Use of Next-Generation Sequencing in the Pacific Oyster to Discover and Genotype SNP Markers for Building Third-Generation Linkage Maps

Alberto Arias-Perez
Dennis Hedgecock

First-generation Maps

- Microsatellites

  - tandemly repeated motifs (e.g. CACACA... or ATCATCATC...)

  - highly polymorphic

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

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

- Amplified fragment length polymorphisms (AFLPs)

  - Digestion genomic DNA + ligation adaptors + selective amplification

  - No prior sequence information

  - Dominant markers

  - Non transferable

  - Guo et al. 2012 Marine Biotechnology

Second-generation Maps

- Single-nucleotide polymorphisms (SNPs)
    - Most abundant polymorphism
    - Usually biallelic
    - Low mutation rate
    - High-throughput
    - Multiple technologies and platforms

- Illumina Golden Gate assay
    - 1536 SNPs in coding sequences typed
    - 1095 mapped in one or more of five families
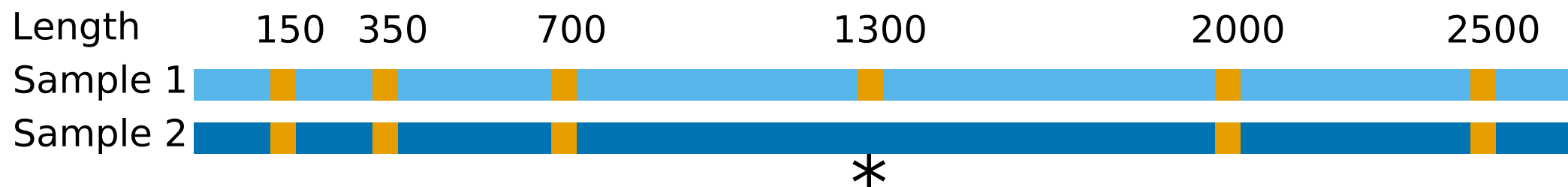   - Hedgecock et al. 2011 JSR, Hedgecock et al. in prep.

DNA Sequencing Costs

Cost per Raw Megabase of DNA Sequence
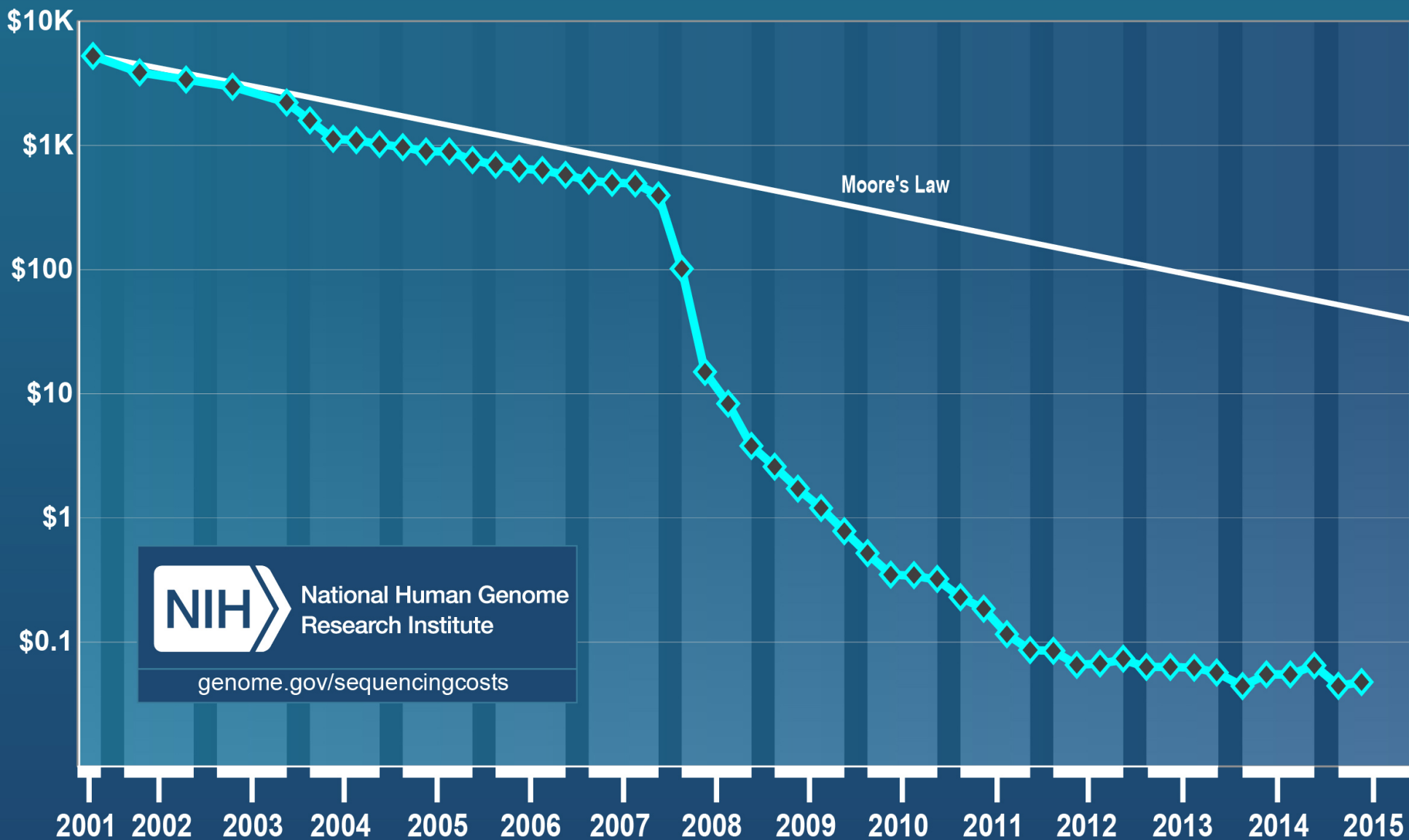
Moore's Law

National Human Genome Research Institute

genome.gov/sequencingcosts

"Genotyping by sequencing" Elshire et al. (2011)
PLoS ONE 6, no. 5 : e19379

"Genotyping by sequencing" Elshire et al. (2011) PLoS ONE 6, no. 5 : e19379

"Genotyping by sequencing" Elshire et al. (2011) PLoS ONE 6, no. 5 : e19379

"Genotyping by sequencing" Elshire et al. (2011) PLoS ONE 6, no. 5 : e19379

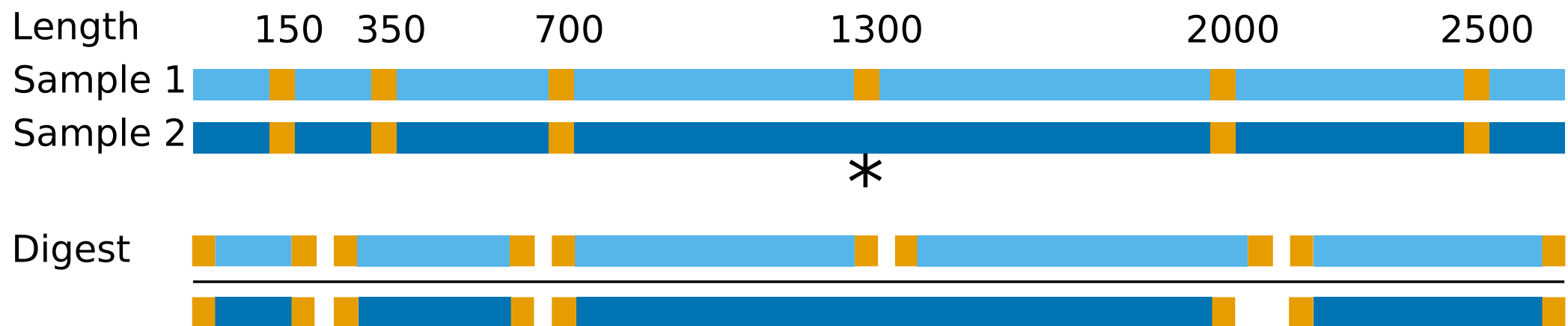Length 150 350 700 1300 2000 2500

Sample 1

Sample 2

Digest

Ligate adaptors
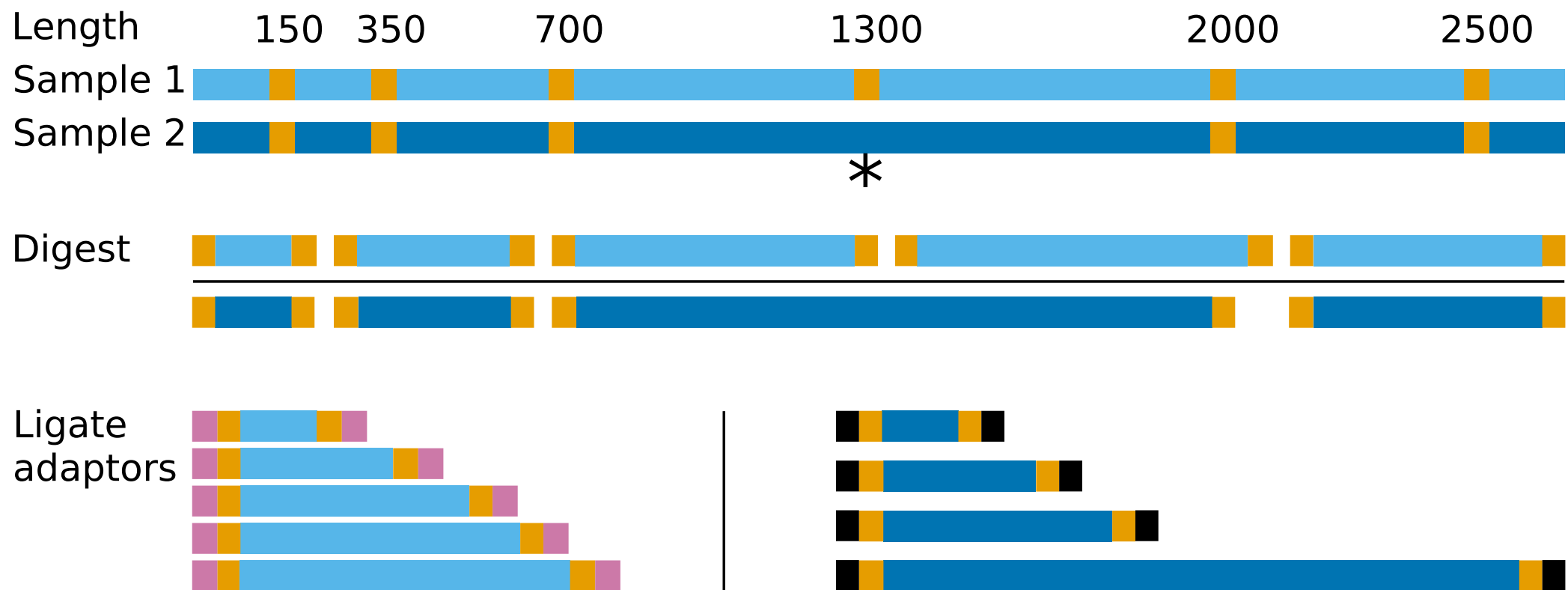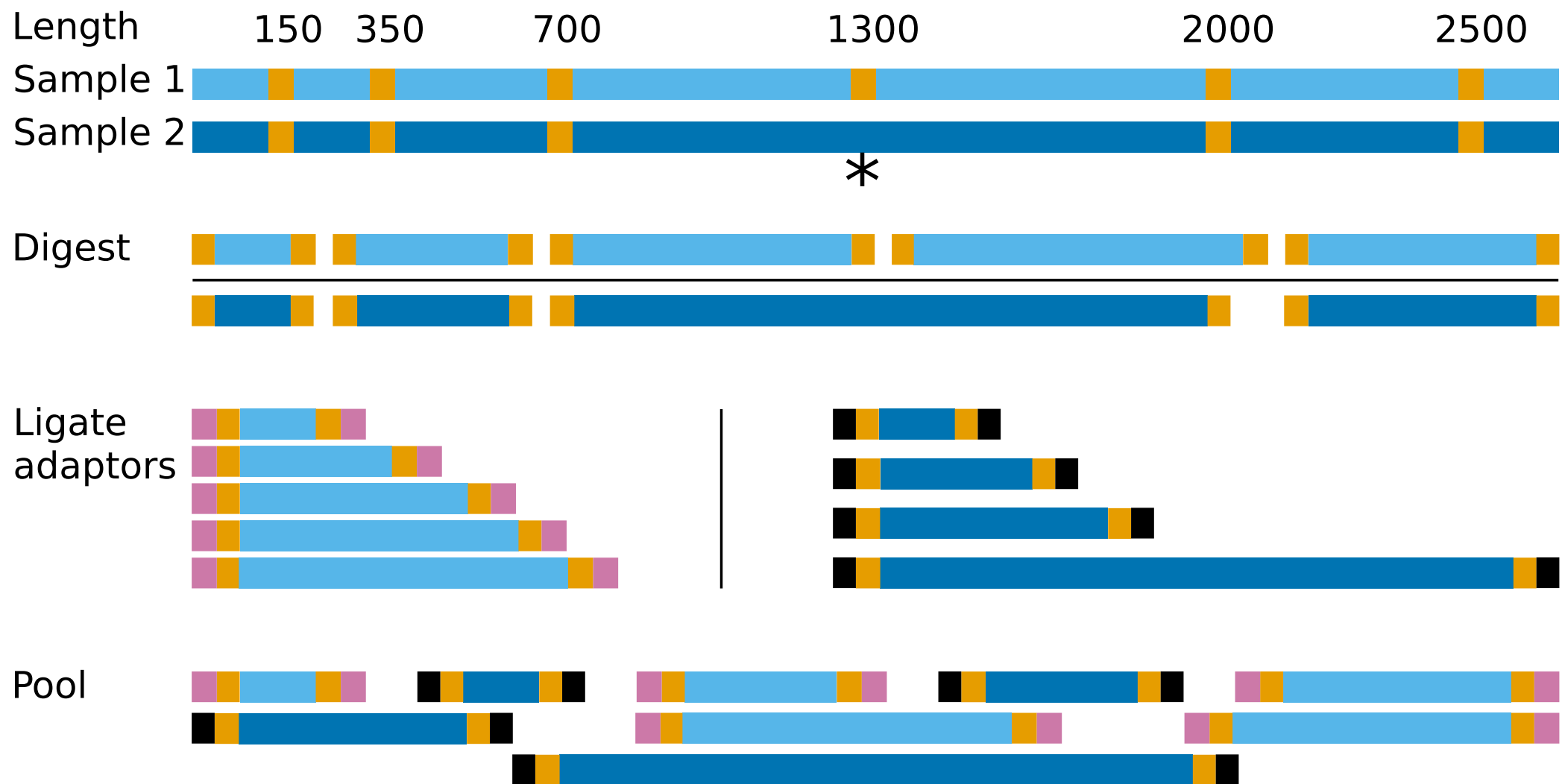
Pool

Modified from Davey et al. (2011) Nature Reviews Genetics 12:499

"Genotyping by sequencing" Elshire et al. (2011)
PLoS ONE 6, no. 5 : e19379
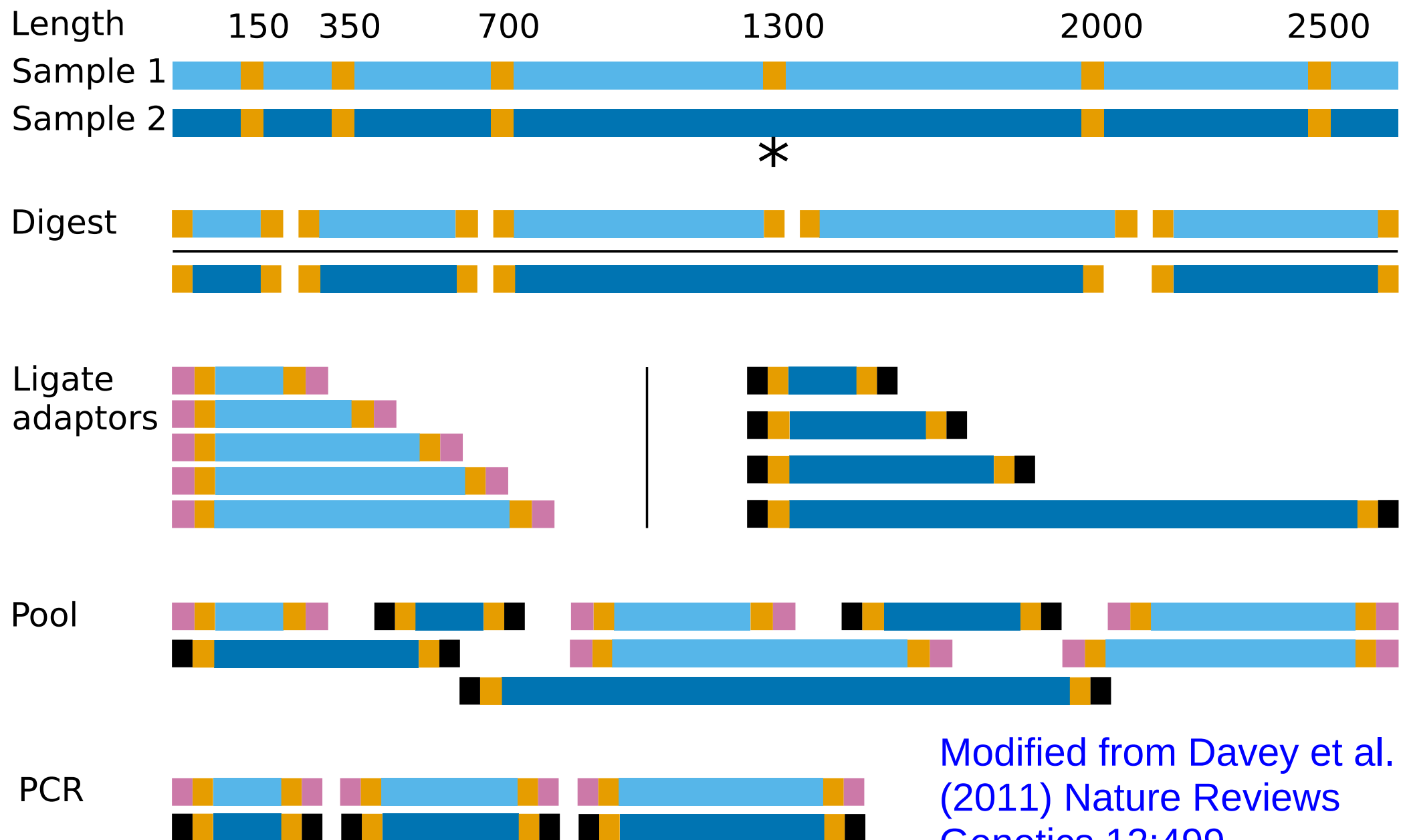
Length 150 350 700 1300 2000 2500

Sample 1

Sample 2

Digest

Ligate adaptors

Pool

PCR

Modified from Davey et al. (2011) Nature Reviews Genetics 12:499

Data Analysis

- Map to Reference

- Indel Realignment

- Base Recalibration

- Genotyping

- Variant Recalibration

Genome Analysis Toolkit (GATK)
Broad Institute
https://www.broadinstitute.org/gatk/

Data Analysis

- Map to Reference

- Indel Realignment

- Base

- Geno

- Varia

Data Analysis

- Map to Reference

- Indel Realignment

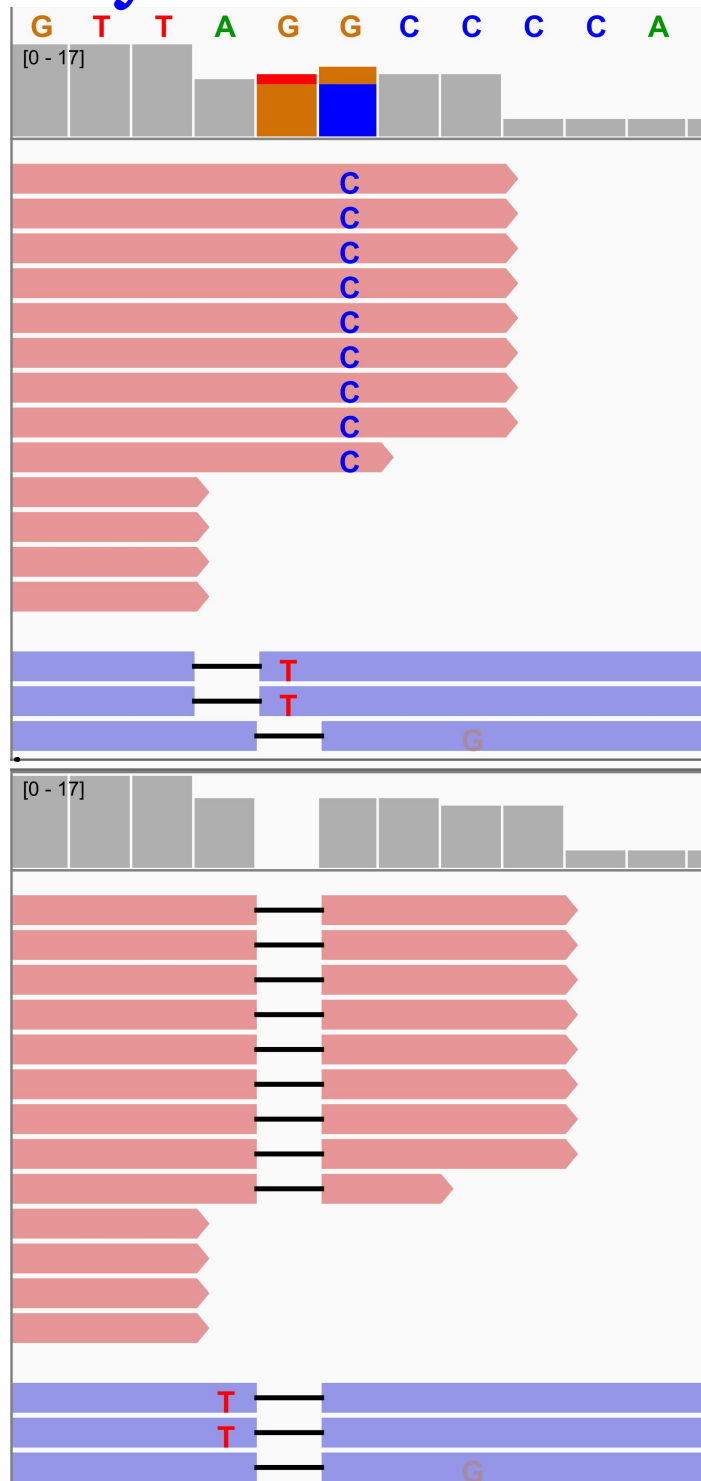- Base Recalibration

- Genotyping

- Variant Recalibration

Data Analysis

- Map to Reference

- Indel Realignment

- Base Recalibration

- Genotyping

- Variant Recalibration

Data Analysis

- Map to Reference

- Indel Realignment

- Base Recalibration

- Genotyping

- Variant Recalibration

```
SCA.  POS       REF  ALT     FORMAT          51x35_001_m
99    110689    C    T       GT:AD:DP:GQ:PL  0/0:51,0:51:99:0,120,1800
99    110723    T    C       GT:AD:DP:GQ:PL  0/1:12,30:42:99:608,0,456
122   76155     G    A       GT:AD:DP:GQ:PL  1/1:1,20:21:38:458,38,0
122   78018     A    G       GT:AD:DP:GQ:PL  0/0:25,0:25:65:0,65,607
125   33062     T    C       GT:AD:DP:GQ:PL  0/1:8,6:14:99:99,0,153
```
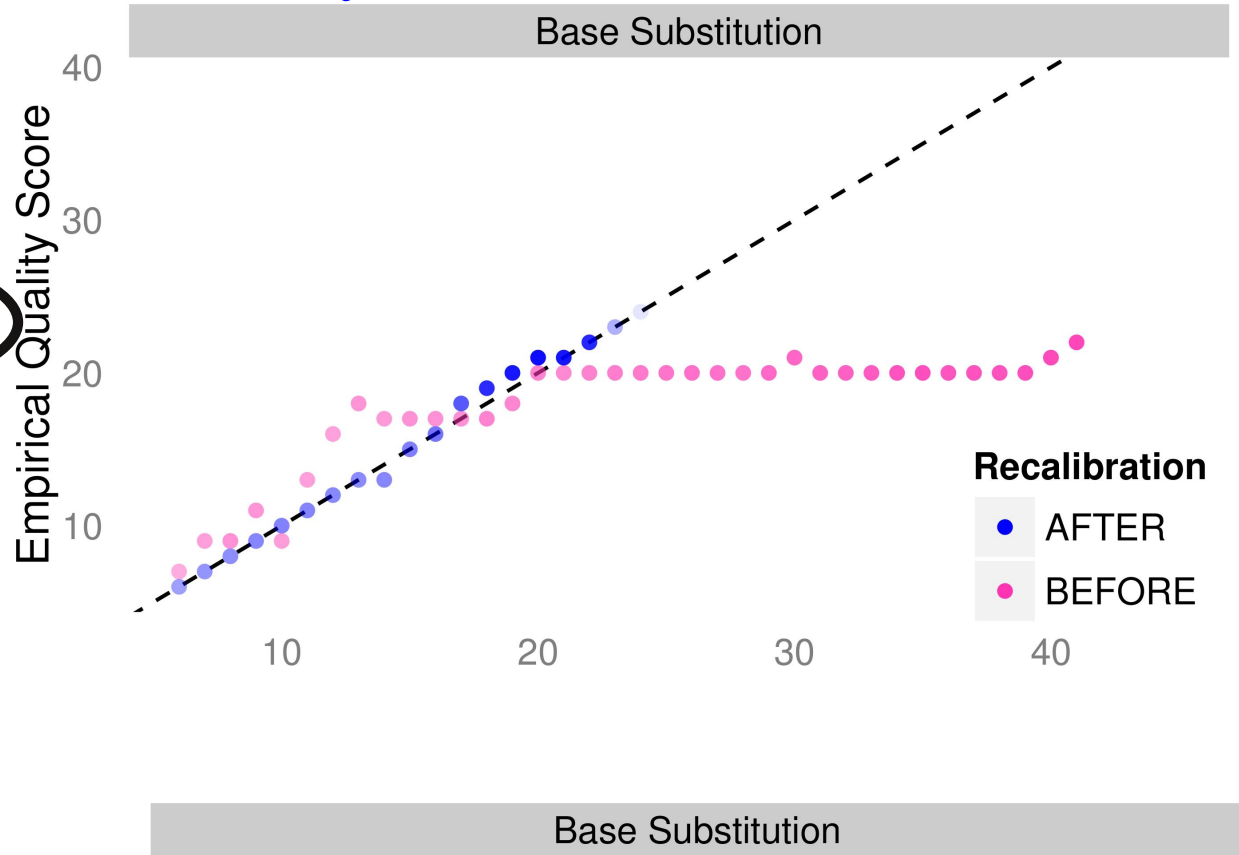
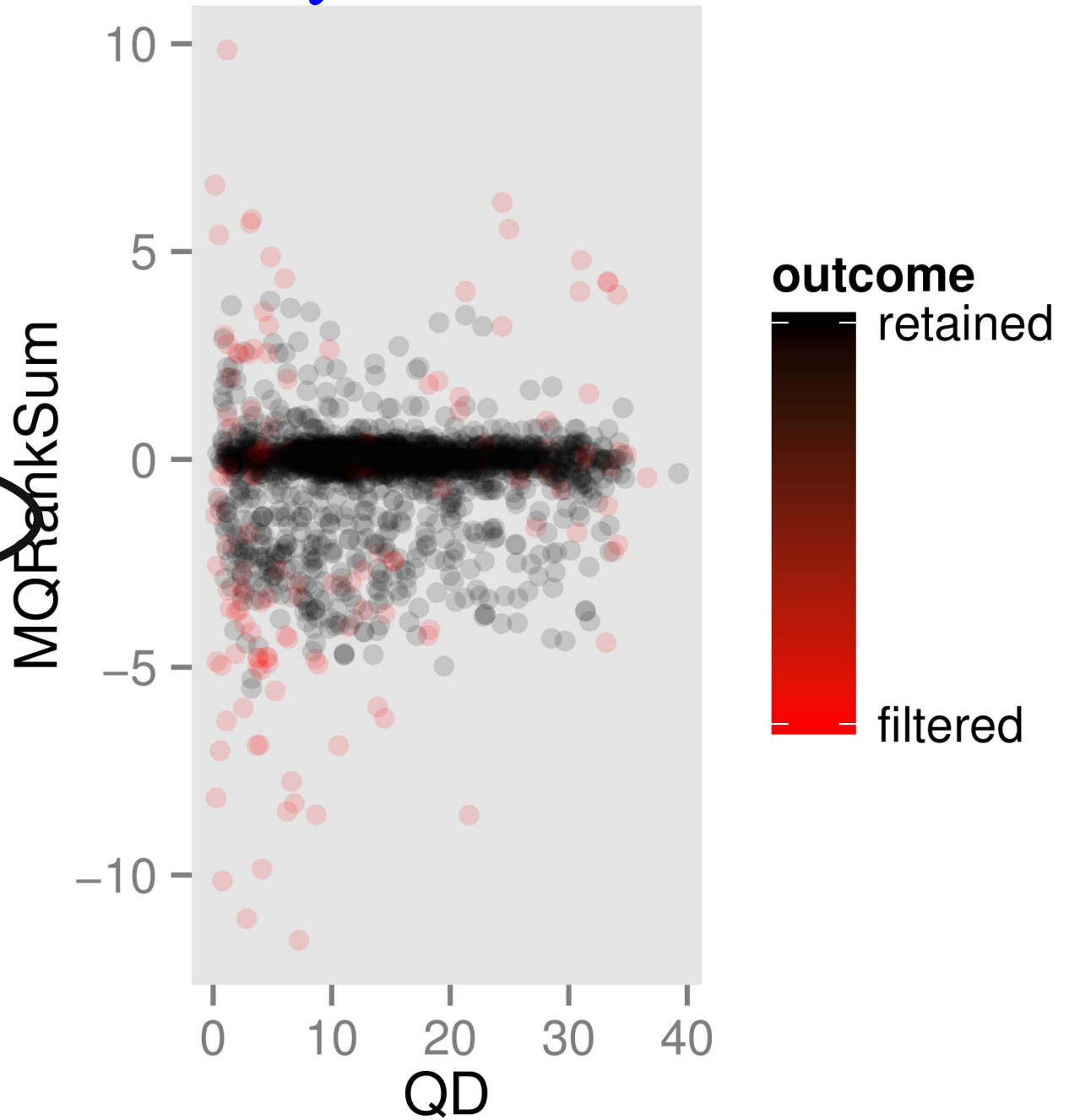# Data Analysis

- Map to Reference

- Indel Realignment

- Base Recalibration

- Genotyping

- Variant Recalibration

Full-sib family (two F1 parents and 129 F2 offspring)

6 libraries (2 24-plex, 4 48-plex)

Full-sib family (two F1 parents and 129 F2 offspring)

6 libraries (2 24-plex, 4 48-plex)
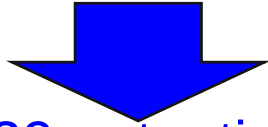
1,922,129 potential variants

5303 / 11,969 scaffolds (Zhang et al 2012 Nature 490:49)

Full-sib family (two F1 parents and 129 F2 offspring)
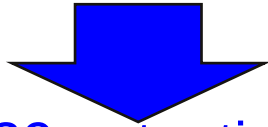
6 libraries (2 24-plex, 4 48-plex)

1,922,129 potential variants

5303 / 11,969 scaffolds (Zhang et al 2012 Nature 490:49)

- low coverage

- non-segregating variants

- variants with unexpected offspring genotypes

- individuals with a low number of sequences

Full-sib family (two F1 parents and 129 F2 offspring)

6 libraries (2 24-plex, 4 48-plex)

1,922,129 potential variants

5303 / 11,969 scaffolds (Zhang et al 2012 Nature 490:49)

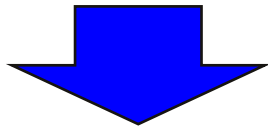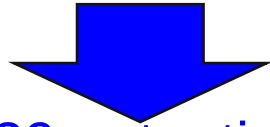
- low coverage

- non-segregating variants

- variants with unexpected offspring genotypes
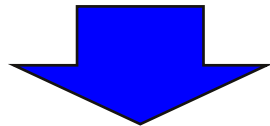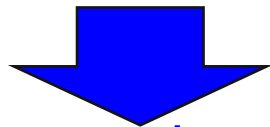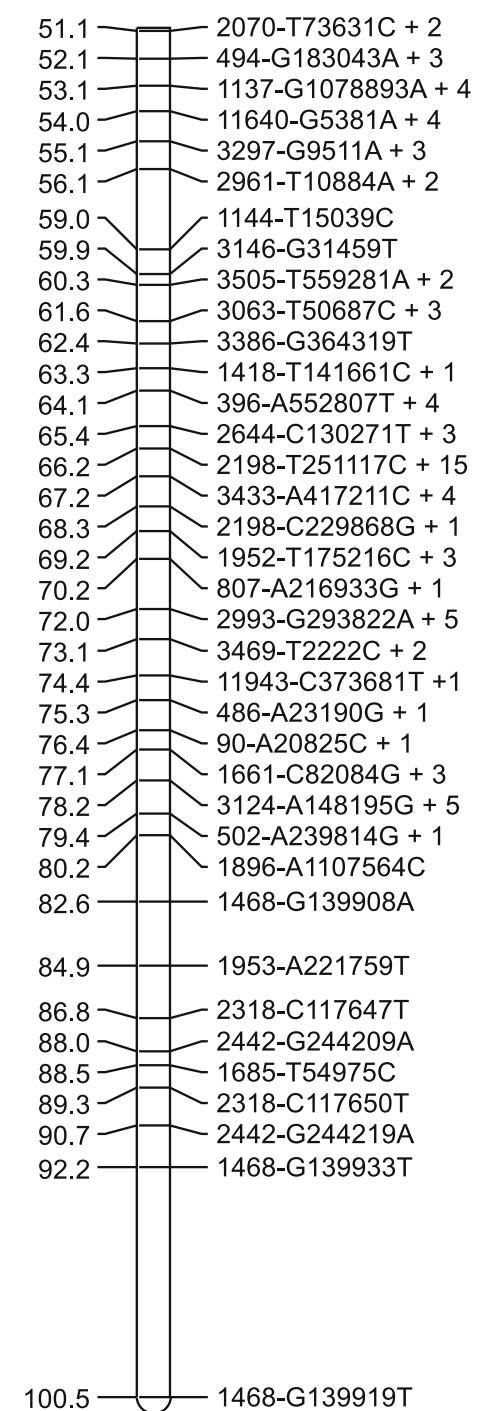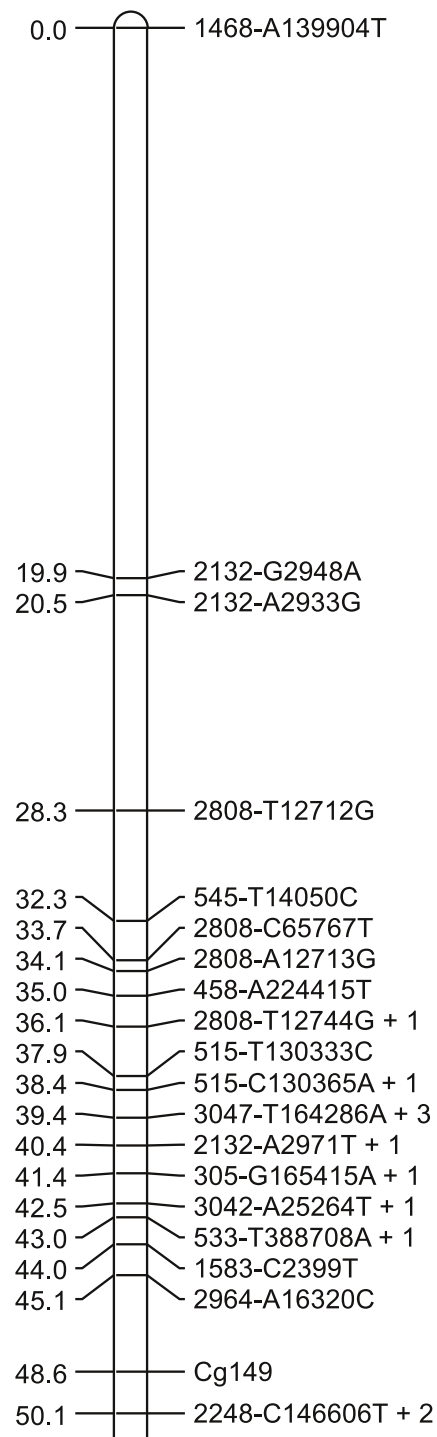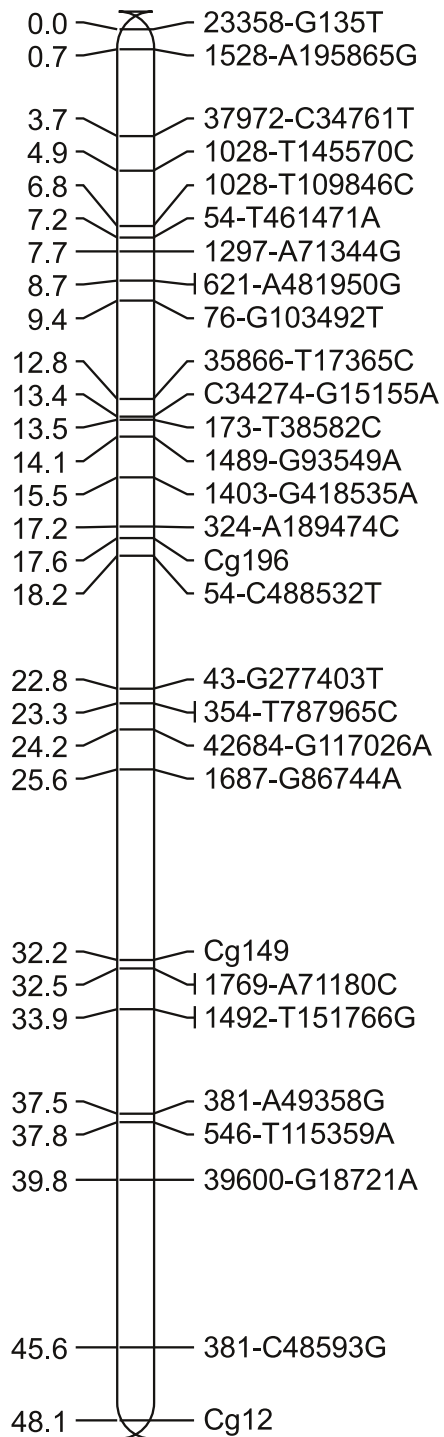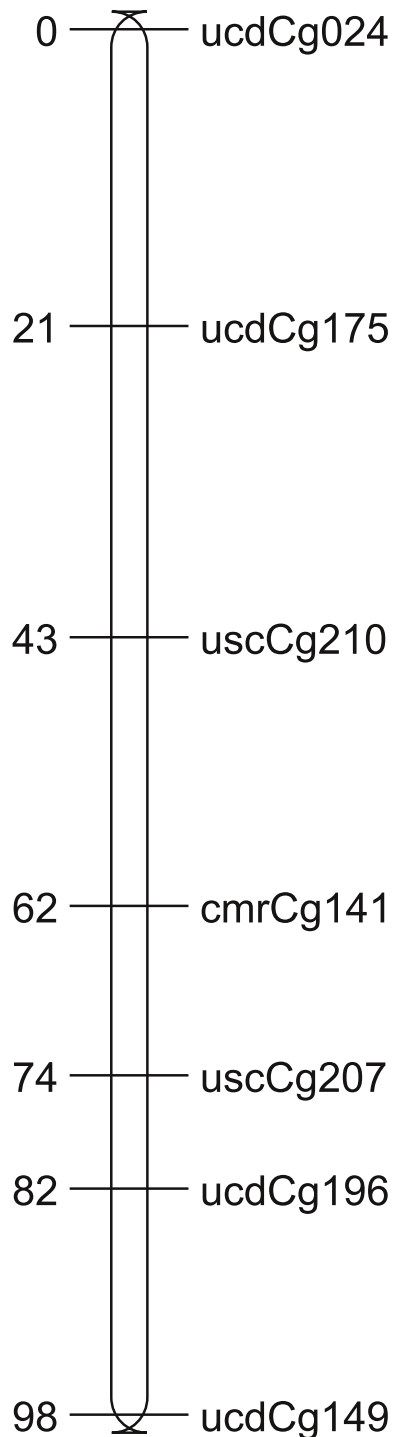
- individuals with a low number of sequences

3664 segregating SNP markers.

ten main linkage groups, 1442 markers, 874.5 cM (average spacing of 0.8 cM)

Maps of LG 08

Map 1

Position	Marker
0	ucdCg024
21	ucdCg175
43	uscCg210
62	cmrCg141
74	uscCg207
82	ucdCg196
98	ucdCg149

Map 2

Position	Marker
0.0	23358-G135T
0.7	1528-A195865G
3.7	37972-C34761T
4.9	1028-T145570C
6.8	1028-T109846C
7.2	54-T461471A
7.7	1297-A71344G
8.7	621-A481950G
9.4	76-G103492T
12.8	35866-T17365C
13.4	C34274-G15155A
13.5	173-T38582C
14.1	1489-G93549A
15.5	1403-G418535A
17.2	324-A189474C
17.6	Cg196
18.2	54-C488532T
22.8	43-G277403T
23.3	354-T787965C
24.2	42684-G117026A
25.6	1687-G86744A
32.2	Cg149
32.5	1769-A71180C
33.9	1492-T151766G
37.5	381-A49358G
37.8	546-T115359A
39.8	39600-G18721A
45.6	381-C48593G
48.1	Cg12

Map 3

Position	Marker
0.0	1468-A139904T
19.9	2132-G2948A
20.5	2132-A2933G
28.3	2808-T12712G
32.3	545-T14050C
33.7	2808-C65767T
34.1	2808-A12713G
35.0	458-A224415T
36.1	2808-T12744G + 1
37.9	515-T130333C
38.4	515-C130365A + 1
39.4	3047-T164286A + 3
40.4	2132-A2971T + 1
41.4	305-G165415A + 1
42.5	3042-A25264T + 1
43.0	533-T388708A + 1
44.0	1583-C2399T
45.1	2964-A16320C
48.6	Cg149
50.1	2248-C146606T + 2

Map 4

Position	Marker
51.1	2070-T73631C + 2
52.1	494-G183043A + 3
53.1	1137-G1078893A + 4
54.0	11640-G5381A + 4
55.1	3297-G9511A + 3
56.1	2961-T10884A + 2
59.0	1144-T15039C
59.9	3146-G31459T
60.3	3505-T559281A + 2
61.6	3063-T50687C + 3
62.4	3386-G364319T
63.3	1418-T141661C + 1
64.1	396-A552807T + 4
65.4	2644-C130271T + 3
66.2	2198-T251117C + 15
67.2	3433-A417211C + 4
68.3	2198-C229868G + 1
69.2	1952-T175216C + 3
70.2	807-A216933G + 1
72.0	2993-G293822A + 5
73.1	3469-T2222C + 2
74.4	11943-C373681T +1
75.3	486-A23190G + 1
76.4	90-A20825C + 1
77.1	1661-C82084G + 3
78.2	3124-A148195G + 5
79.4	502-A239814G + 1
80.2	1896-A1107564C
82.6	1468-G139908A
84.9	1953-A221759T
86.8	2318-C117647T
88.0	2442-G244209A
88.5	1685-T54975C
89.3	2318-C117650T
90.7	2442-G244219A
92.2	1468-G139933T
100.5	1468-G139919T

Mapping Genome Scaffolds

| SNPs per scaffold	Linkage groups to which scaffolds map				
	1	2	3	5	Sum
1	378	0	0	0	
2	100	37	0	0	137
3	25	28	6	0	59
4	7	12	4	0	23
5	2	5	2	0	9
6	1	2	1	0	4
7	0	0	0	1	1
8	0	0	1	0	1
9	0	0	0	1	1
11	2	0	0	0	2
Sum	515	84	14	2	615

Of 237 genome scaffolds with two or more SNPs, 100 (42.2%) have SNPs that map to different linkage groups.

Analysis of Adjacent Markers

Distance (bp) between adjacent markers in a scaffold
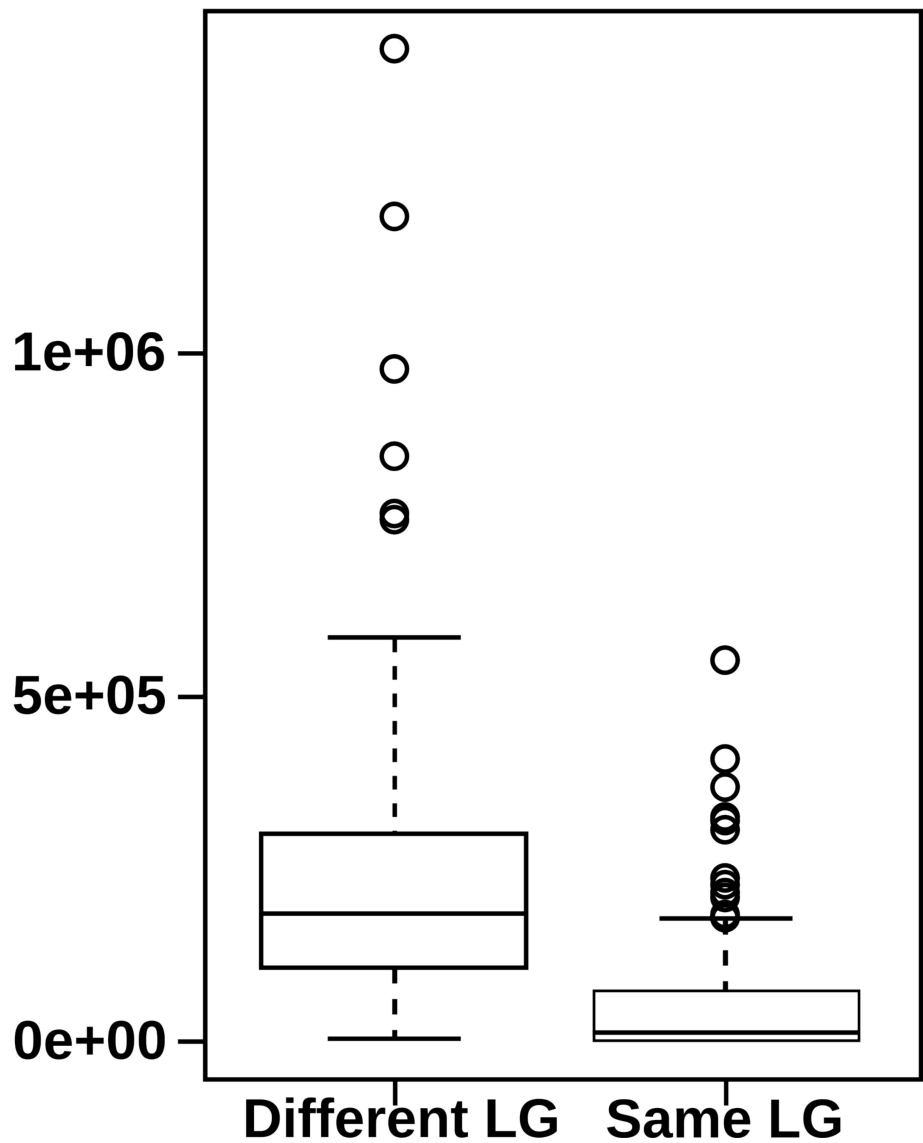
Number of ambiguous bases between adjacent markers in a scaffold
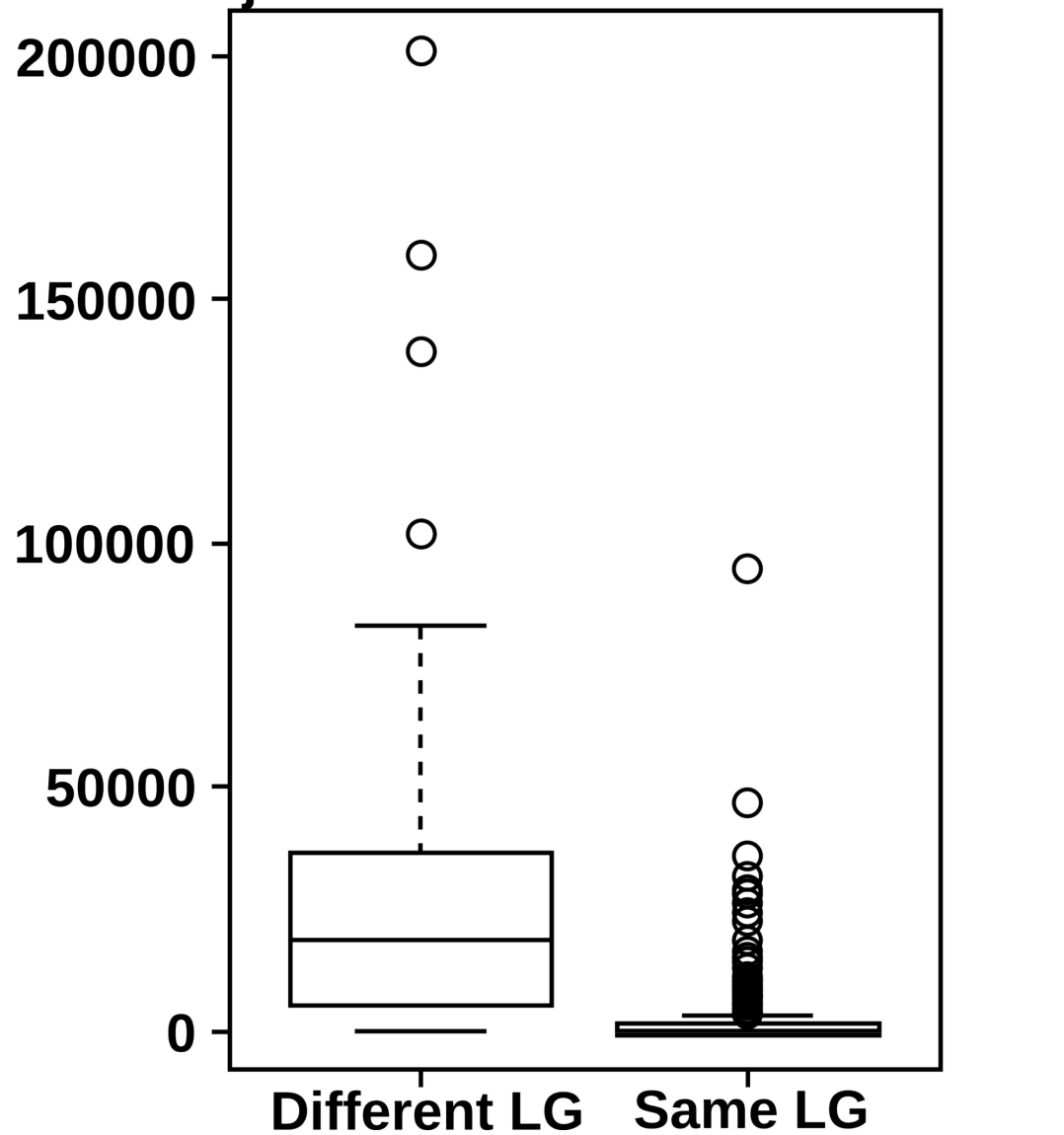
Conclusions

• Linkage map information will help to identify misassembled scaffolds and to improve genome assembly

• The high density maps will allow to study quantitative traits in detail and would help future practical breeding programs

Acknowledgements

Taylor Shellfish Farms

Joth Davis

Edward Buckler's Lab - Cornell University

Daniel Campo Falgueras - UPC Genome & Cytometry Core (University of Southern California)

Charles Nicolet - Epigenome Center (University of Southern California)

Mikhail Matz - University of Texas at Austin

Use of Next-Generation Sequencing in the Pacific Oyster to Discover and Genotype SNP Markers for Building Third-Generation Linkage Maps

Alberto Arias-Perez
Dennis Hedgecock