Second-generation Linkage Maps Reveal Errors in the Assembly of the Pacific Oyster Genome

Dennis Hedgecock, USC*
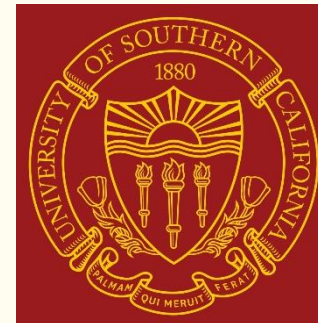
*USC

- Universidad de Santiago de Compostela

- University of South Carolina

- University of the Sunshine Coast (Australia)
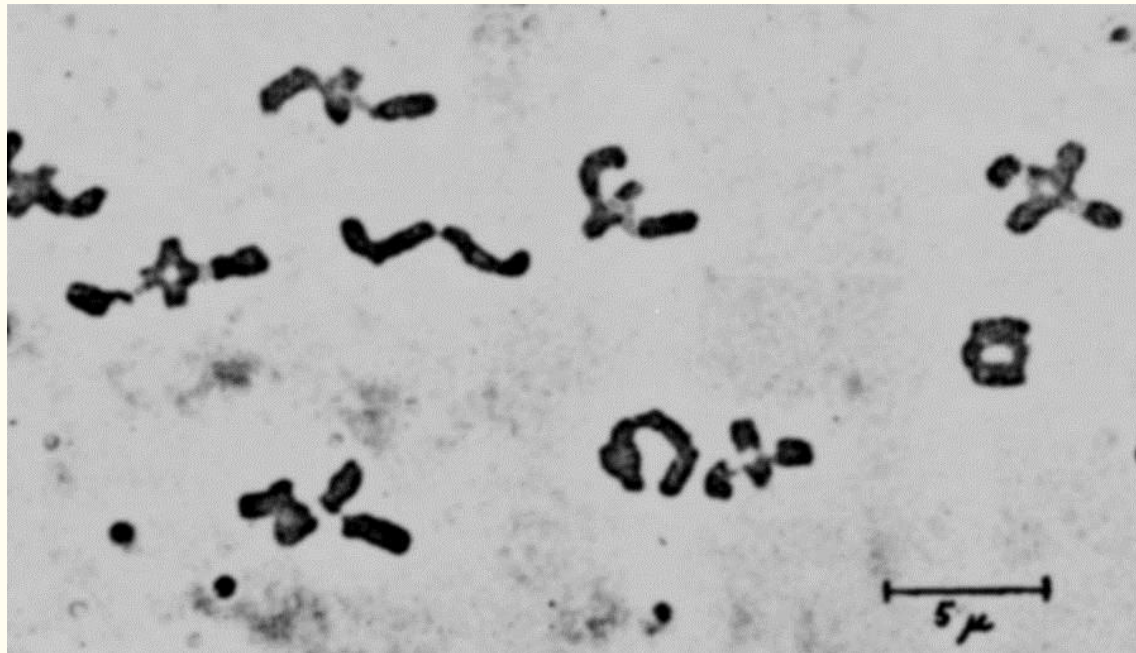
- University of Southern California

Oyster Chromosomes & Genome Size

10 linkage groups, ~600 cM map units, 559 Mbp

(1.2 crossovers per chromosome = 60 cM)

from Longwell, Stiles, & Smith 1967 Can. J. Genet. Cytol. 9:845-856

First- & Second-generation Maps

## First-generation maps

- Used microsatellite DNA or Simple-Sequence Repeat markers

- >100 markers, but only 50-60 mapped in any one family

- Hubert & Hedgecock 2004 *Genetics*; Hubert, Cognard, Hedgecock 2009 *Aquaculture*

Second-generation maps

- Coding SNPs added

- 1536 SNPs typed in 5 families by Illumina Golden Gate assay

- Hedgecock et al. 2011 *JSR;* Hedgecock et al. 2015, submitted
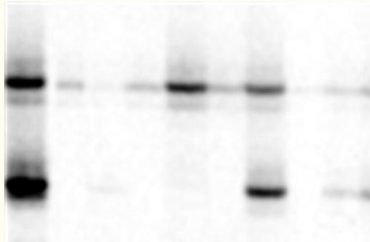
Lengths of 2nd Generation Maps

Markers grouped by pairwise G-tests of independent segregation

LG	No. of SNPs	12	45	20	2x10	51x35	Mean
1	154	43.3	56.7	65.4	96.4	57.4	63.8
2	18	37.1	41.0	28.7	42.4	51.7	40.2
3	119	53.1	48.8	45.4	100.0	51.6	59.8
4	120	57.0	73.4	95.0	71.8	60.8	71.6
5	60	49.3	49.2	55.4	53.1	61.9	53.8
6	152	47.0	68.4	42.8	65.4	58.0	56.3
7	153	54.0	44.8	30.0	94.8	67.0	58.1
8	82	198.4	33.4	78.1	68.8	48.1	85.3
9	29	38.8	39.0	42.1	35.1	57.8	42.6
10	198	50.8	52.9	52.2	62.0	65.7	56.7
Totals	**1085**	**628.8**	**507.6**	**535.0**	**689.8**	**579.9**	**588.2**
cM/marker		**1.06**	**0.79**	**0.88**	**1.35**	**1.11**	**1.04**

Markers by Mating Types

Mating type	G_0 families			F_2 families	
	12	*45*	*20*	*2×10*	*51×35*
np×nn	219	232	223	122	61
ll×lm	221	246	240	146	67
hk×hk	134	134	132	236	396
ef×eg	13	10	8	15	8
ab×cd	15	28	13	2	0
All	602	650	616	521	532

Distribution of SNP mating types among parents reflects breeding history and an ascertainment bias

Distributions of SNPs over Families

No. of maps	No. of SNPs
1	262
2	285
3	290
4	180
5	68
Sum	1085

= 823

76% of SNPs are mapped in two or more families

Assignment of 1078 SNPs to 618 Genome Scaffolds

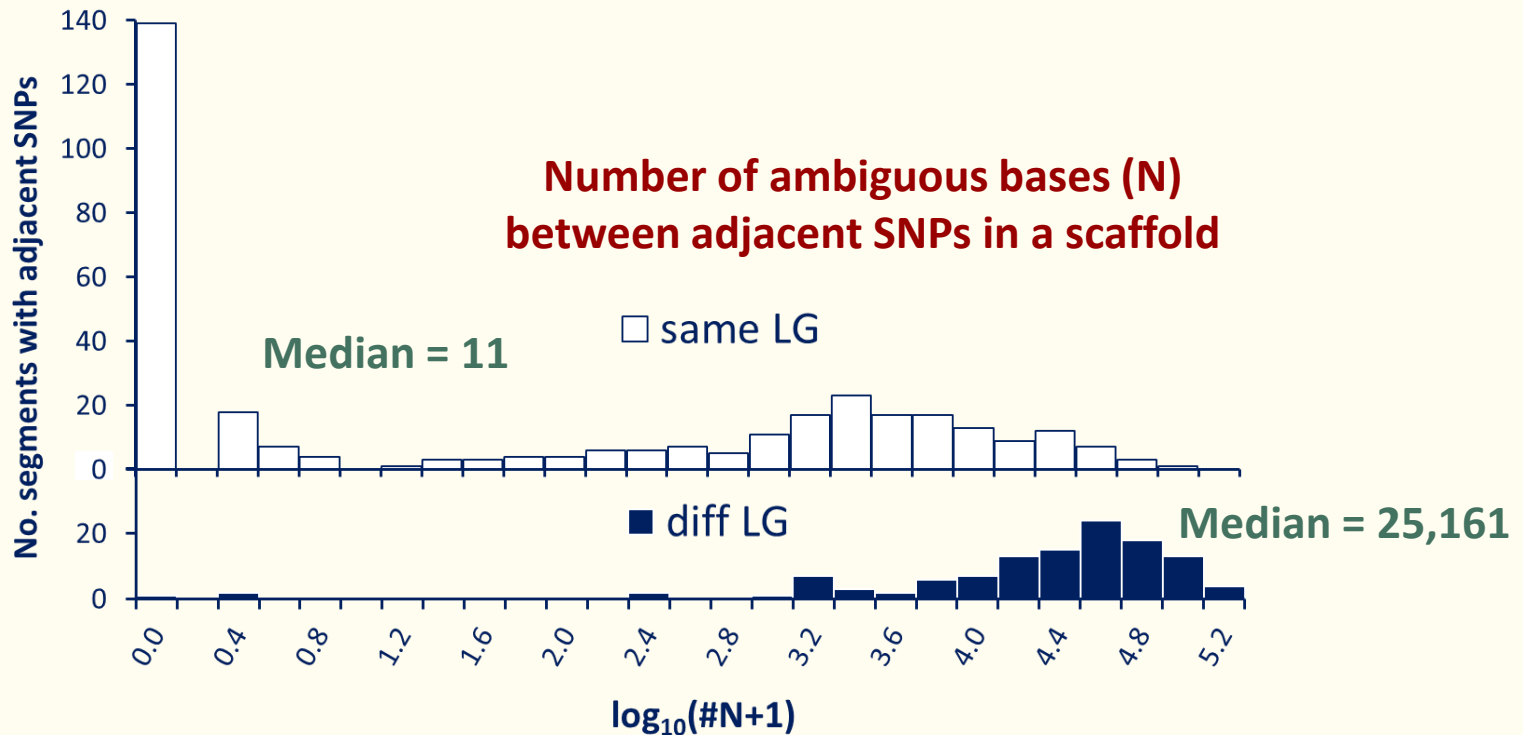| No. of SNPs per scaffold	No. of linkage groups to which scaffolds map				
	1	2	3	4	Sum
1	358	-	-	-	358
2	106	37	0	0	143
3	39	27	3	0	69
4	11	13	2	0	26
5	3	4	5	1	13
6	0	3	2	0	5
7	1	0	3	0	4
Sums, ≥2 SNPs/scaffold	**160**	**84**	**15**	**1**	**260**

Of 260 genome scaffolds with two or more mapped SNPs, 100 (38%) have SNPs that map to different linkage groups

Mapping SNPs in Genome Scaffolds

Example SNPs	12	45	20	2x10	51x35
1004-G495596A	10	—	10	—	10
1004-T617736C	—	—	6	—	6
1009-A469251G	—	8	8	8	—
1009-C1006340T	10	10	10	—	10
1009-C1020482T	10	—	10	—	10
100-A378029G	6	6	6	—	—
100-C260475A	3	—	3	—	—

Statistics on Scaffolds & Adjacent SNPs That Map to 1 or > 1 Linkage Group

Scaffold maps to:	No. of scaffolds	Length (kbp)
1 LG	160	403
≥ 2 LG	100	735

Adjacent SNPs map to:	No. of adjacent SNPs	Ave. inter-SNP distance (bp)
Same LG	337	58,224
Different LG	118	273,322

Number of ambiguous bases (N) between adjacent SNPs in a scaffold

□ same LG

Median = 11

■ diff LG

Median = 25,161

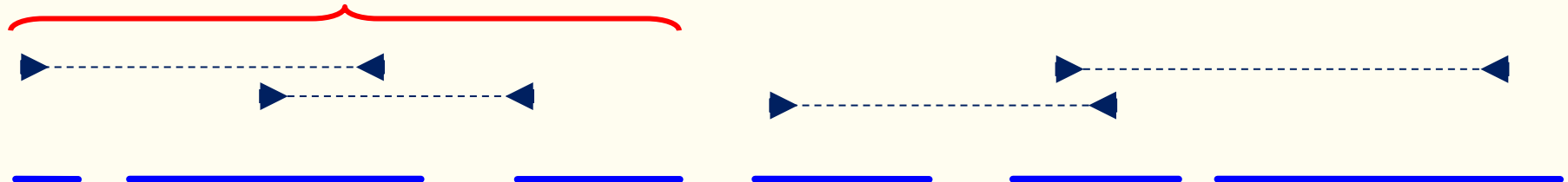No. segments with adjacent SNPs

$\log_{10}(\#N+1)$
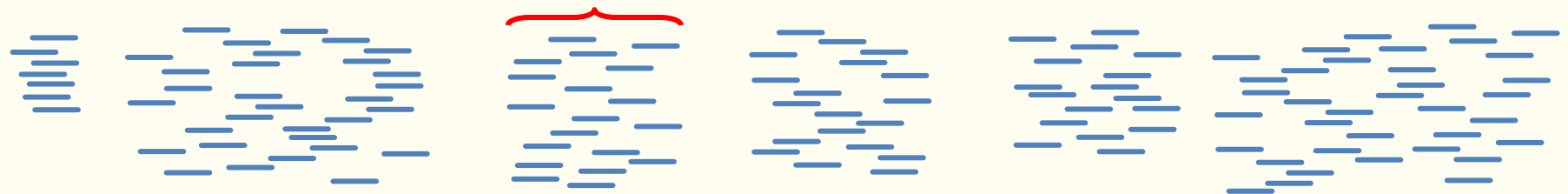
Implications for Genome Assembly

Ultimately, map scaffolds into the 10 chromosomes

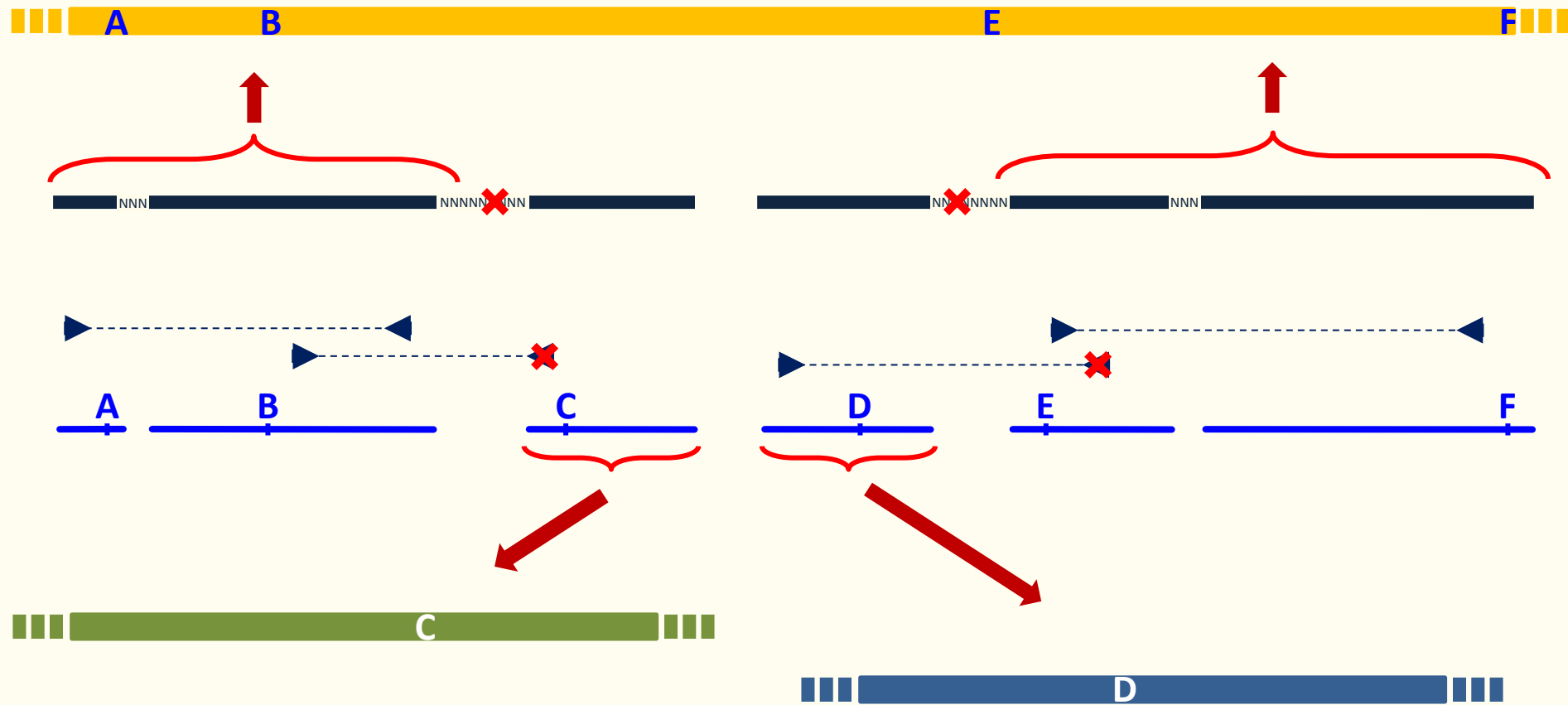Long PE-reads join contigs into scaffolds (N_{90} = 1,669)

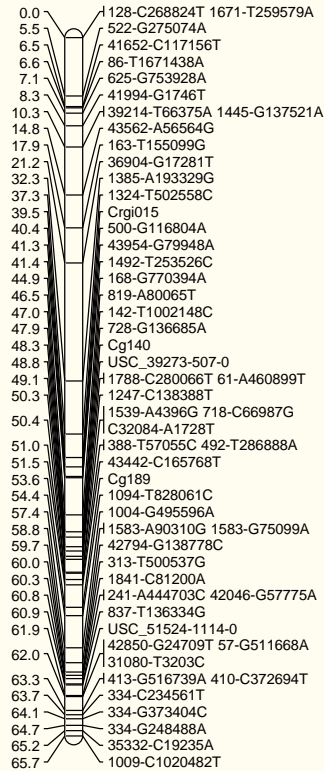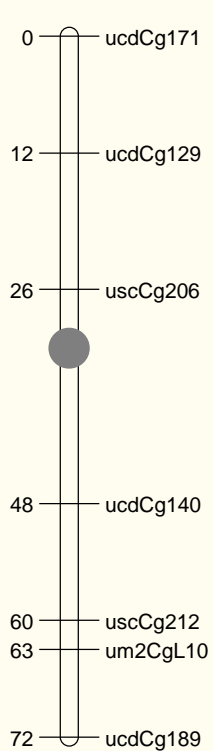Reads aligned into contigs (N_{90} = 25,658)

Short reads from Illumina sequencing

Hypothesis about Genome Assembly

Second-generation Map of LG 10
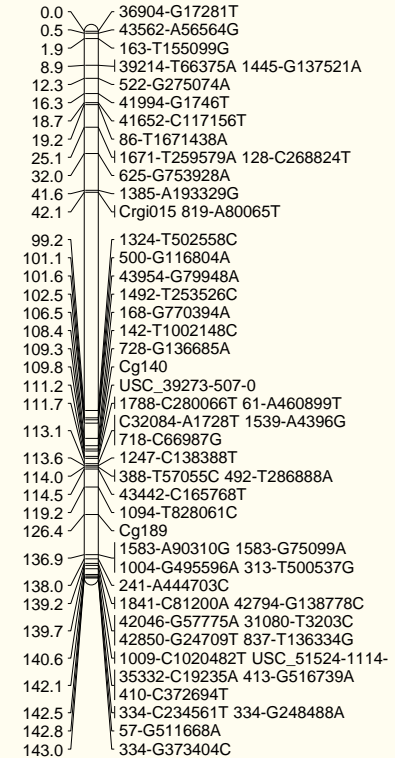

First-generation map, centromere located

Regression map

65.7 cM

Maximum likelihood map

143.0 cM

Plausibility of Marker Order on LG 10

Locus	Segregation	Position
57-C511666T	<hkxhk>	
31080-G3201A		
1583-C75097T	<hkxhk>	0.0
140-G1...		
...-G3201A	<nnxnp>	0.0
...83-C75097T	<hkxhk>	0.6
140-G184494A	<hkxhk>	2.3
1788-C509084T	<hkxhk>	4.0
43122-G186710A	<nnxnp>	490.9
42634-C59095G	<hkxhk>	5015.7
...4-C83122T	<nnxnp>	5020.9
...-C390738T	<nnxnp>	5026.2
522-C1...		
43358-G1572...		
1032-A418047C	<nnxnp>	5026.2
40246-A49615G		
36904-C17279A	<hkxhk>	
1729-C168040T	<hkxhk>	5...
...hk>	5034.9
1445-T120774A	<nnxnp>	5301.5
168-A770394G	<nnxnp>	5336.2
1492-T253526C	<nnxnp>	5336.2
1539-G6263T	<hkxhk>	5398.4
1247-C138306T	<hkxhk>	5422.7
425-C64538T	<hkxhk>	5427.6
1104-C108927A	<hkxhk>	5427.6
Cg140	<abxcd>	5430.2
42528-T105540C	<hkxhk>	5443.8
42528-C145563T	<hkxhk>	5444.3
140-A196754G	<nnxnp>	5445.5
L10	<nnxnp>	5449.7
1583-T90308C	<nnxnp>	5454.0
1009-T1006340C	<nnxnp>	5455.1
396-T391712A	<hkxhk>	10471.1
574-T25910G	<hkxhk>	10471.1
43630-T96559A	<hkxhk>	10472.2
86-A1671436T	<hkxhk>	10472.8
625-G805250A	<hkxhk>	10474.5
951-A40849C	<hkxhk>	10475.2
985-T40887C	<hkxhk>	10475.7
1108-G292604A	<hkxhk>	10475.7

Distortion from early genetic inviability

Log$_{10}$ (NN Fit, cM)

-Log$_{10}$ (P...

Fit of markers to observed recombination

Framework Marker Order on LG 10
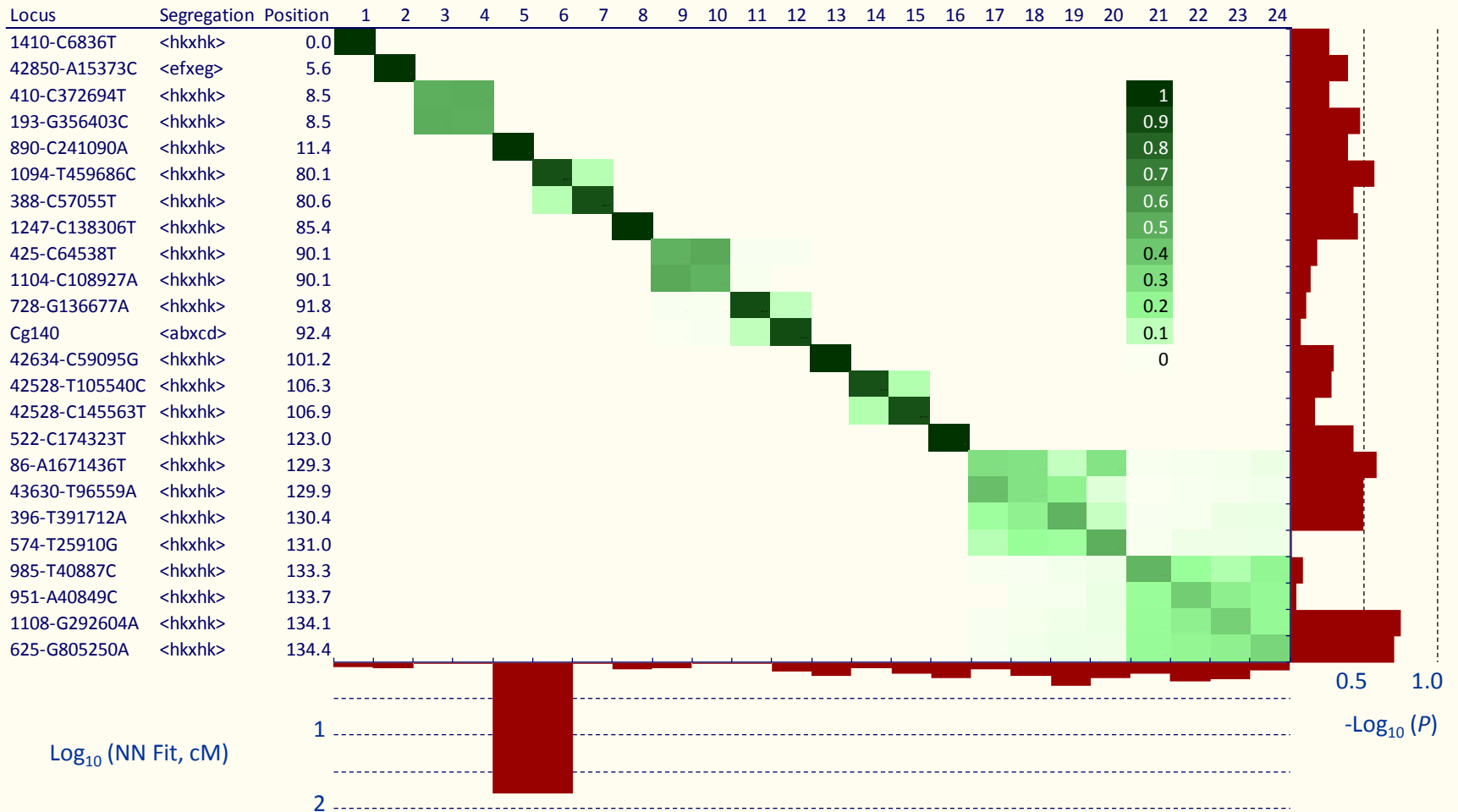
Improvement in Marker Orders

All markers

$y = 0.8183x$
$r^2 = 0.657$

Reg marker rank order

ML marker rank order

Framework markers

$y = 0.9945x$
$r^2 = 0.9523$

RG bi-parental marker rank order

ML bi-parental marker rank order
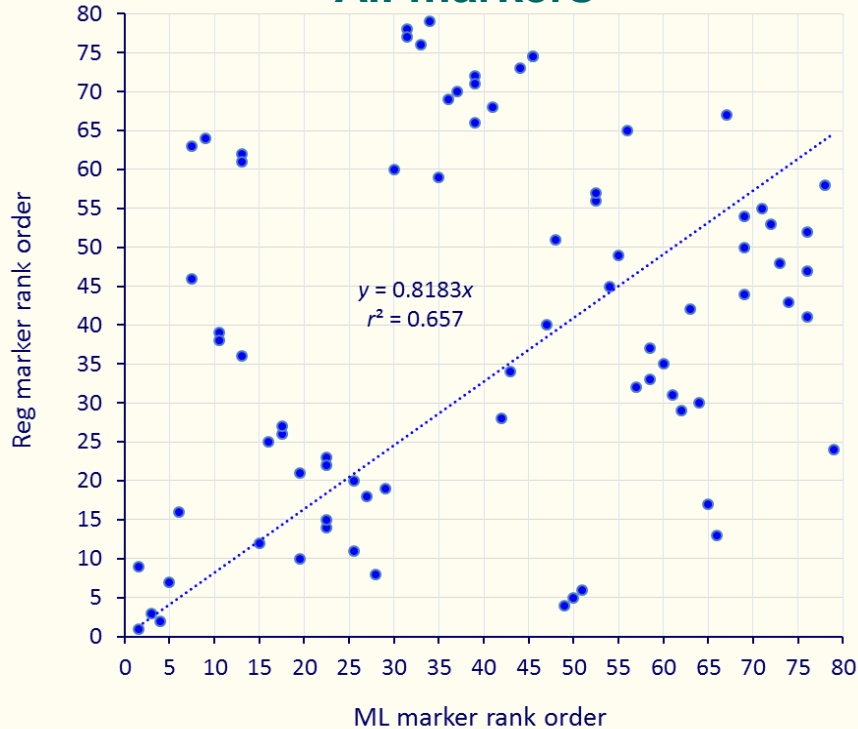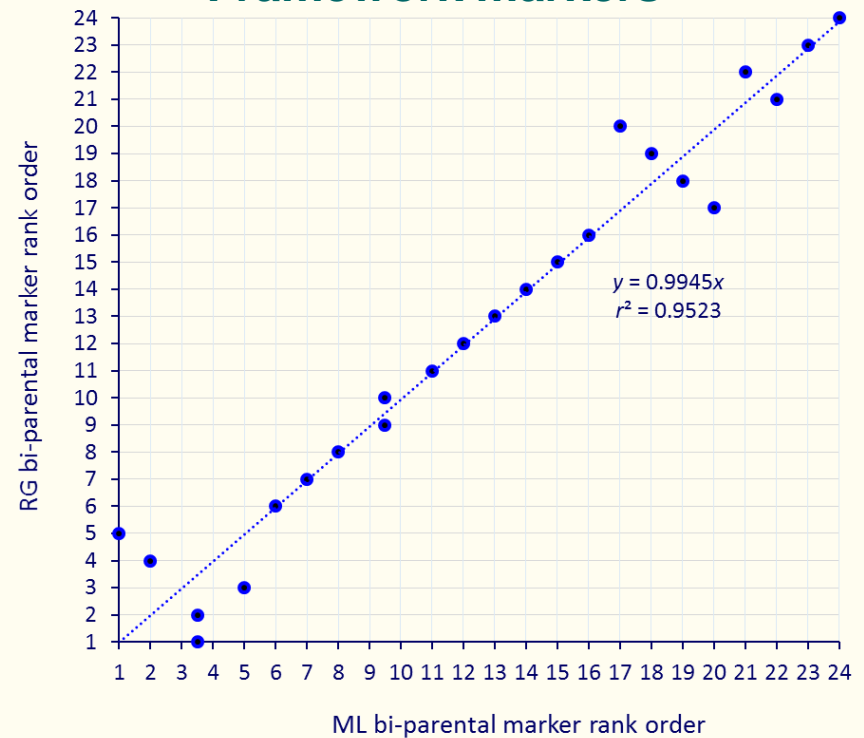
Causes of discrepancies between RG and ML maps
1. Markers segregating from one parent
2. Early viability selection, leading to distorted ratios
3. Genotyping errors (microsatellites >>> SNPs)

Acknowledgements

- Pat Gaffney, Ximing Guo, GIGASNP (USDA AFRI)

- Andrew Gracey, Grace Shin (USC); Manoj Samanta (Systemix)

- Guofan Zhang, Li Li, Haigang Qi (IOCAS); Jun Wang et al. (BGI)

- Dan Rokhsar (UC Berkeley, DOE JGI)

- Members of the Oyster Genome Consortium (Greg Warr, Kim Reece, Pierre Boudry)

USDA

United States Department of Agriculture

National Institute of Food and Agriculture