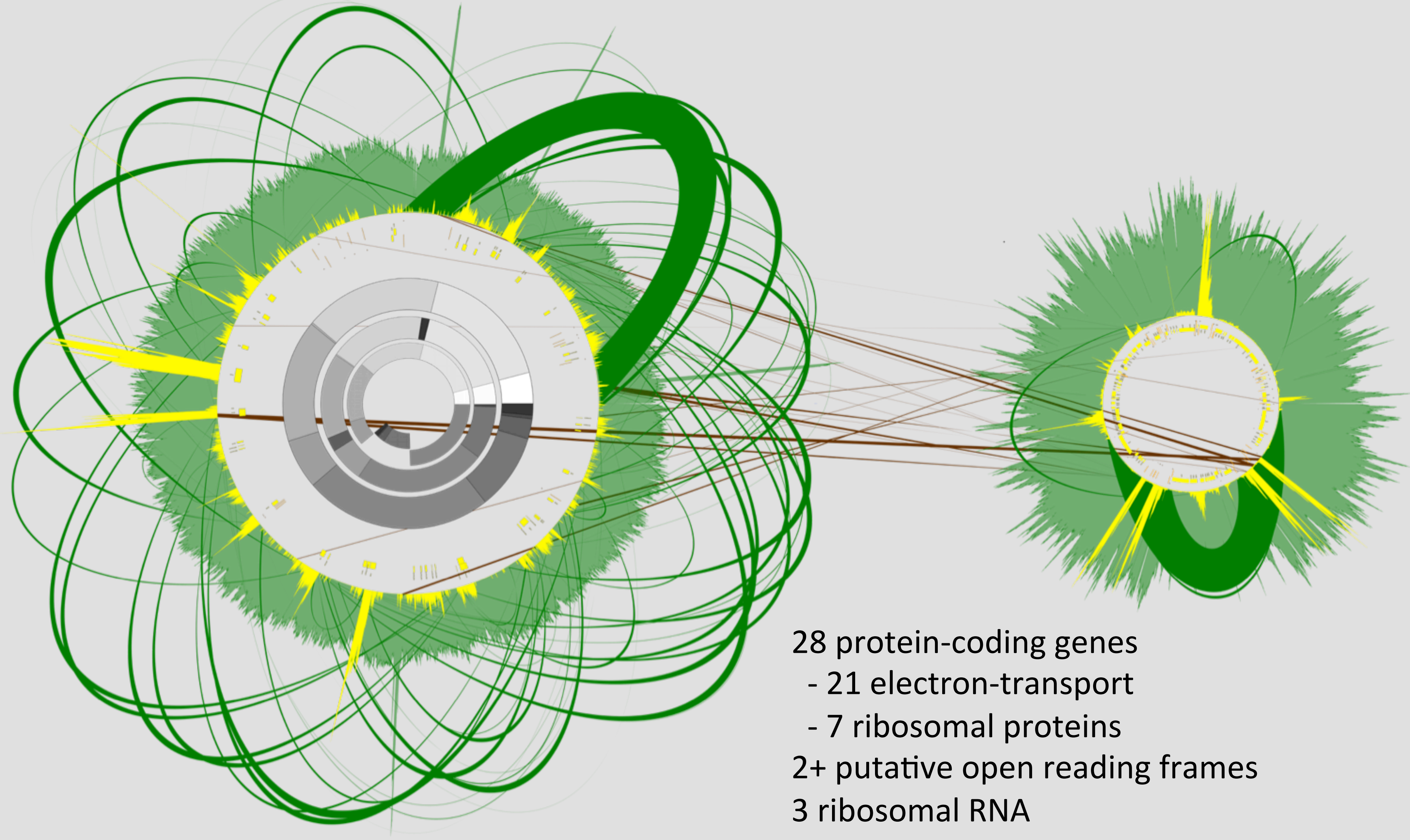
Dan Ebert

Restriction site maps





Siculelal & Palmer, 1988

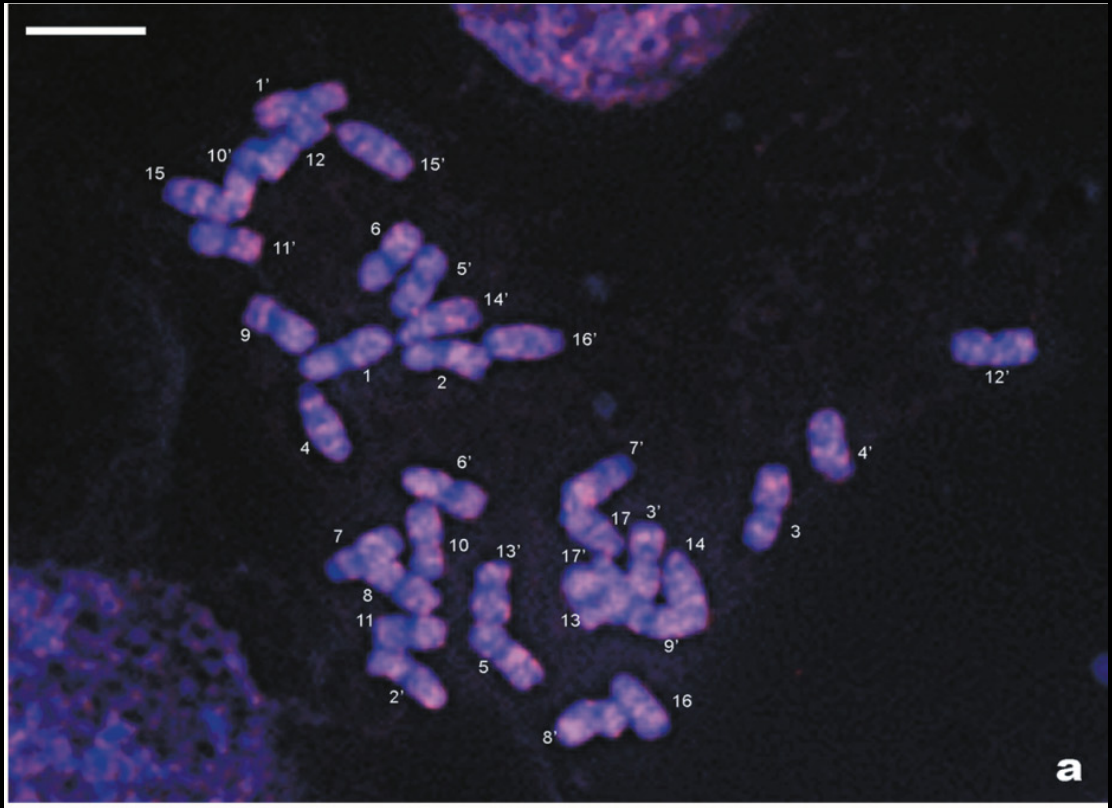


Mitochondrial Genome

Plastid Genome

28 protein-coding genes
  - 21 electron-transport
  - 7 ribosomal proteins
2+ putative open reading frames
3 ribosomal RNA
14 tRNA
11.87% of genome

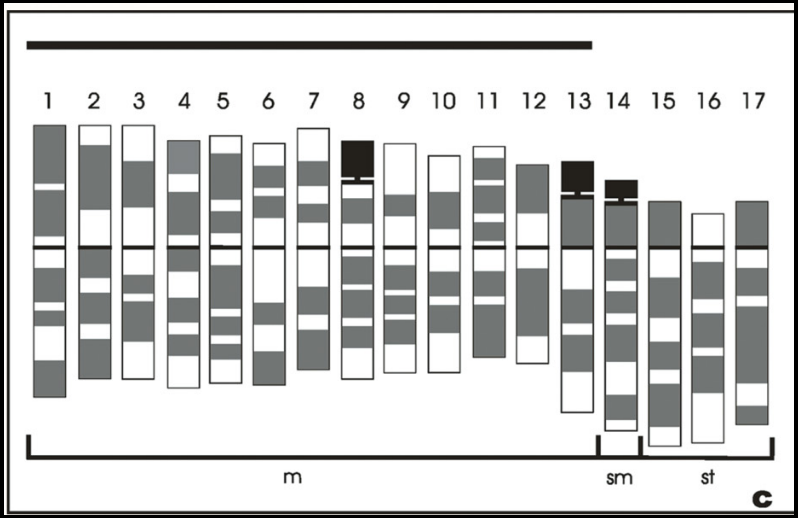
Bars represent 10 μm.

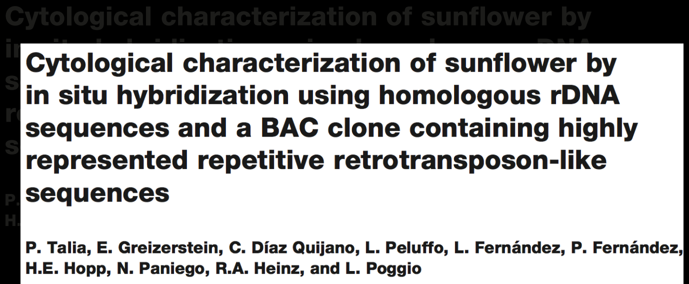


Nuclear Genome:

~3.6 Gbp

17 chromosomes

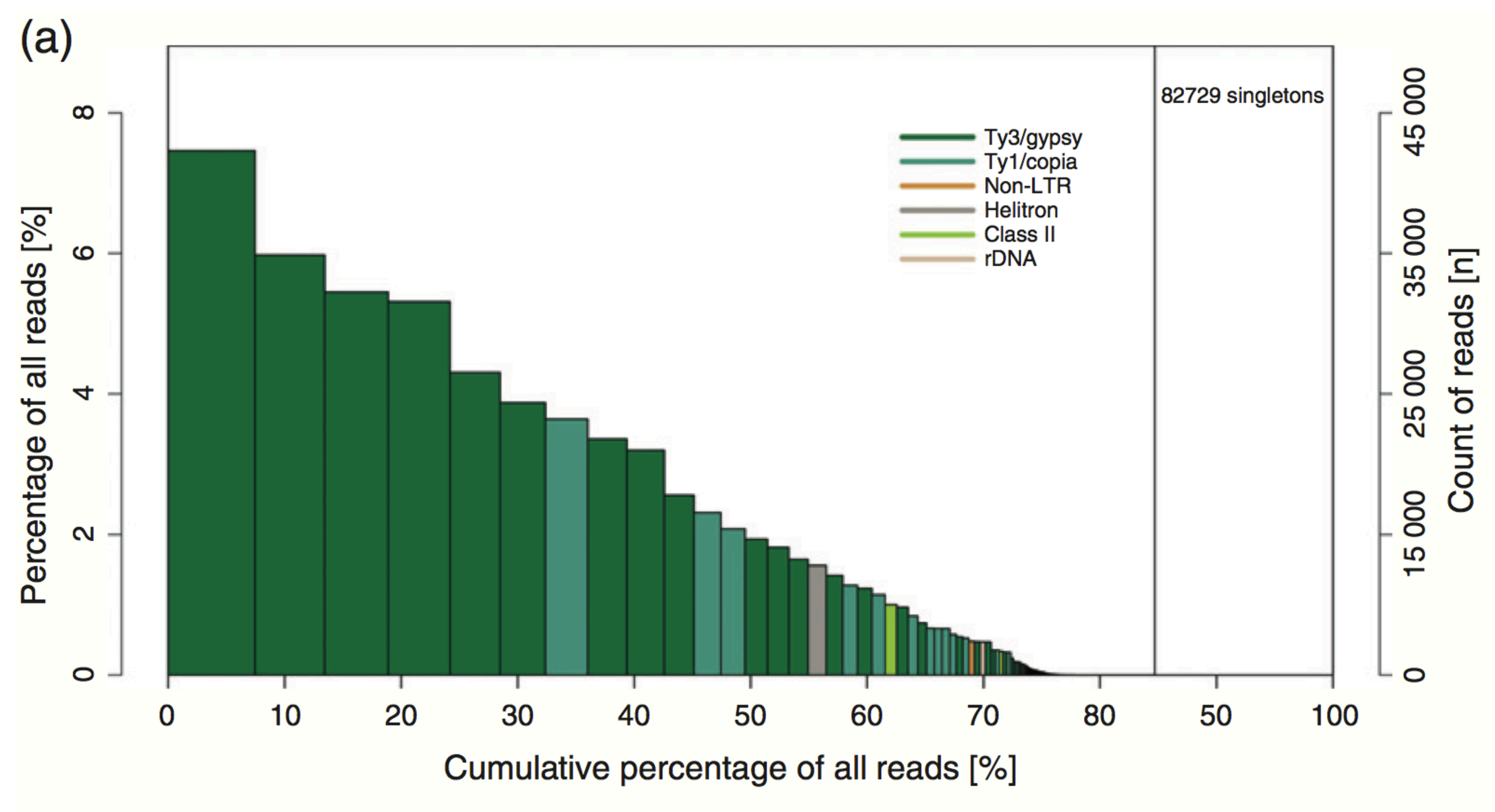



Cytological characterization of sunflower by

Cytological characterization of sunflower by in situ hybridization using homologous rDNA sequences and a BAC clone containing highly represented repetitive retrotransposon-like sequences

P. Talia, E. Greizerstein, C. Díaz Quijano, L. Peluffo, L. Fernández, P. Fernández, H.E. Hopp, N. Paniego, R.A. Heinz, and L. Poggio

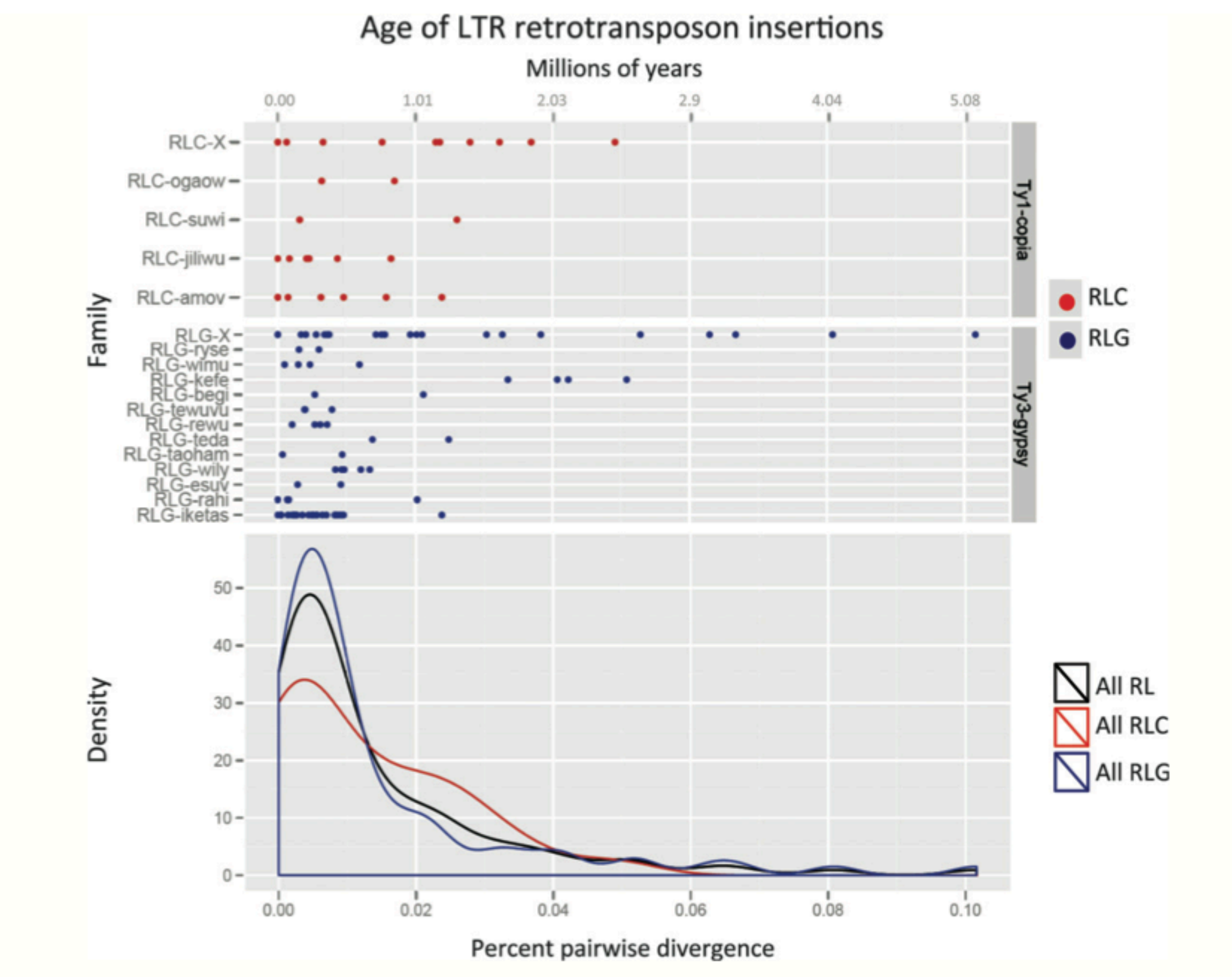
Highly Repetitive Genome



The sunflower (*Helianthus annuus* L.) genome reflects a recent history of biased accumulation of transposable elements

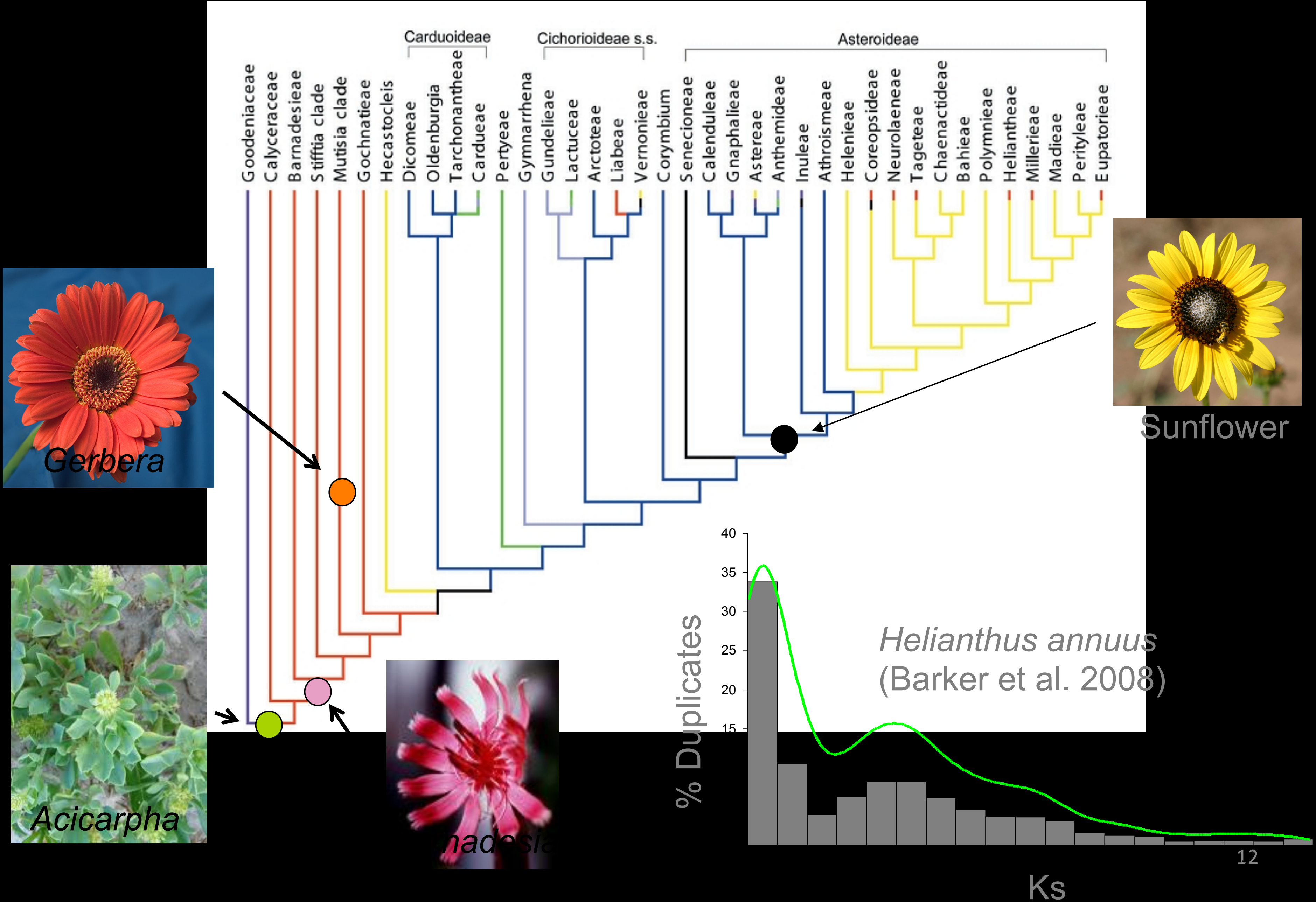
S. Evan Staton[1], Bradley H. Bakken[2], Benjamin K. Blackman[3,†], Mark A. Chapman[4,‡], Nolan C. Kane[5], Shunxue Tang[6,§], Mark C. Ungerer[2], Steven J. Knapp[6,¶], Loren H. Rieseberg[5] and John M. Burke[4,*]

Young Repeats, Low Divergence

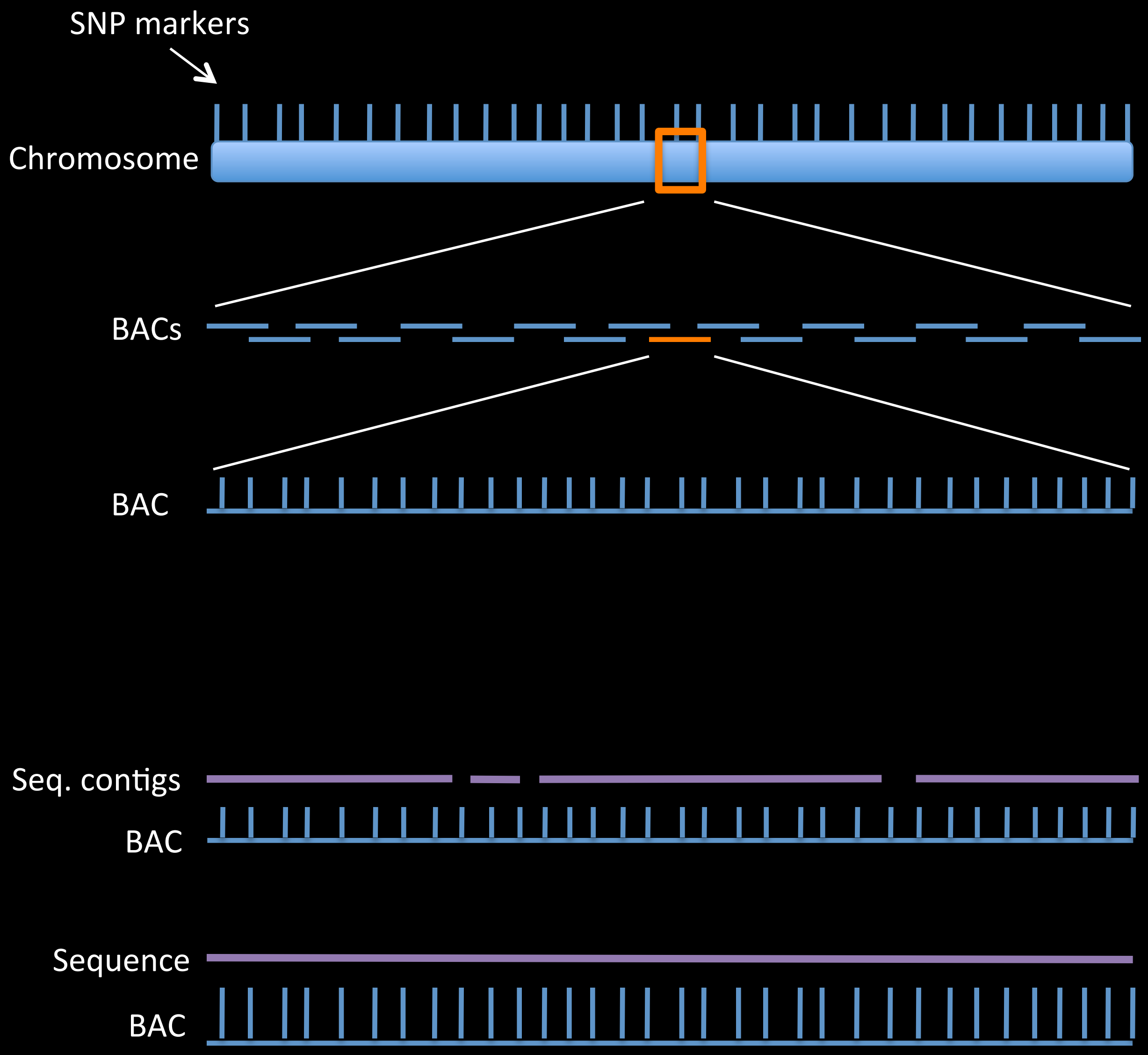


Evan Staton

Sunflower is ancient octaploid



Strategy



High-density genetic map

Sequence-based physical map

Sequence-tags every 5-6 kb

Assembly of WGS sequence

Sequence contigs aligned against genetic & physical map

Local scaffolding with mate pairs

Annotation

43 Read Libraries

	Illumina ~60x			454 ~24x		
	LIBRARY_NAME	MEAN	STDDEV	LIBRARY_NAME	MEAN	STDDEV
Short Fragment	A1	136	23	01V17GRL2	368	133
	A2	139	28	MPS004761454RL	373	124
	A5	160	34	MPS006655454RL	383	120
	200bp_HA0001	192	21	01V17G454RL	384	136
	500bp_HA0002	408	46	MPS004762454RL	394	125
				HA412Long	592	169
				MAYha412long	648	205
Mate Pair	2kbp_HA0003_61YEJAAXX_1	1510	321			
	MP1	2062	1744	01V17G454PE1	2890	485
	MP2.BD0TEHACXX_3	2451	295	01V17G454PE2	2929	515
	MP3.AC0C9VACXX_4	2500	272	MPS008920454PE55kb	3259	1148
	LBM11326_GFI-529_3kb_LJD	2550	760	MPS008921454PE6kb	3517	1412
	MP4.BD0TEHACXX_5	3320	443	MPS006655454PE20Kb	4491	3390
	HA412_GGCTAC_40kb_LJD	3458	1910	MPS004761454PE38kb	7272	1060
	MP5.AC0C9VACXX_6	3848	339	MPS008922454PE8kb	7584	1360
	5kbp_HA0004_626E6AAXX_5	4418	846	MPS004761454PE210kb	7897	1146
	MP6.BD0TEHACXX_7	4653	468	MPS008923454PE10kb	10041	1587
	INX517*	4394	321	MPS004761454PE10kb	10441	1932
	LBM11326_GFI-546_40kb_LJD	5084	3642	MPS004761454PE15kb	11157	5060
	INX518*	5286	2016	MPS009917454PE20kb	12507	4667
	LBM11325_GFI-530_8kb_LJD	7114	1090	MPS008924454PE20kb	12955	4944
	LBM_CAGATC_8kb.LJD	7132	1057	MPS009918454PE20kb	13463	5190
	LBM_GATCAG_20kb.LJD	13887	5153			
	LBM1481_GFI-531_20kb_LJD	16863	4578			

The Eulerian path/de Bruijn graph approach

1. kmer hash table
2. de Bruijn graph
3. simplification of the graph and Eulerian path search

Examples: Euler, Velvet, Allpath, Abyss, SOAPdenovo. . .

10bp read: ATTCGACTCC

for k=5,
6 kmers:

ATTCG
TTCGA
TCGAC
CGACT
GACTC
ACTCC

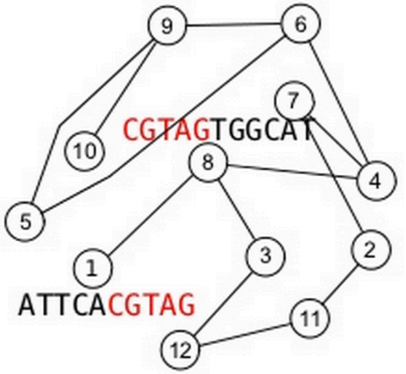
de Bruijn Graph

Tristan Lefébure

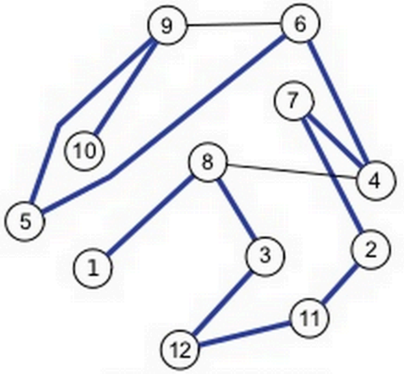
The Overlap-layout-consensus (OLC) approach

1. Pairwise alignments and overlap graph
2. Graph Layout: search of a single path in the graph (i.e. the Hamiltonian path)
3. Multiple sequence alignments and consensus

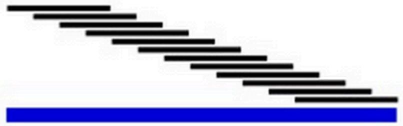
Examples: Newbler, Celera, Arachne. . .



Overlap graph



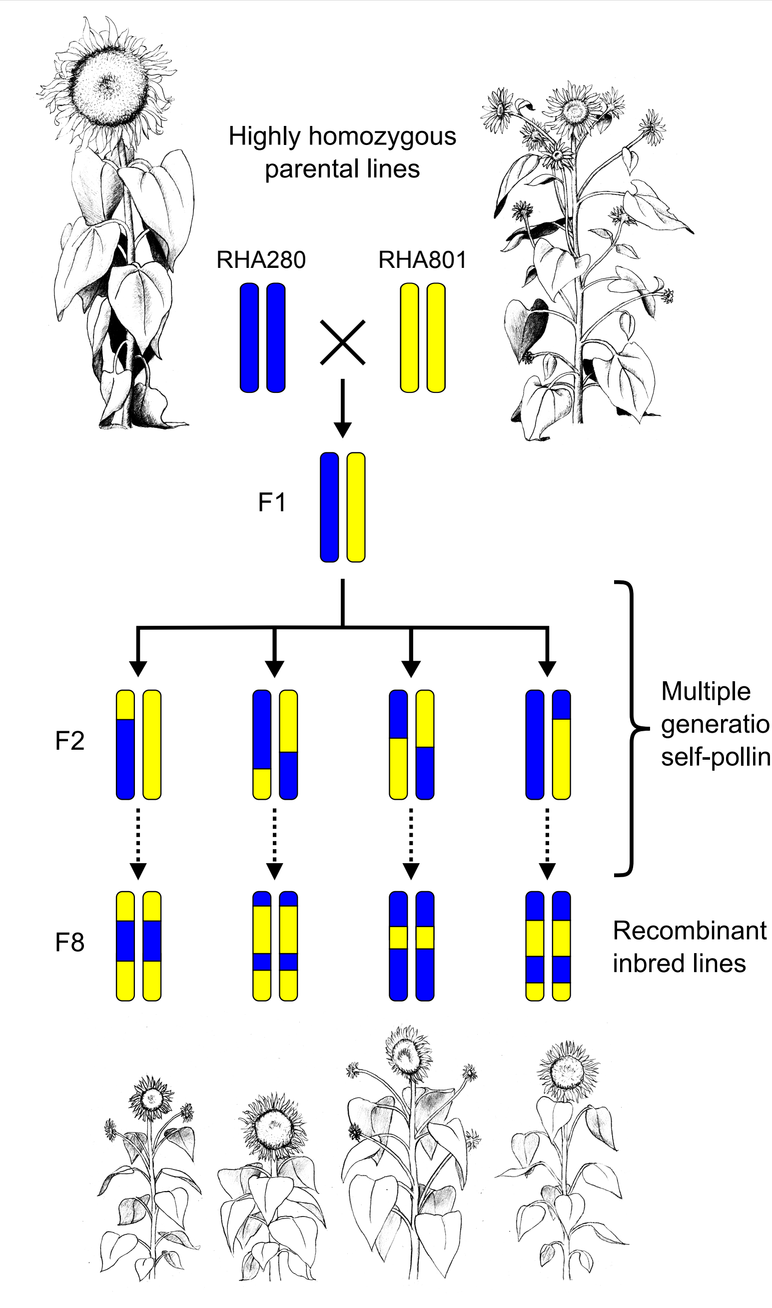
Graph layout

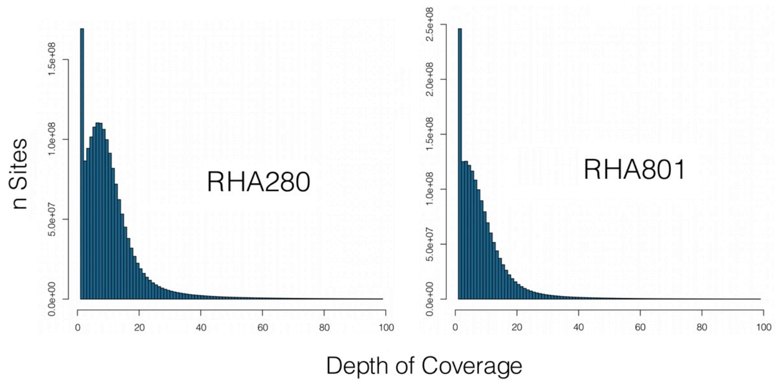


Consensus

Tristan Lefébure

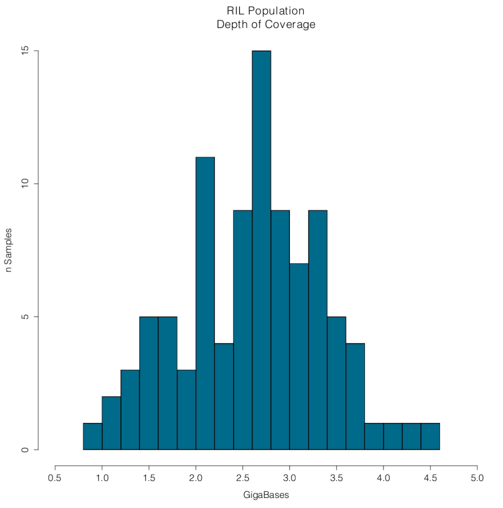
Sequence Based Genetic Map



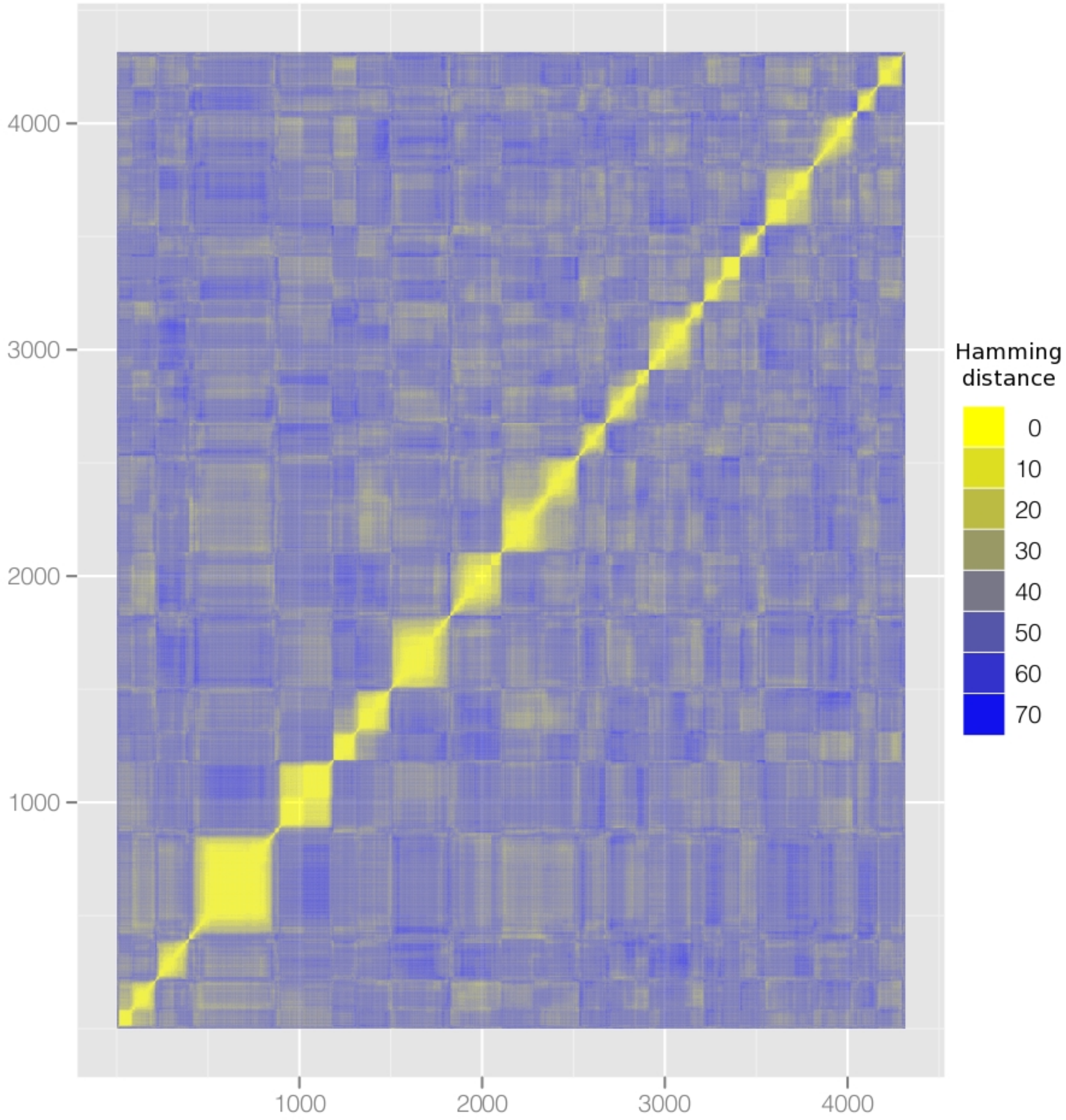


10x WGS Parental Lines

~.8x WGS of 93 RILs



Template Map



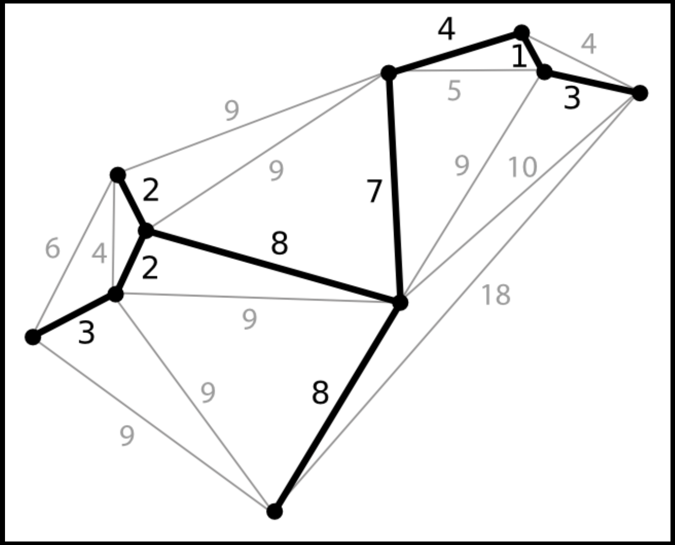

OPEN ACCESS Freely available online

PLoS GENETICS

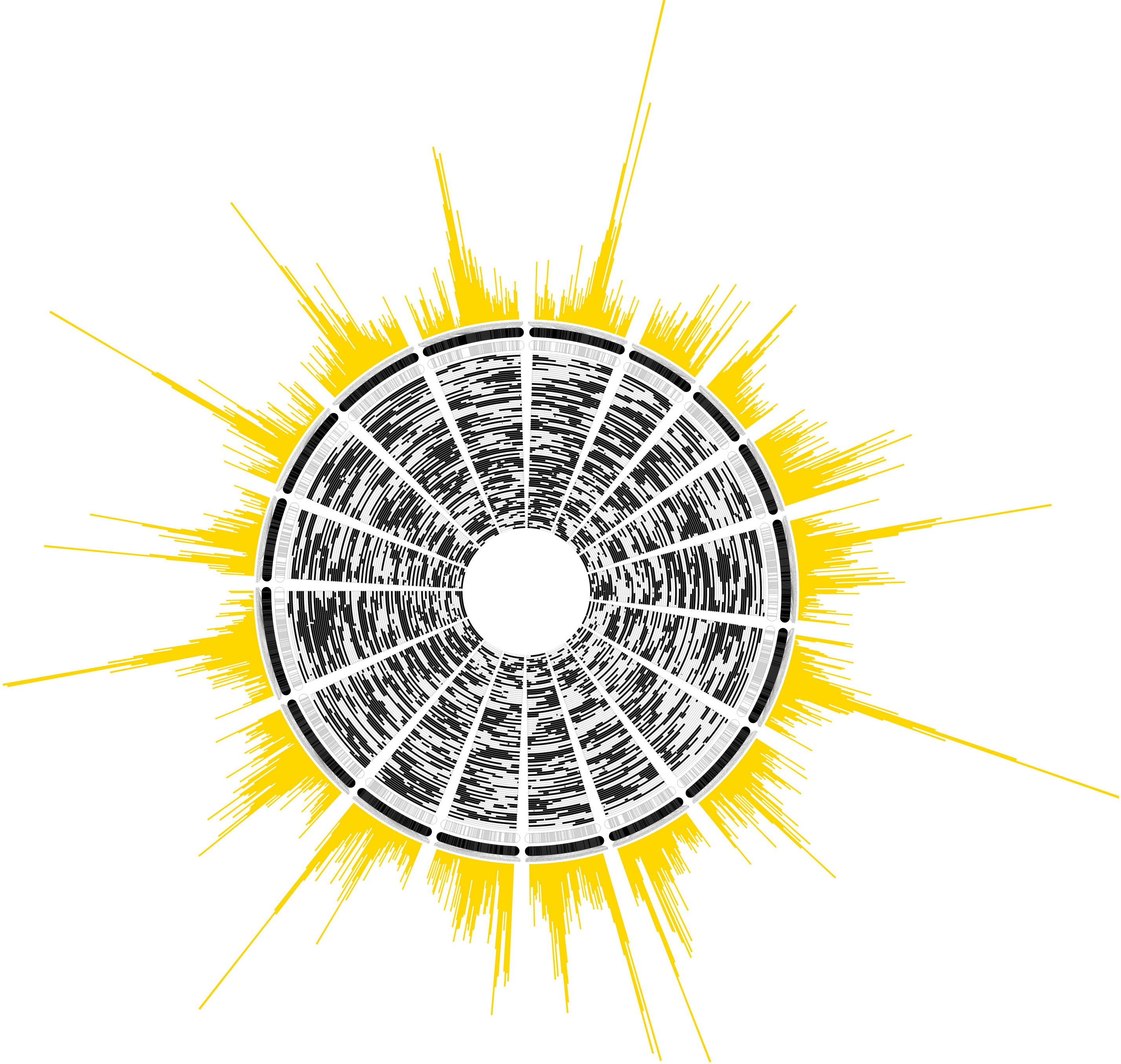
Efficient and Accurate Construction of Genetic Linkage Maps from the Minimum Spanning Tree of a Graph

Yonghui Wu[1], Prasanna R. Bhat[2], Timothy J. Close[2], Stefano Lonardi[1]*

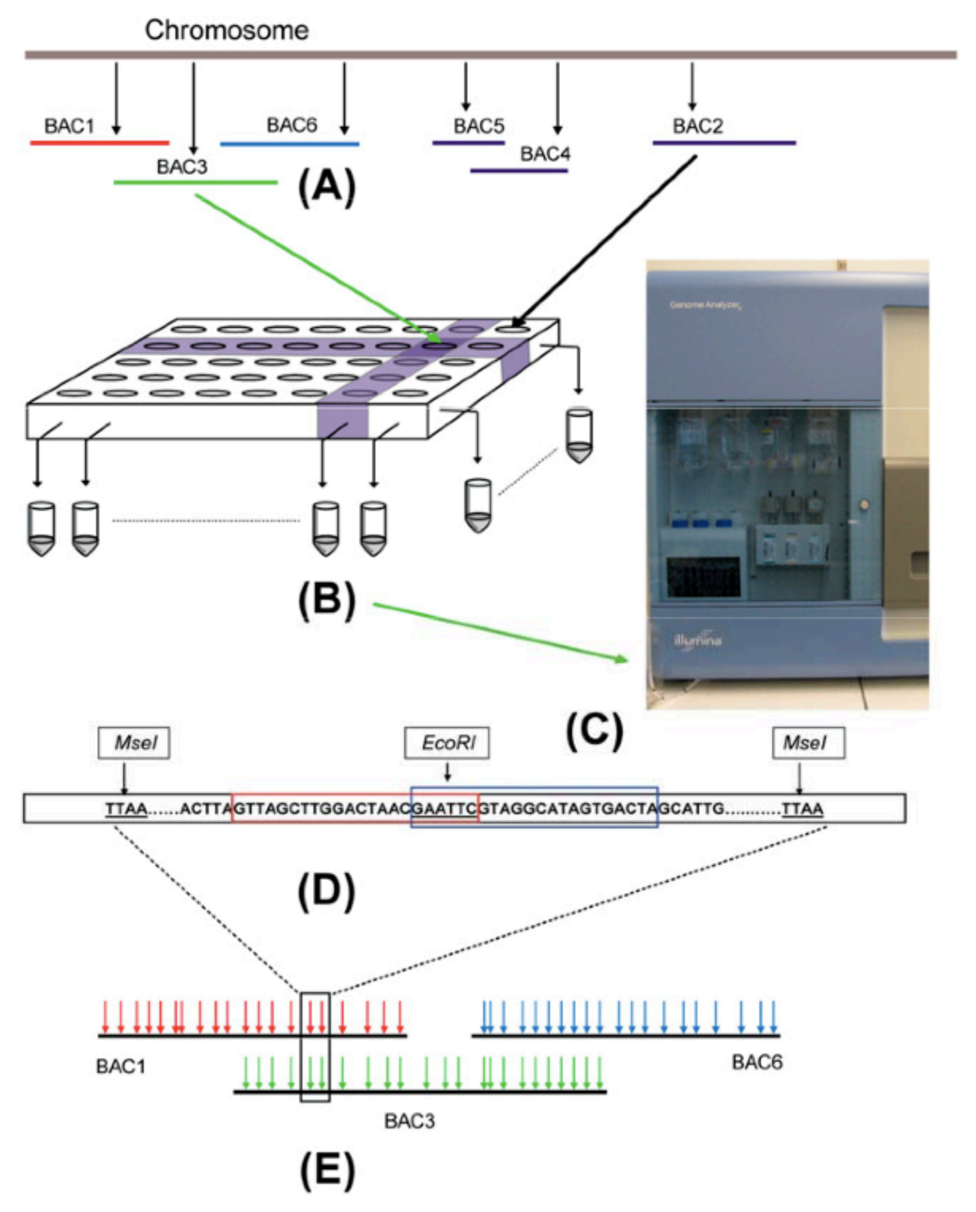
1 Department of Computer Science and Engineering, University of California Riverside, Riverside, California, United States of America, 2 Department of Botany and Plant Sciences, University of California Riverside, Riverside, California, United States of America




47.878
47.340
46.803
46.265
45.727
45.190
43.576
43.039
42.501
41.963
40.888
40.350
39.813
39.275
38.737
36.586
36.048
35.510
34.973
34.435
33.897
32.284
31.746
31.209
30.671
30.133
29.596
29.058
28.520
27.983
27.445
26.907
26.370
25.294
24.757
24.219
23.681
23.144
22.606
22.068
21.531
16.133
15.596
15.058
13.445
12.907
10.755
10.217
9.680
9.142
8.604
8.067
6.453
5.916
4.840
4.303
3.765
2.690
2.152
0.000

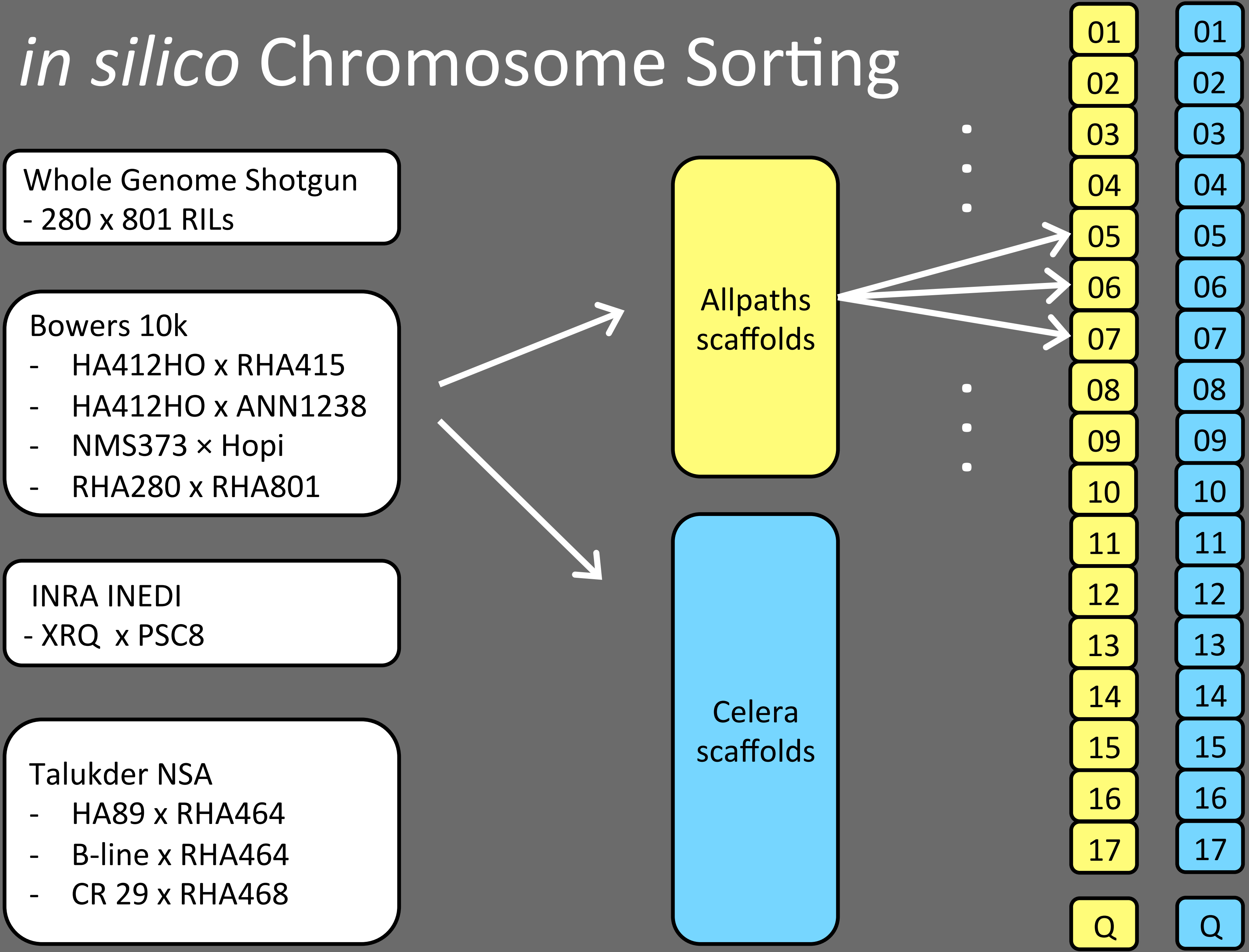


Whole Genome Profiling Physical Map



A) 12x BAC library
B) Barcoded BAC Pools
C) Restriction Digest and Sequencing
D) Deconvolution of BAC tags
E) BAC-contig assembly

in silico Chromosome Sorting



Whole Genome Shotgun
- 280 x 801 RILs

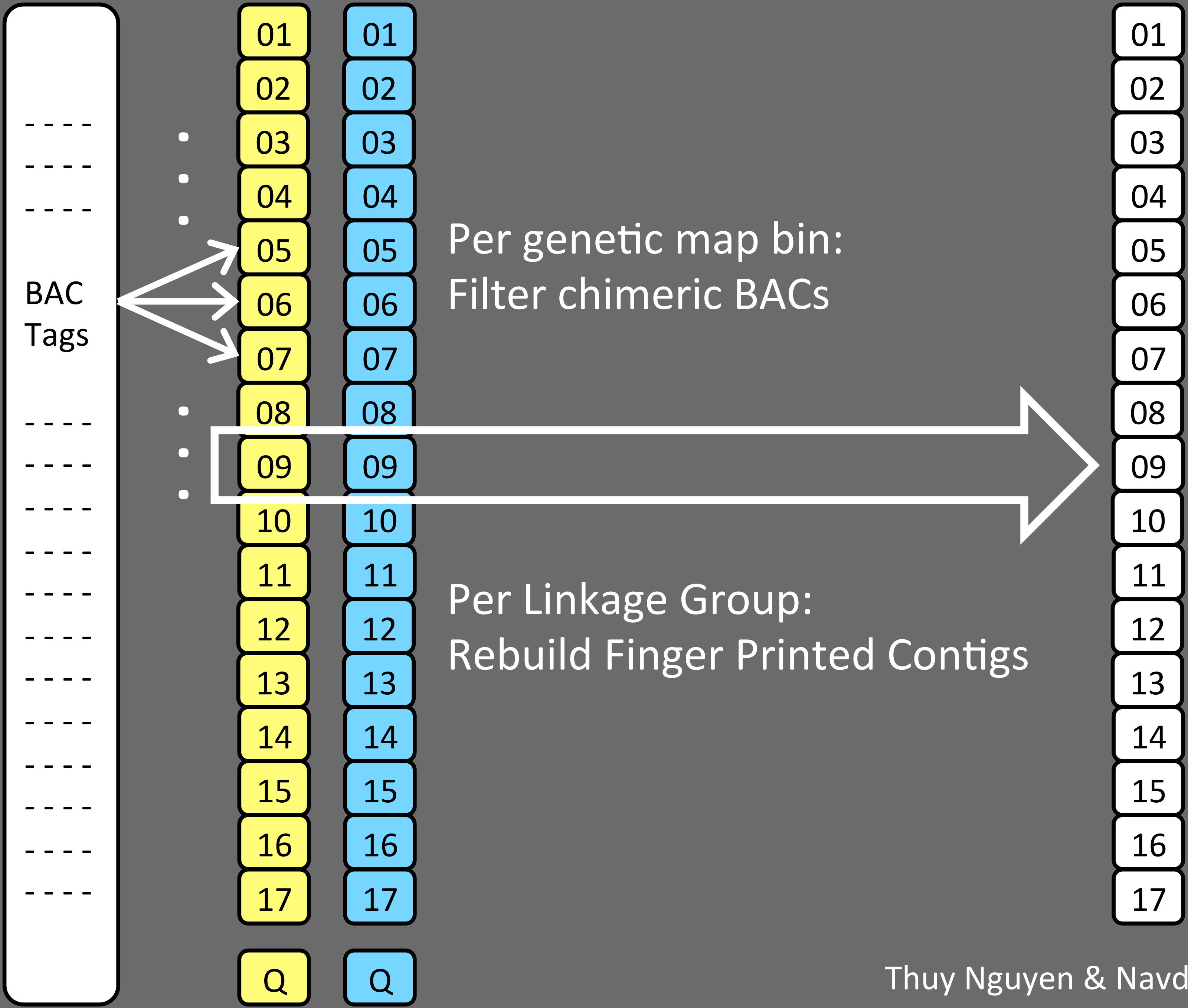
Bowers 10k
- HA412HO x RHA415
- HA412HO x ANN1238
- NMS373 × Hopi
- RHA280 x RHA801

INRA INEDI
- XRQ x PSC8

Talukder NSA
- HA89 x RHA464
- B-line x RHA464
- CR 29 x RHA468

Allpaths scaffolds

Celera scaffolds

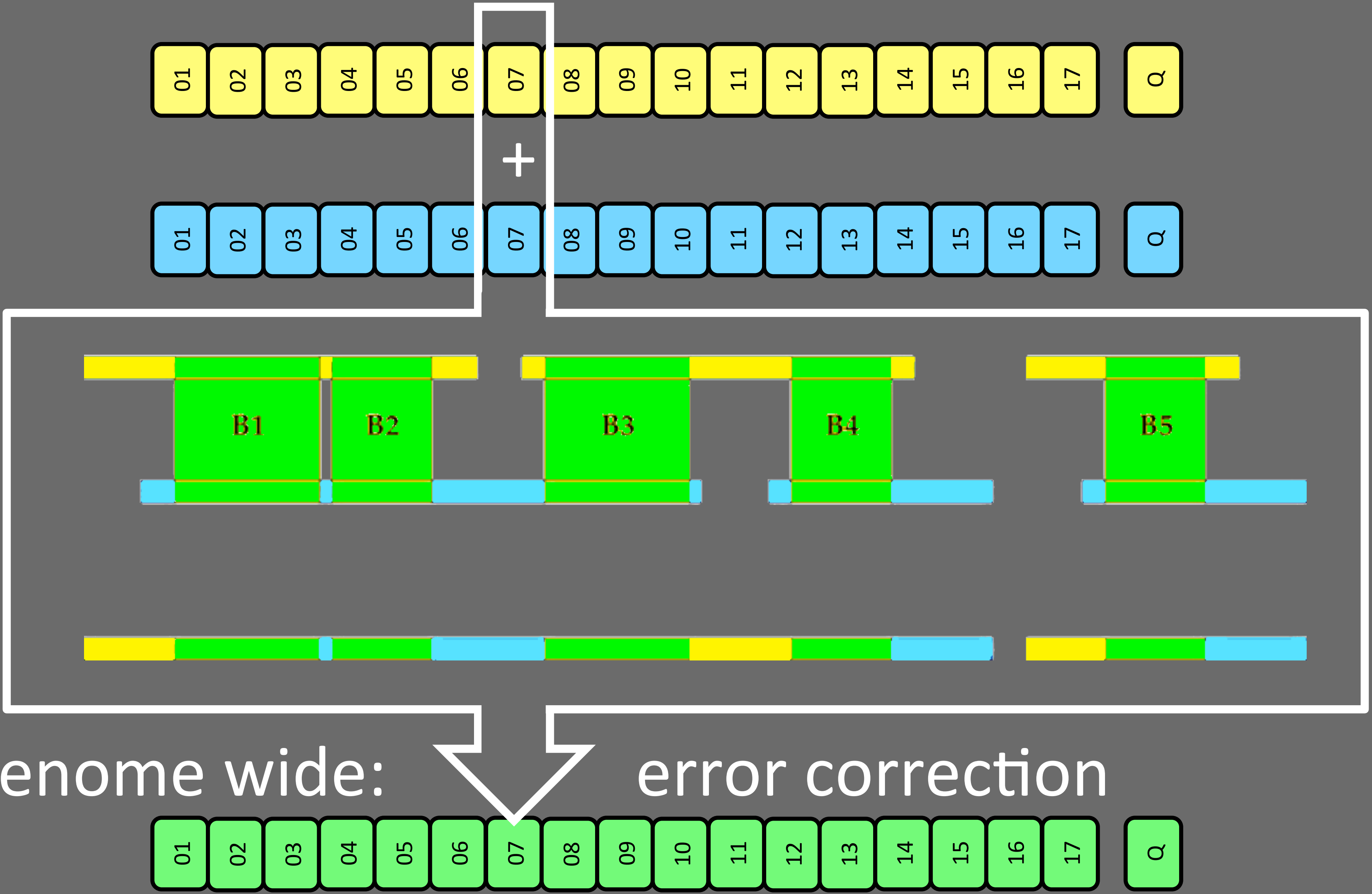


Per genetic map bin:
Filter chimeric BACs

Per Linkage Group:
Rebuild Finger Printed Contigs

Thuy Nguyen & Navdeep Gill

per LG: Assembly merge & SSPACE scaffolding



Genome wide: error correction

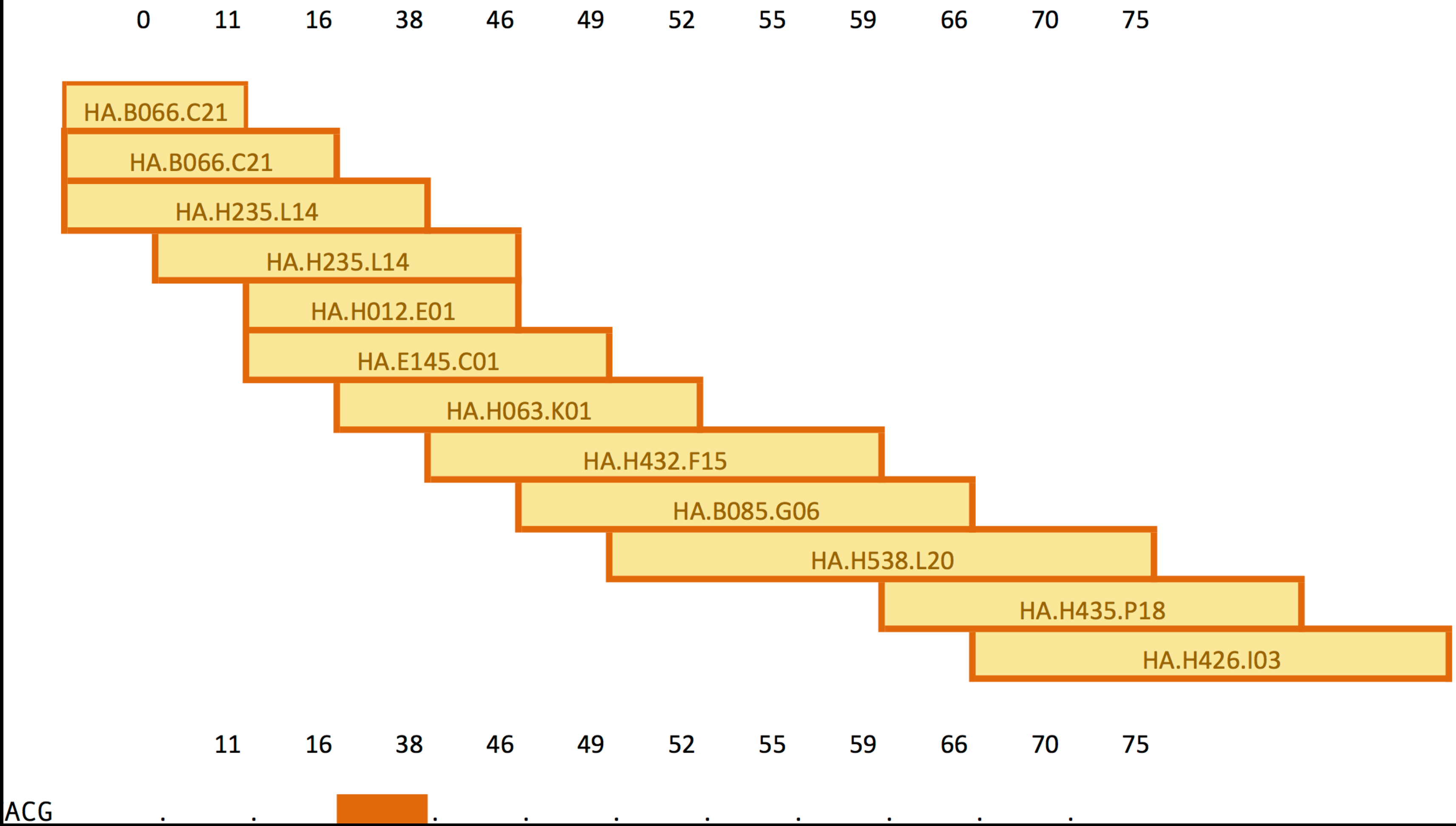
Sariel Hubner

```
#### vvv LG08|scaffold403|size86505 TO ctg66    vvv ####
#            LG:08 cM: 66.93
#
#pMap_TAG               scaffold_bp pMap_CTG_fpc
#                                    100 102 115
GAATTCCGAACACACTGATGTGATTA      17724   ..X
GAATTCGTTGTAAAACAGAGATATGATTTC  17746   ..X
GAATTCTAGAATATCCTTGAATACAACCAT  18170   ..X
GAATTCAAGGAAACACGAAATGAGTGGTTT  18974   ..X
GAATTCATTTTCATCAACATGCATCATCTT  20734   ..X
GAATTCAAGGTTGATTTTGAAGAAGAACTG  20758   ..X
GAATTCGAGCTAGCTCGGCTTGGCTCGATC  25585   .X.
GAATTCTAATCAAGCCGAGCTCGAGCCTCA  25609   X..
#### ^^^ LG08|scaffold403|size86505 forward: 2 ^^^ ####
GAATTCTAATCAAGCCGAGCTCGAGCCTCA  25609   X..
GAATTCGAGCTAGCTCGGCTTGGCTCGATC  25585   .X.
GAATTCAAGGTTGATTTTGAAGAAGAACTG  20758   ..X
GAATTCATTTTCATCAACATGCATCATCTT  20734   ..X
GAATTCAAGGAAACACGAAATGAGTGGTTT  18974   ..X
GAATTCTAGAATATCCTTGAATACAACCAT  18170   ..X
GAATTCGTTGTAAAACAGAGATATGATTTC  17746   ..X
GAATTCCGAACACACTGATGTGATTA      17724   ..X
#### ^^^ LG08|scaffold403|size86505 reverse: 10 ^^^ ####
```

```
## alignment_scoring_scheme:
## no change in index = +1
## move forward one   = +2
## move forward >1    =  0
## move reverse       = -1
```

length	n	N50	N90
3131940255	14940	476491	89759

LG04-CTG80 ~632,000 bp



0 11 16 38 46 49 52 55 59 66 70 75

HA.B066.C21

HA.B066.C21

HA.H235.L14

HA.H235.L14

HA.H012.E01

HA.E145.C01

HA.H063.K01

HA.H432.F15

HA.B085.G06

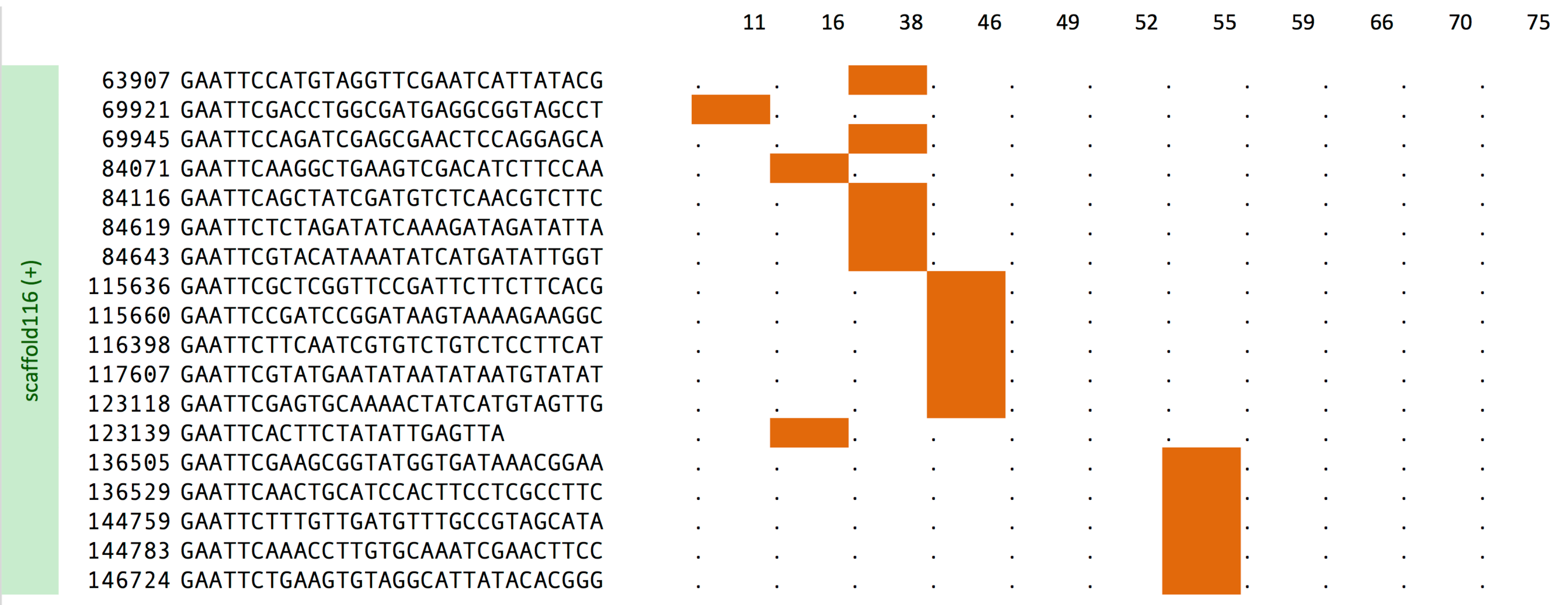
HA.H538.L20

HA.H435.P18

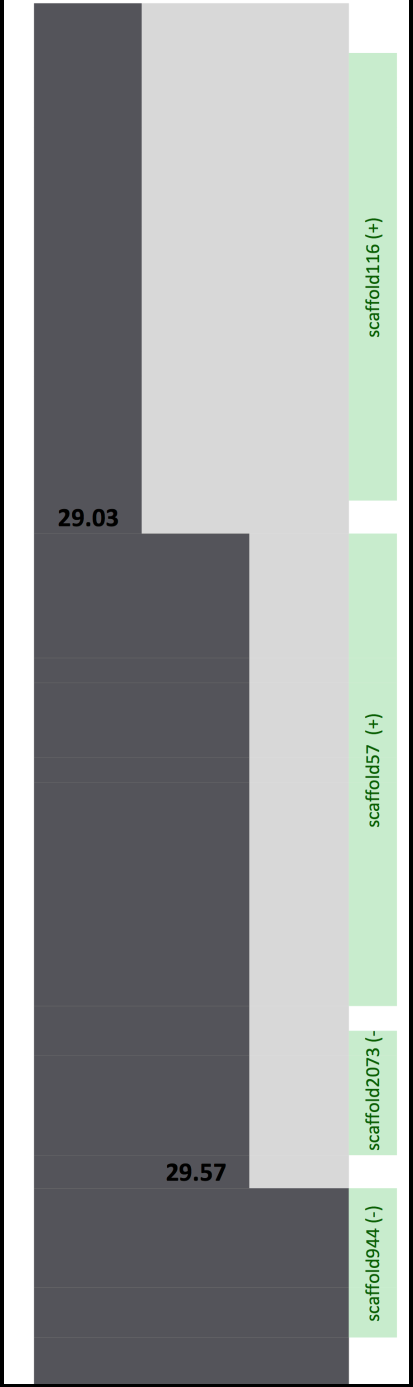
HA.H426.I03

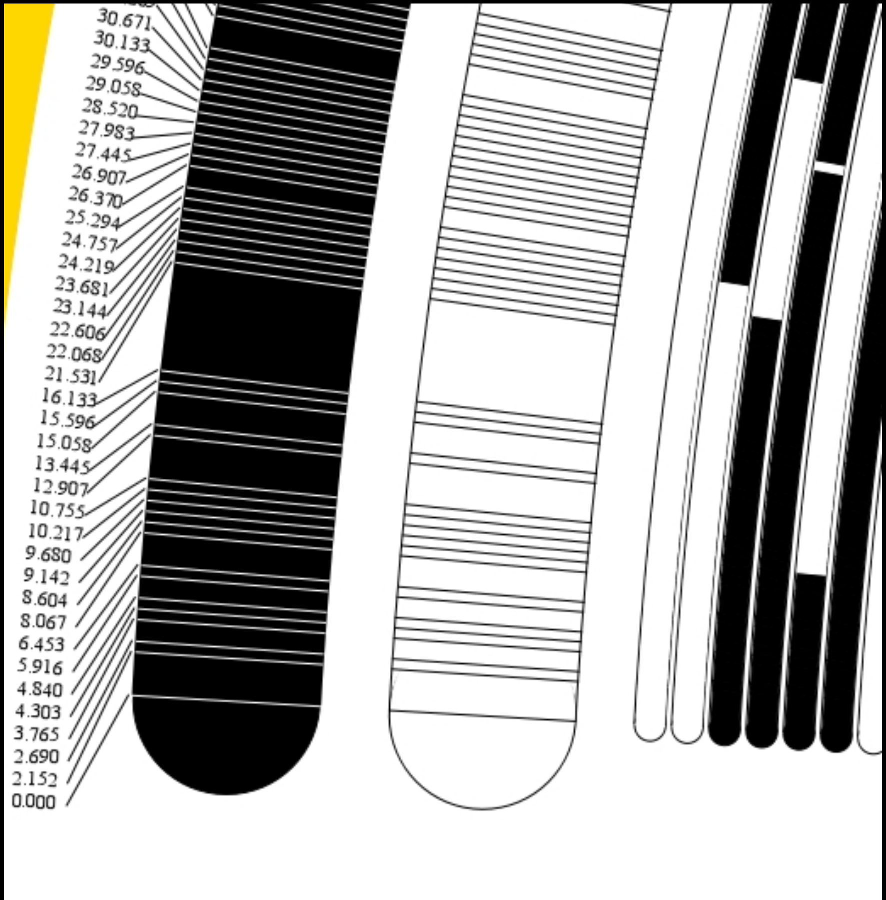
11 16 38 46 49 52 55 59 66 70 75

ACG



			11	16	38	46	49	52	55	59	66	70	75
scaffold116 (+)	63907	GAATTCCATGTAGGTTCGAATCATTATACG	.	.	🟧	.	.	.	.	.	.	.	.
	69921	GAATTCGACCTGGCGATGAGGCGGTAGCCT	🟧	.	.	.	.	.	.	.	.	.	.
	69945	GAATTCCAGATCGAGCGAACTCCAGGAGCA	.	.	🟧	.	.	.	.	.	.	.	.
	84071	GAATTCAAGGCTGAAGTCGACATCTTCCAA	.	🟧	.	.	.	.	.	.	.	.	.
	84116	GAATTCAGCTATCGATGTCTCAACGTCTTC	.	.	🟧	.	.	.	.	.	.	.	.
	84619	GAATTCTCTAGATATCAAAGATAGATATTA	.	.	🟧	.	.	.	.	.	.	.	.
	84643	GAATTCGTACATAAATATCATGATATTGGT	.	.	🟧	.	.	.	.	.	.	.	.
	115636	GAATTCGCTCGGTTCCGATTCTTCTTCACG	.	.	.	🟧	.	.	.	.	.	.	.
	115660	GAATTCCGATCCGGATAAGTAAAAGAAGGC	.	.	.	🟧	.	.	.	.	.	.	.
	116398	GAATTCTTCAATCGTGTCTGTCTCCTTCAT	.	.	.	🟧	.	.	.	.	.	.	.
	117607	GAATTCGTATGAATATAATATAATGTATAT	.	.	.	🟧	.	.	.	.	.	.	.
	123118	GAATTCGAGTGCAAAACTATCATGTAGTTG	.	.	.	🟧	.	.	.	.	.	.	.
	123139	GAATTCACTTCTATATTGAGTTA	.	🟧	.	.	.	.	.	.	.	.	.
	136505	GAATTCGAAGCGGTATGGTGATAAACGGAA	.	.	.	.	.	.	🟧	.	.	.	.
	136529	GAATTCAACTGCATCCACTTCCTCGCCTTC	.	.	.	.	.	.	🟧	.	.	.	.
	144759	GAATTCTTTGTTGATGTTTGCCGTAGCATA	.	.	.	.	.	.	🟧	.	.	.	.
	144783	GAATTCAAACCTTGTGCAAATCGAACTTCC	.	.	.	.	.	.	🟧	.	.	.	.
	146724	GAATTCTGAAGTGTAGGCATTATACACGGG	.	.	.	.	.	.	🟧	.	.	.	.


29.03
29.57
scaffold116 (+)
scaffold57 (+)
scaffold2073 (-
scaffold944 (-)


30.671
30.133
29.596
29.058
28.520
27.983
27.445
26.907
26.370
25.294
24.757
24.219
23.681
23.144
22.606
22.068
21.531
16.133
15.596
15.058
13.445
12.907
10.755
10.217
9.680
9.142
8.604
8.067
6.453
5.916
4.840
4.303
3.765
2.690
2.152
0.000

Stats

assembly	nATCG	length	n	N50	N90
gold-superscaffolds	1658168430	3131940255	14940	476491	89759
bronze-pseudomolecules	1989332786	3641596391	17		
bronze-superscaffolds	2037102379	4168038957	31392	210251	54824
AllPathsLG	993696150	1156891308	100027	20687	6018
Nov22k22	2163238910	3082784204	453539	24876	1541
454AllContigs	1946374353	1946374353	1603248	2009	509

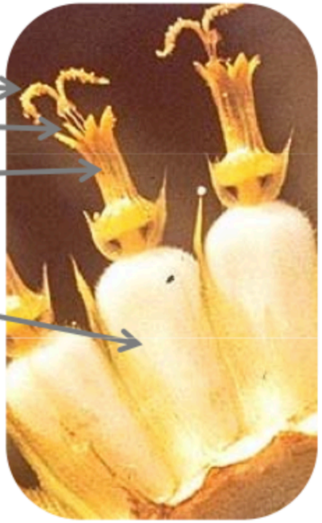
mRNA discovery strategy

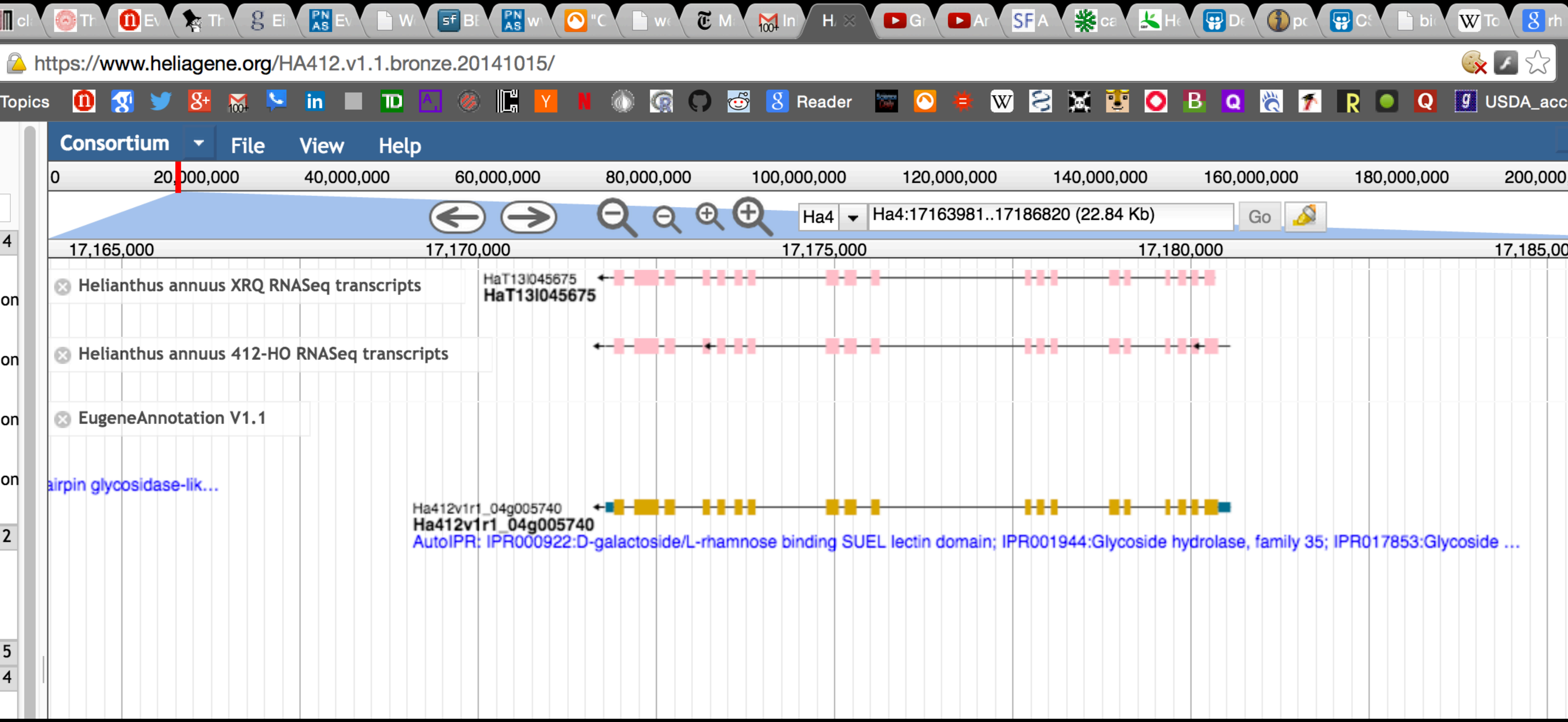
- Genotype XRQ (B)
  - Illumina 100bp

- 3 vegetative tissues

  Leaves, roots and stem
- 8 floral tissues
  - Bracts
  - Ray florets: Ligule (tube and lamina)
  - *Ray florets: Sterile ovary*

  - Disc florets: Style (stigmata + style)
  - Disc florets: Stamens (anthers + filaments)
  - Disc florets: Corolla(Disc floret petals)
  - Disc florets: Seeds (Disc floret seeds)
  - Pollen
- collected at 3-5 different developmental stages





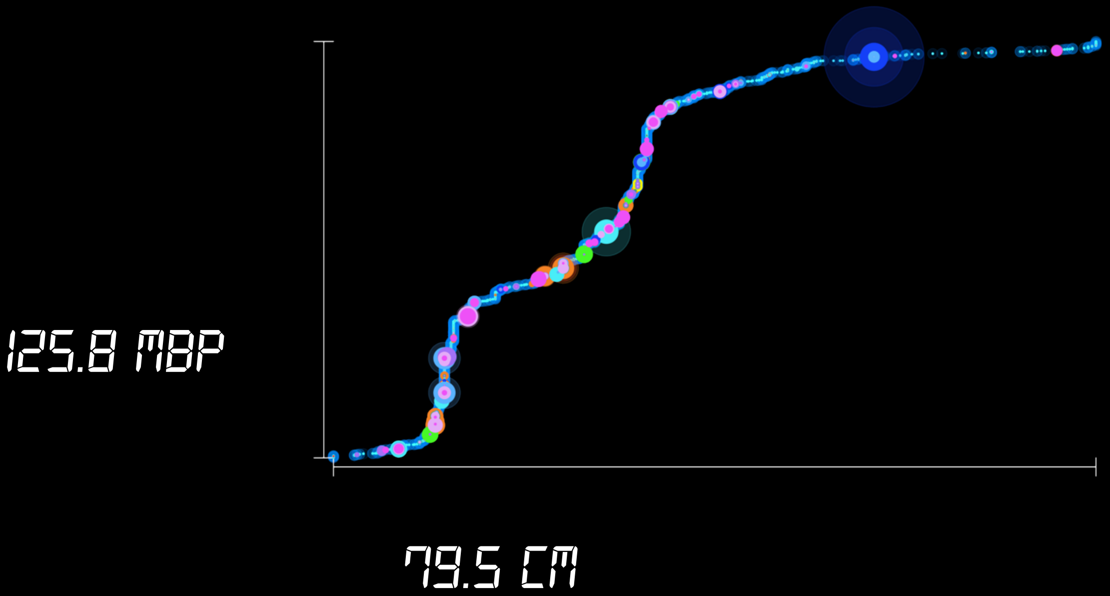
https://www.heliagene.org/HA412.v1.1.bronze.20141015/

Ha4:17163981..17186820 (22.84 Kb)

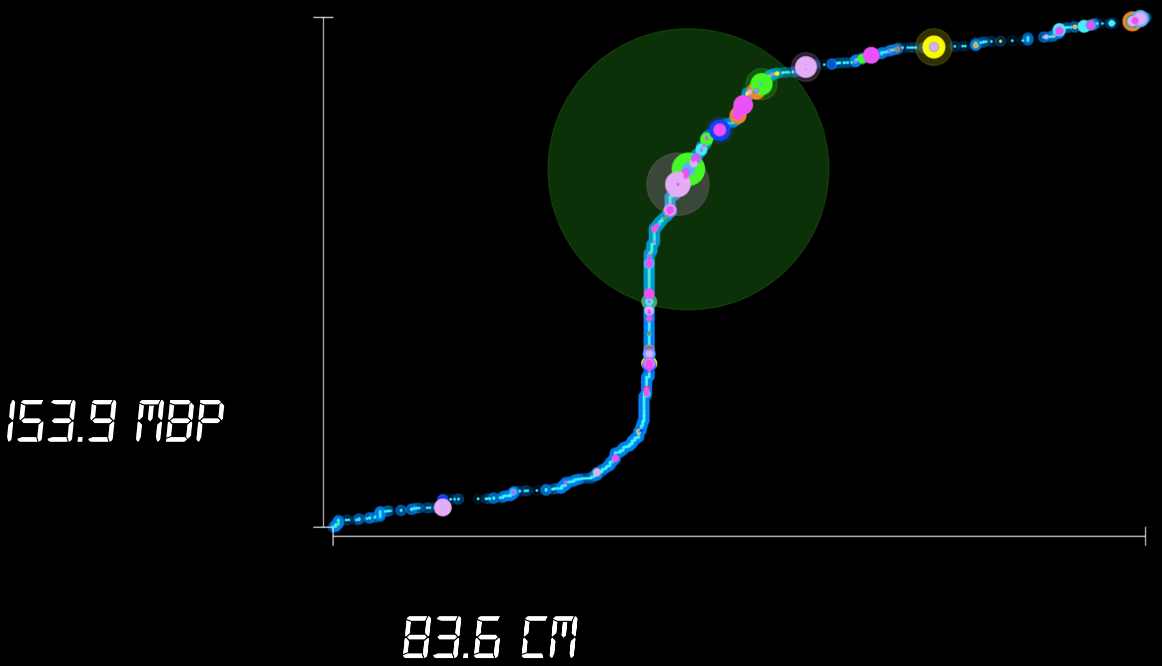
AutoIPR: IPR000922:D-galactoside/L-rhamnose binding SUEL lectin domain; IPR001944:Glycoside hydrolase, family 35; IPR017853:Glycoside …

Jerome Gouzy

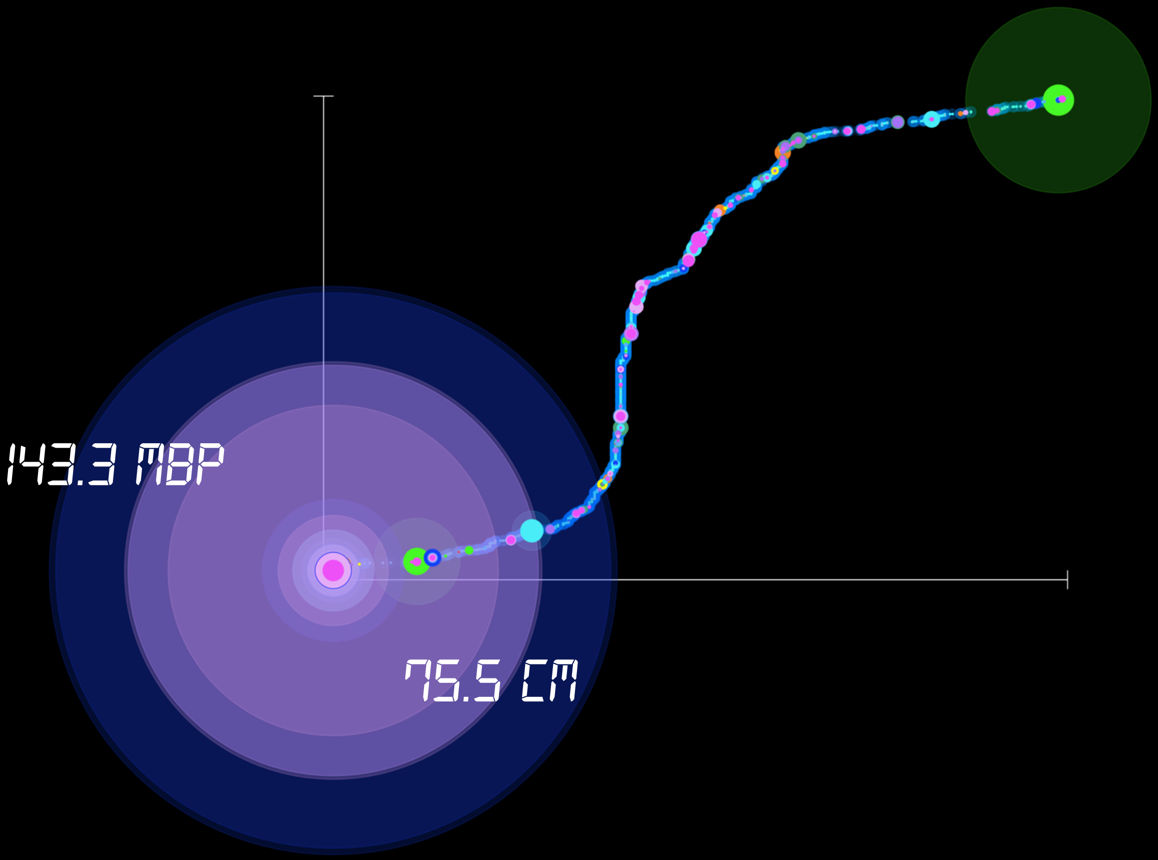
LG01



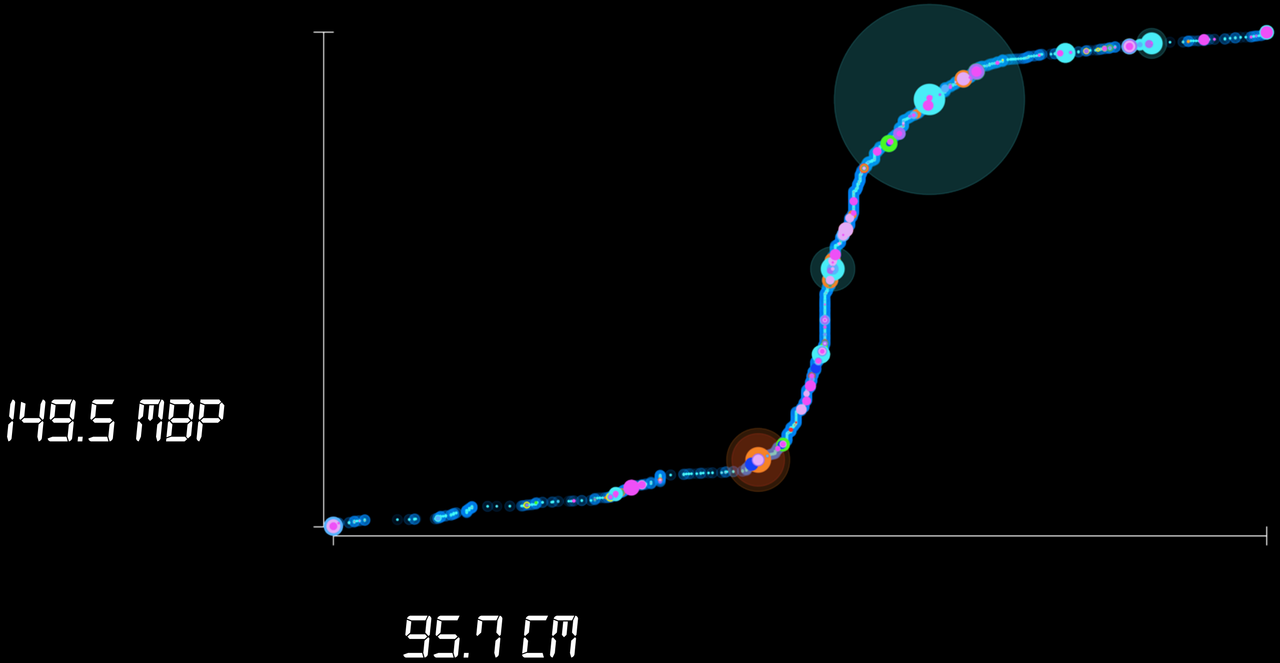
LG02



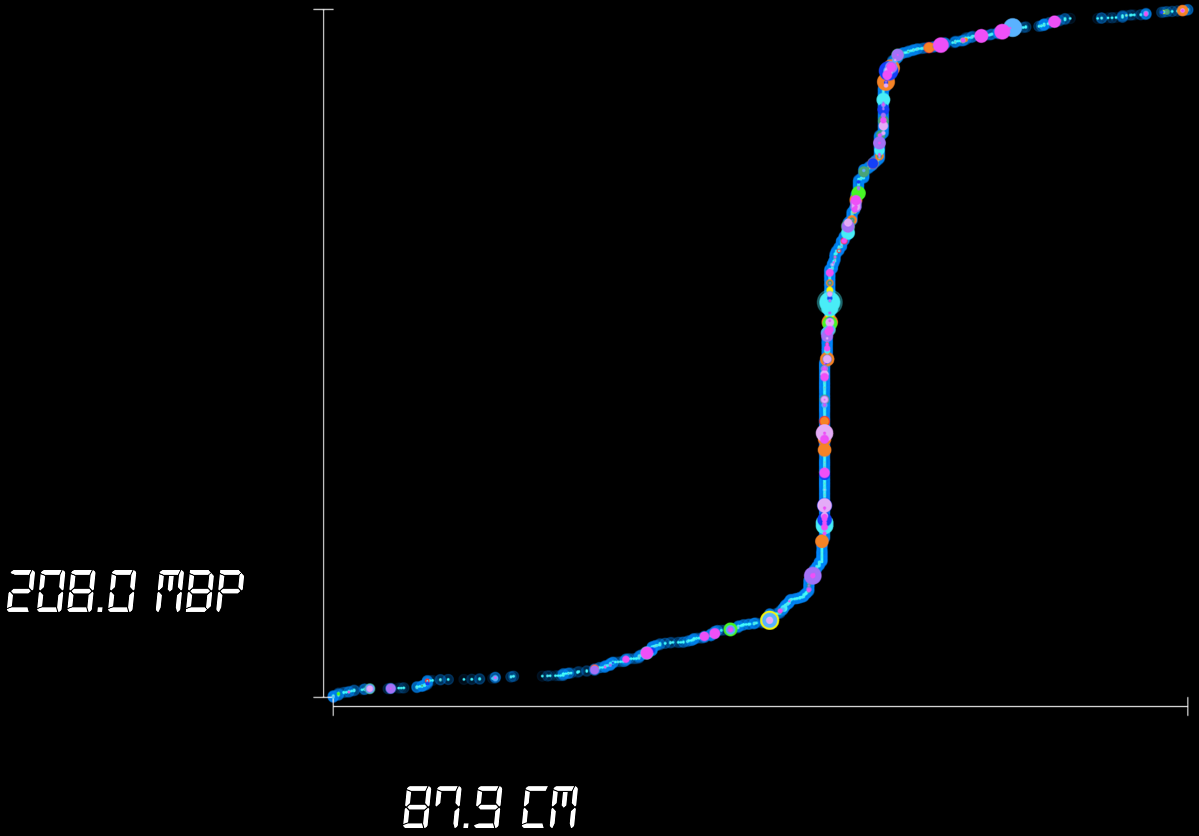
LG03



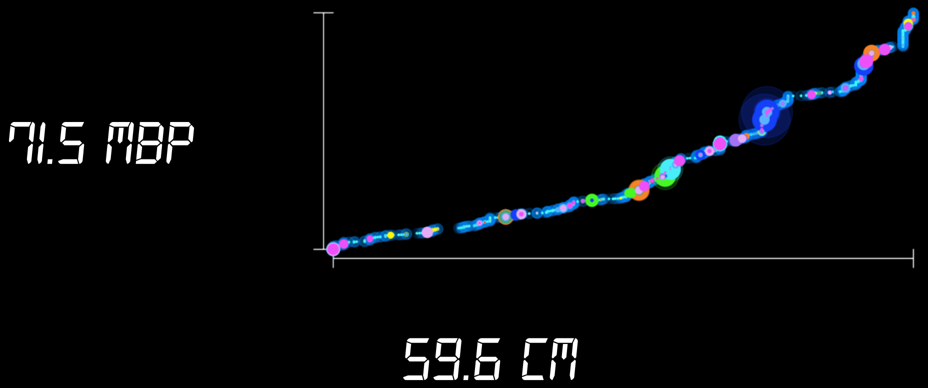
LG04



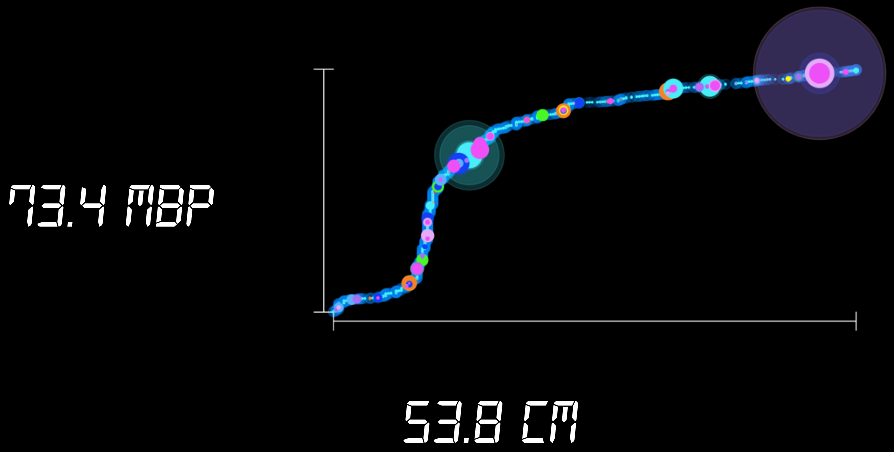
LG05



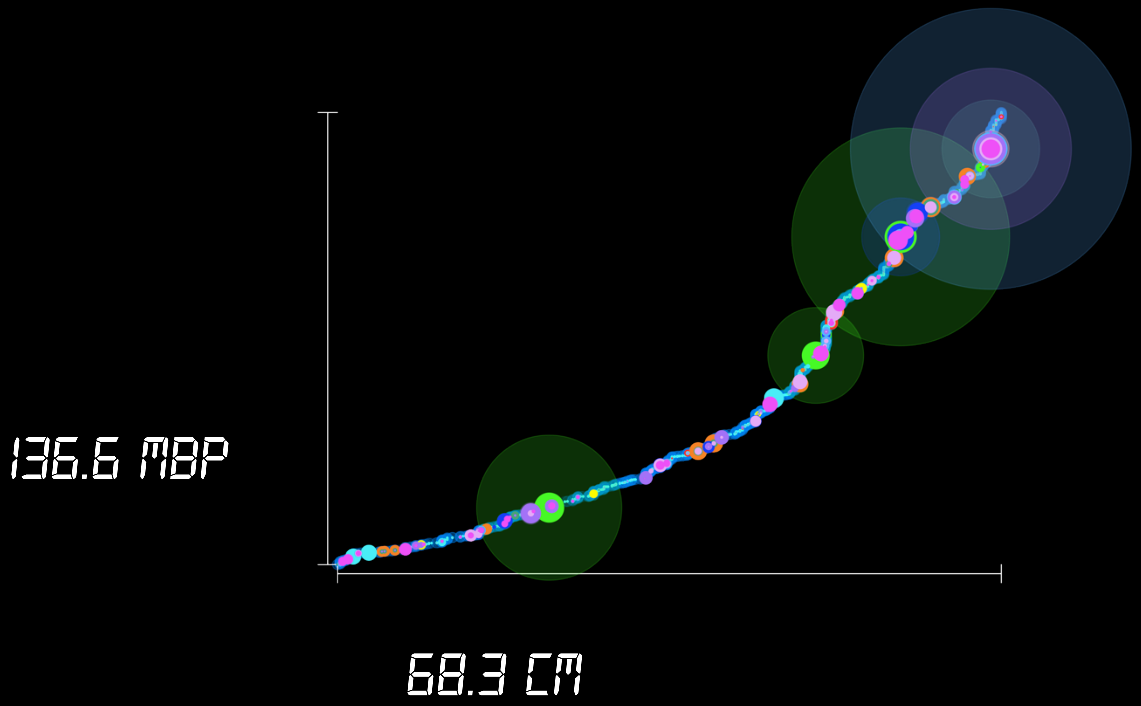
LG06



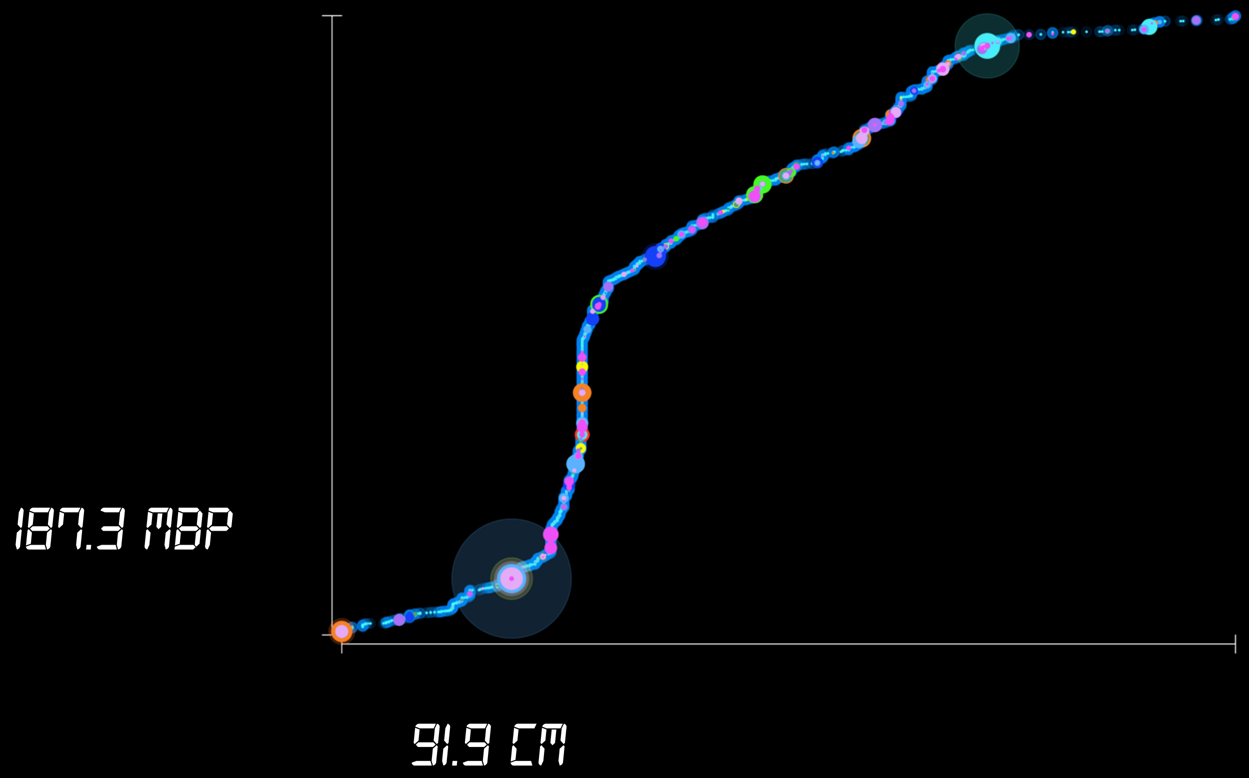
LG07



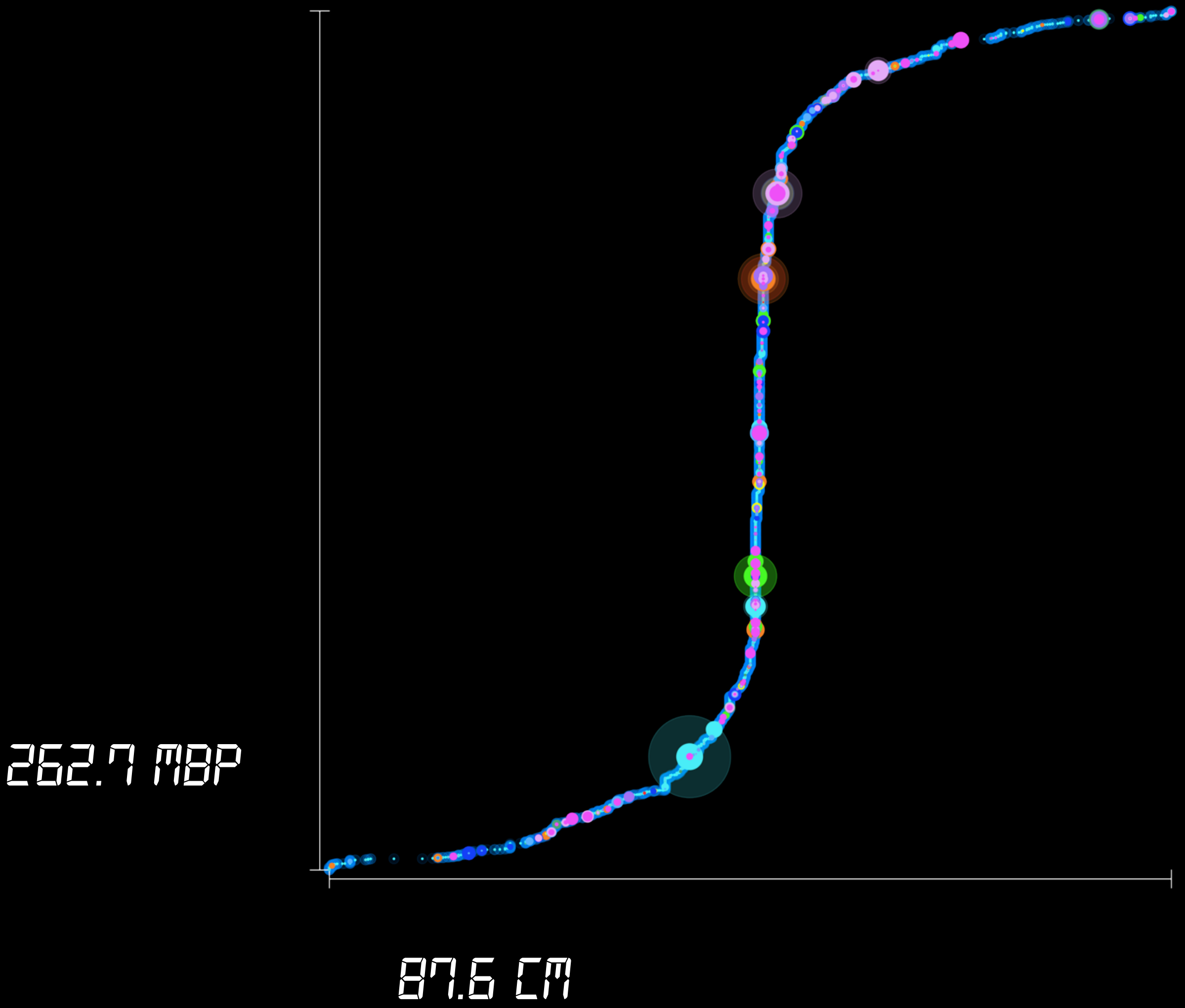
LG08



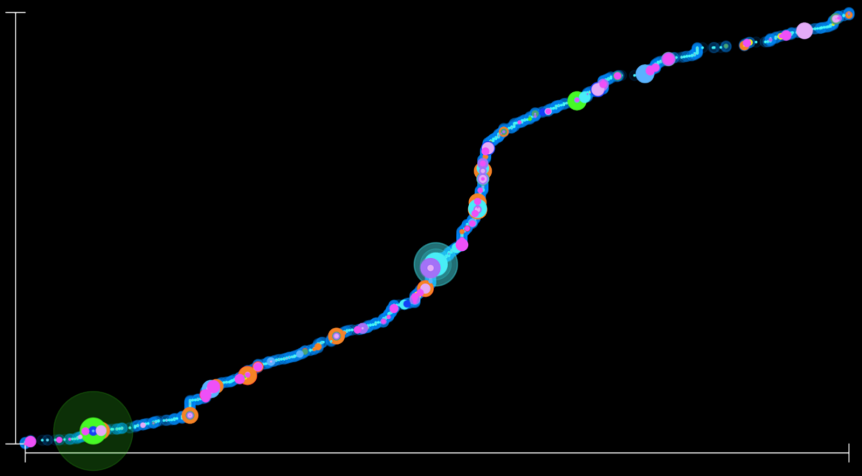
LG09



LG10



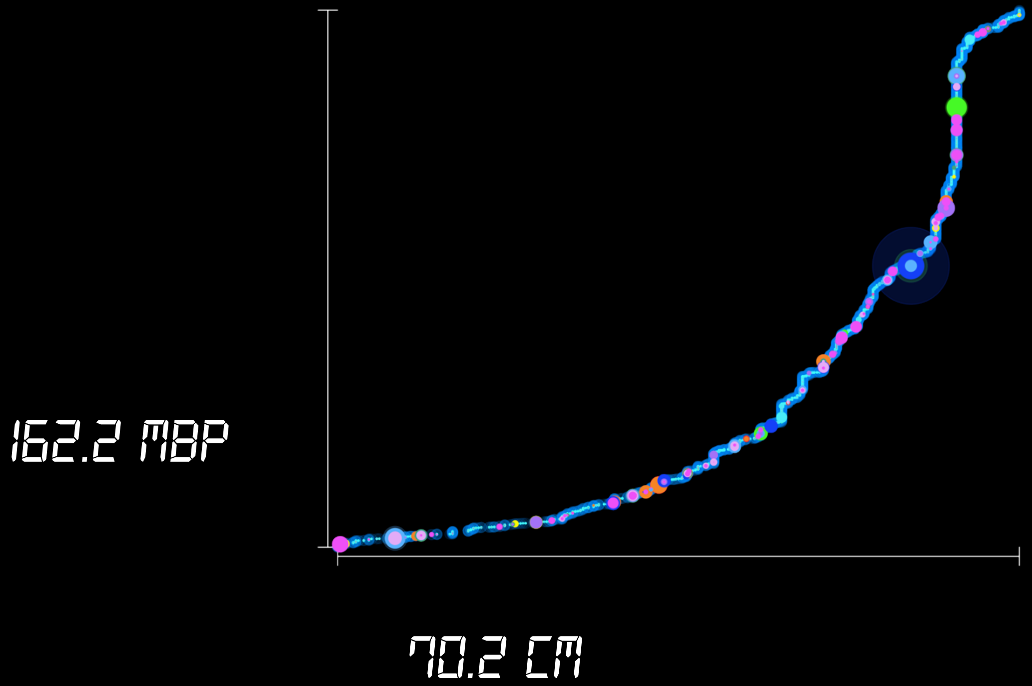
LG11



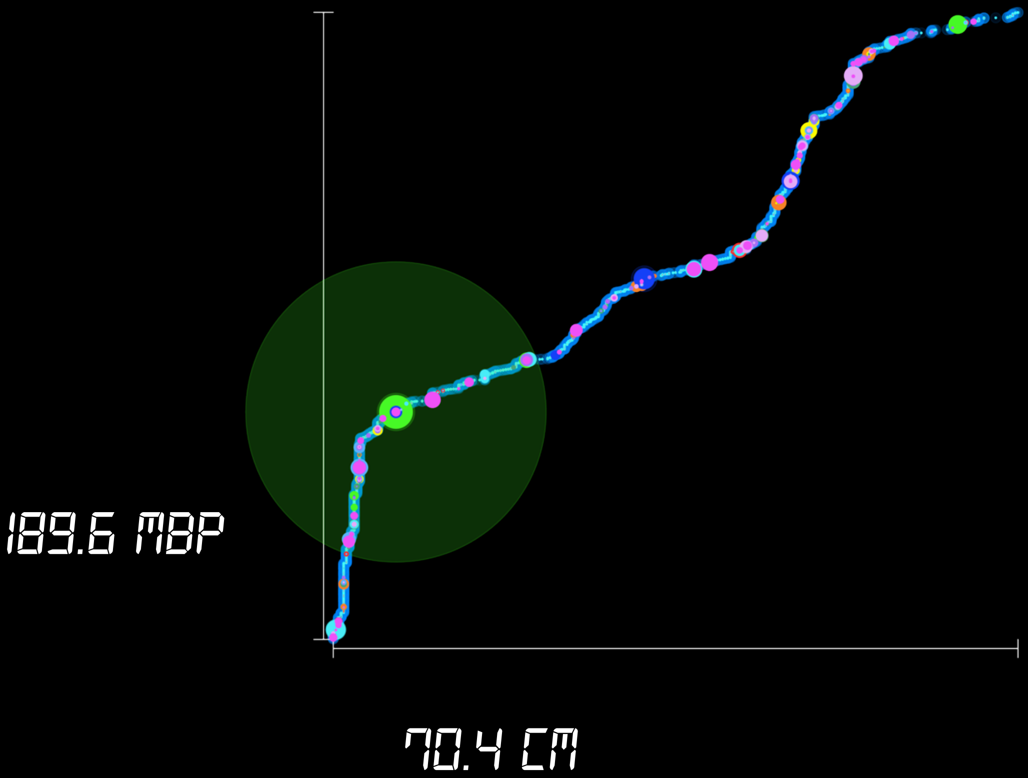
130.5 MBP

84.7 CM

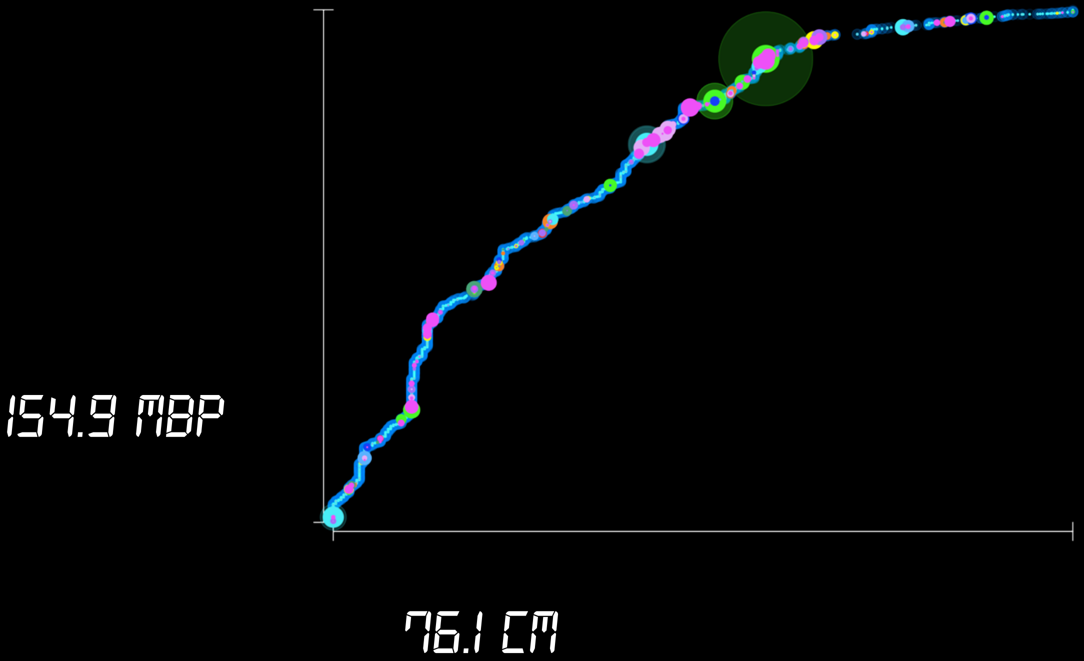
LG12



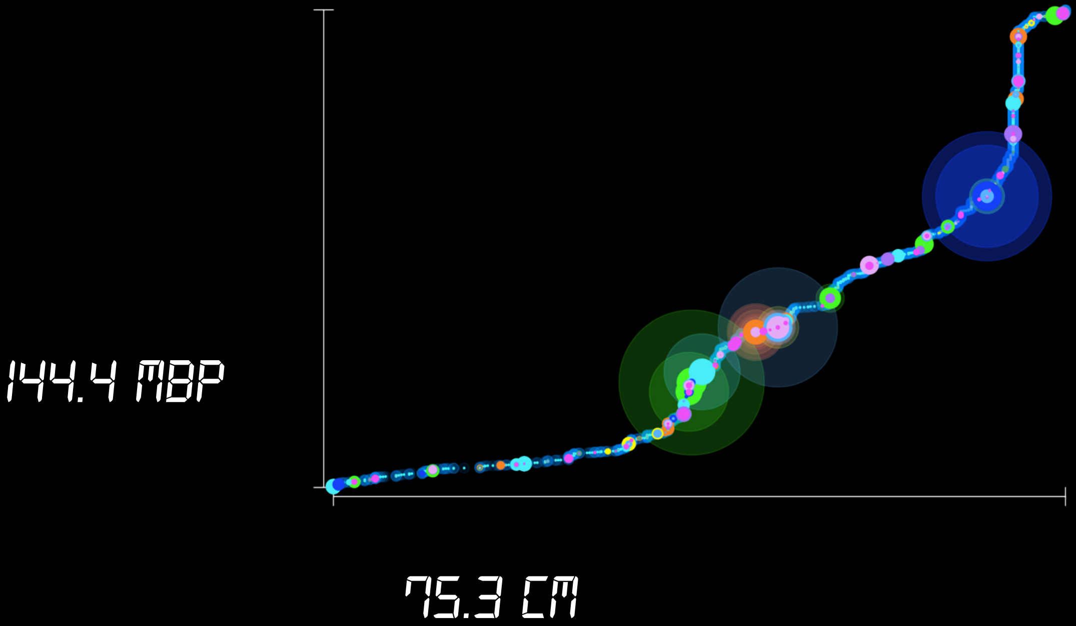
LG13



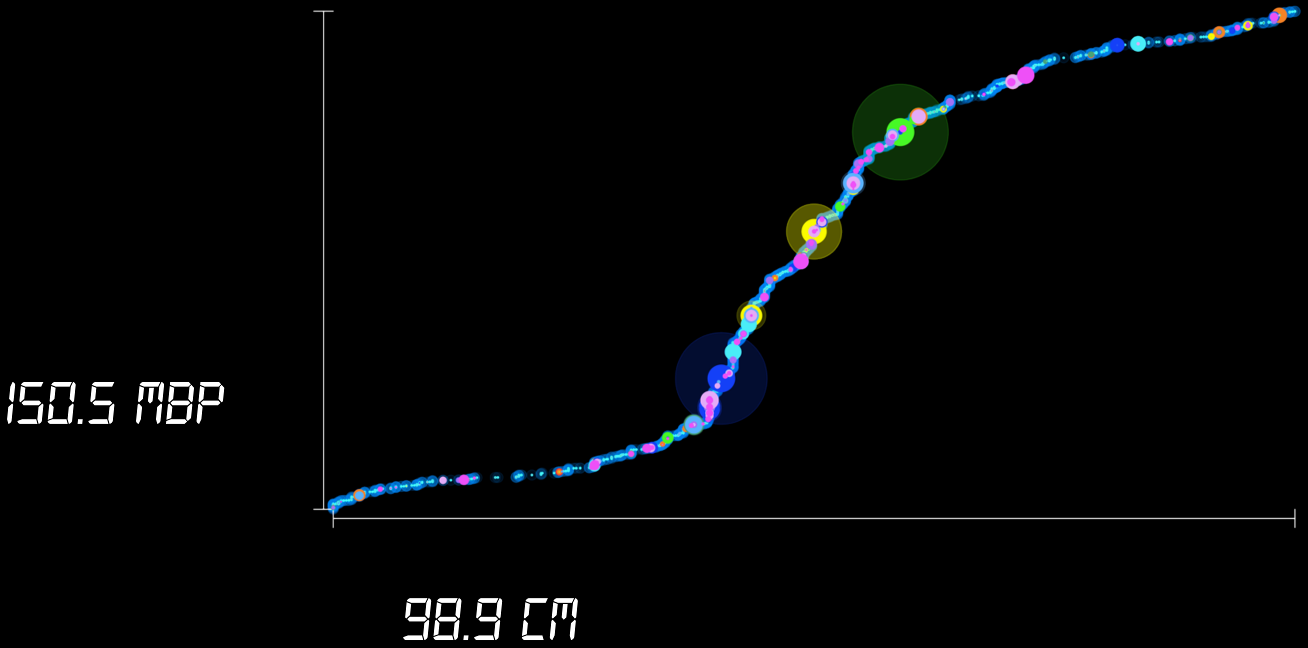
LG14



LG15



LG16



LG17