Supplementary Information for:

Properties of different selection signature statistics and a new strategy for combining them

Simulation

We have used the software msms (Ewing and Hermisson 2010) to simulate neutral scenarios and scenarios with selection. In this study, the Structured Population Models of msms were employed to simulate the scenario with multiple subpopulations.

The command line for the neutral scenarios is: java –Xmx500g -jar msms.jar -N 10000 -ms 2***sample_size** 1000 -s **SNP_number** -r 4000 -I 3 **sample_size sample_size** 0 0 -ma x 0 5 0 x 5 0 0 x > Neu, where **sample_size** is the sample size and **SNP_number** denotes the number of SNPs. In our case **sample_size** = {10, 30, 50, 70, 90} and **SNP_number** = {160, 800, 4000, 20000, 100000} that corresponded to the marker interval {62.5 kb, 12.5 kb, 2.5 kb, 0.5 kb, 0.1 kb} in 10 Mb simulated genome fragment. In this case, first two subpopulations were separately defined as Neu_1 and Neu_2, the migration is forbidden among them.

The command line for the divergent selection scenarios is: java -Xmx500g –jar msms.jar -N 10000 -ms **sample_size** 1000 -s **SNP_number** -r 4000 -seed num. -SAA **0** -SAa **0** -Saa **0** -Sp 0.5 -SF 0 > noSel and java -Xmx500g –jar msms.jar -N 10000 -ms **sample_size** 1000 -s **SNP_number** -r 4000 -seed num. -SAA **selection_coe** -SAa **selection_coe**/2 -Saa 0 -Sp 0.5 -SF 0 **allele_frequecy** > Sel, Where **selection_coe** is the selection coefficient and **allele_frequecy** denotes the data for analysis were sampled when the frequency of the selected allele reached a predefined value. In our case **allele_freque**ncy={0.2, 0.4, 0.6, 0.8, 1.0}, **selection_coe** = {200, 400, 800, 1600, 3200} that corresponded to selection coefficient {0.005, 0.01, 0.02, 0.04, 0.08} and num. is a 64 bit number that can be specified either in hex with a 0x prefix or normal decimal. The same random number seed was used in both no selection and selection scenarios in hope of sharing the same initial frequency between two subpopulations. In this case, the position of SNPs was derived from Sel_2 for all scenarios. The divergent selection simulation here is weaker than that with two different selected directions. Note that the initial frequency of the selected allele (p_0) in both subpopulations is 1/2N when selection was introduced (see Introducing Selection in Manual of msms). For the divergent scenarios, we ignored the influence from the variance of SNP position between two subpopulations because we only care the 500Kb window around the selected loci in this case (see Method). In general, the effect of hitchhiking should have a greater impact on the neutral loci in this window than any other factors.

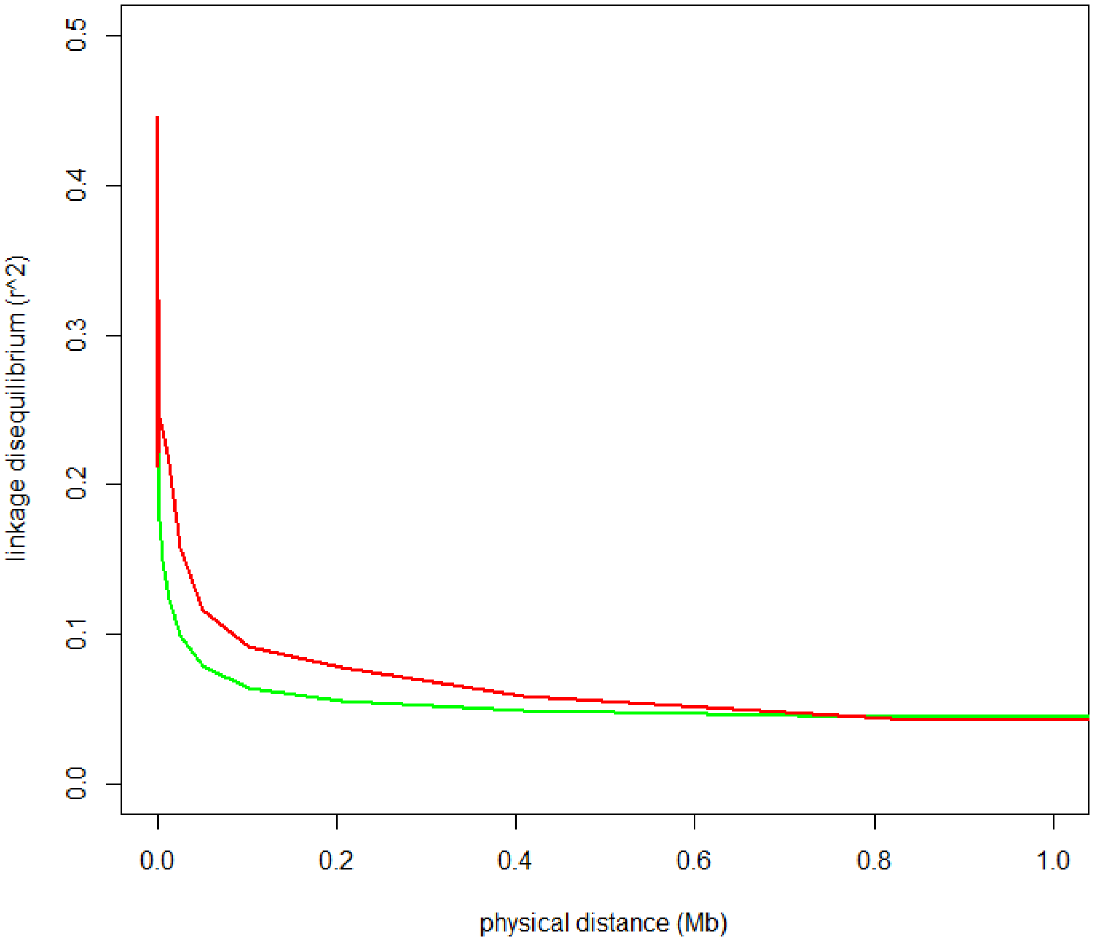
The command line for the parallel selection scenarios is: java –Xmx500g -jar msms.jar -N 10000 -ms 2***sample_size** 1000 -s **SNP_number** -r 4000 -SAA **selection_coe** -SaA **selection_coe**/2 -Saa 0 -Sp 0.5 -SF 0 **allele_frequecy** > Para. Based on the founder population, we further divided the simulated data into two equal subpopulations, which share the same haplotype distribution between two subpopulations at the time of split. In this case, two subpopulations were separately defined as Sel_1 and Sel_2.

In this study, -N 10000 denotes the population size, -r 4000 denotes there are 4000 points on the genome fragment (10 Mb) where recombination may occur, -Sp 0.5 represents that the selection has occurred at the middle of the simulation genomic regions. For within-population analysis tests, there is no comparison between populations. So, the neutral scenarios and selection scenarios were simulated refer to the ideas of Voight *et al*. (2006) and Pavlidis *et al*. (2013). The command line is: java –Xmx500g -jar msms.jar -N 10000 -ms **sample_size** 1000 -s **SNP_number** -r 4000 >Neu and java -Xmx500g –jar msms.jar -N 10000 -ms **sample_size** 1000 -s **SNP_number** -r 4000 -Sp 0.5 -SF 0 **allele_frequecy** –SAA **selection_coe** -SAa **selection_coe**/2 -Saa 0 >Sel.

Additional files

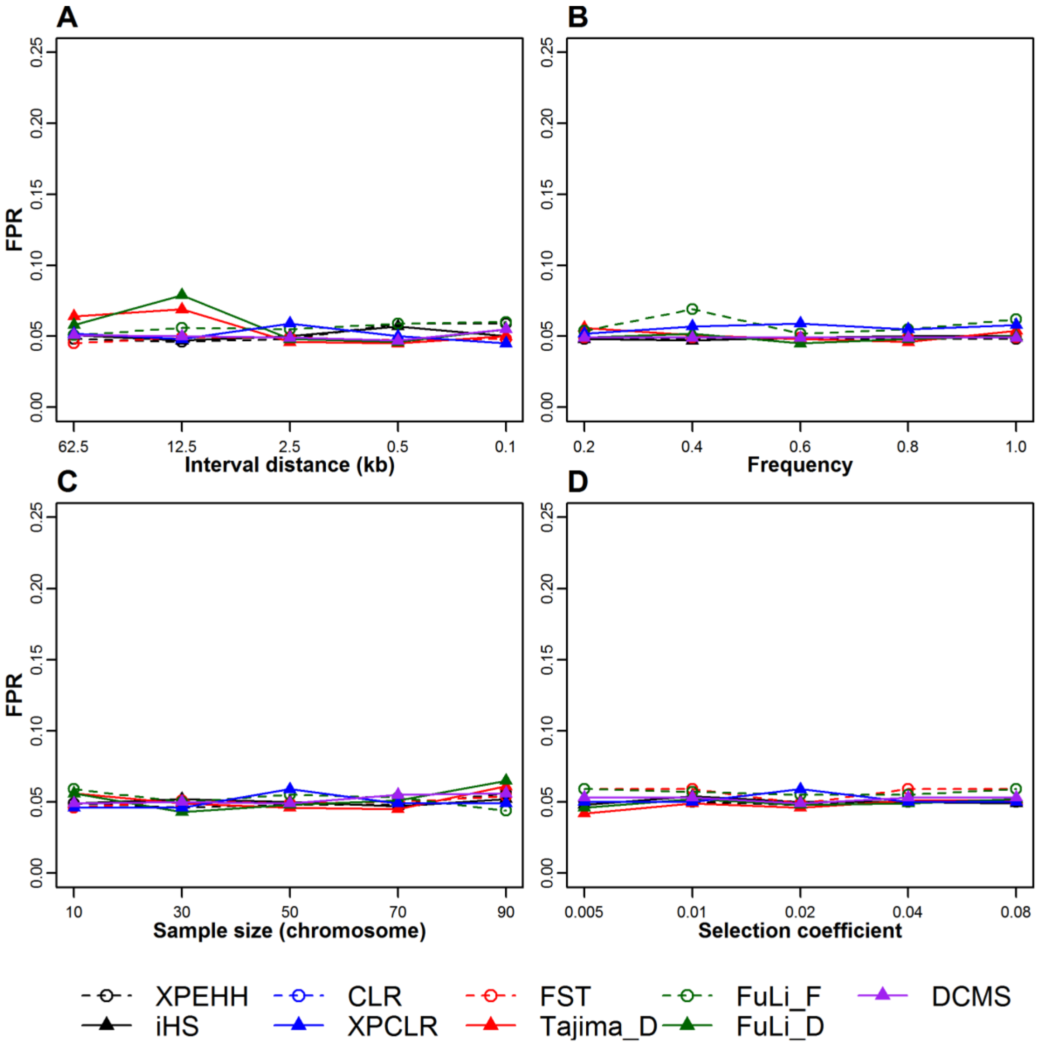
(Additional File 1)

Figure S1: A schematic representation of LD plotted as a function of distance in one repeat of the simulation data. The decay of LD is compared between selected region (the region from 4.5 Mb to 5.5 Mb appeared as red) and the whole simulation fragment (green).



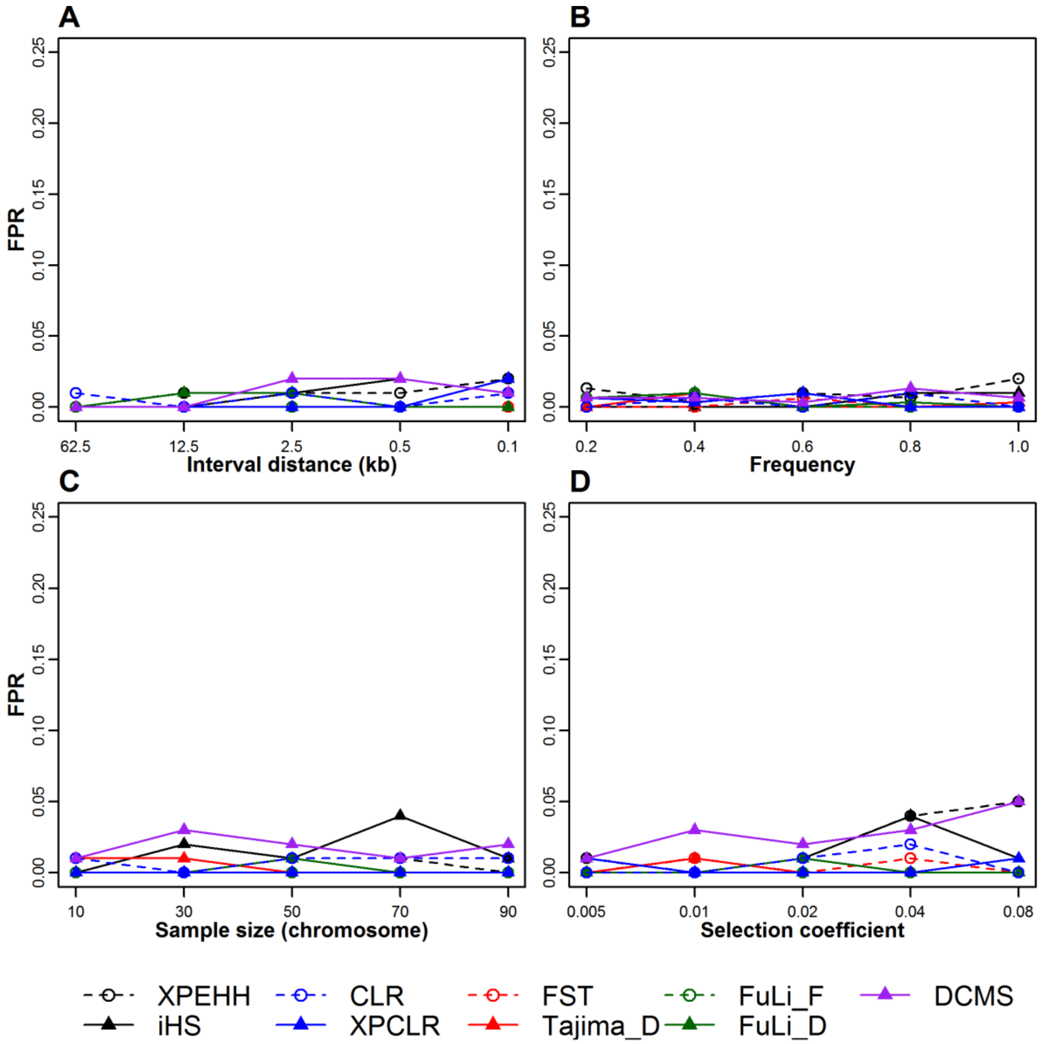
(Additional File 2)

Figure S2: **The False Positive Rate (FPR) of eight different selection signature test statistics and the novel combining strategy. (A) Marker interval distance; (B) Frequency of the selected allele; (C) Sample Size; (D) Selection coefficient.**



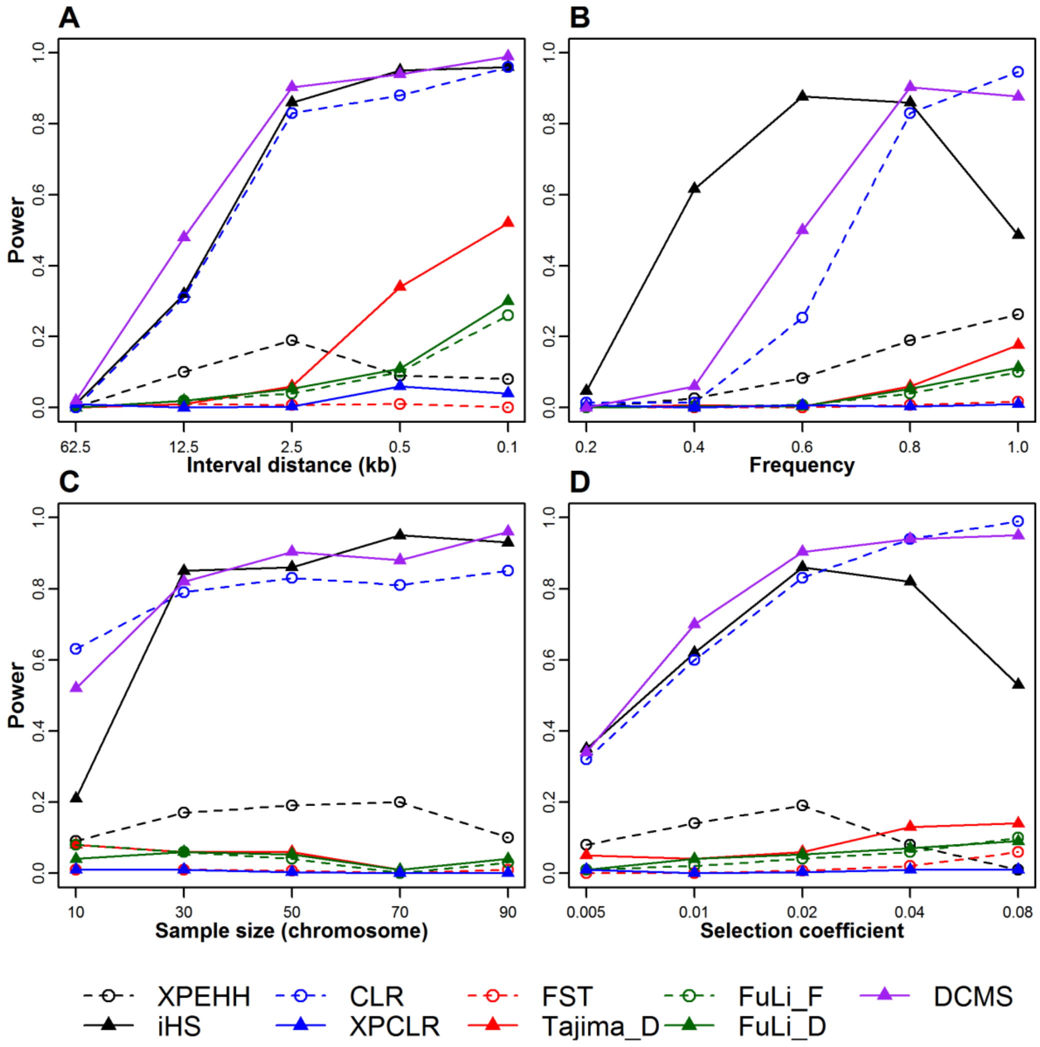
(Additional File 3)

Figure S3: The False Positive Rate (FPR) of eight different selection signature test statistics and the novel combining strategy in selection scenario. (A) Marker interval distance; (B) Frequency of the selected allele; (C) Sample Size; (D) Selection coefficient.



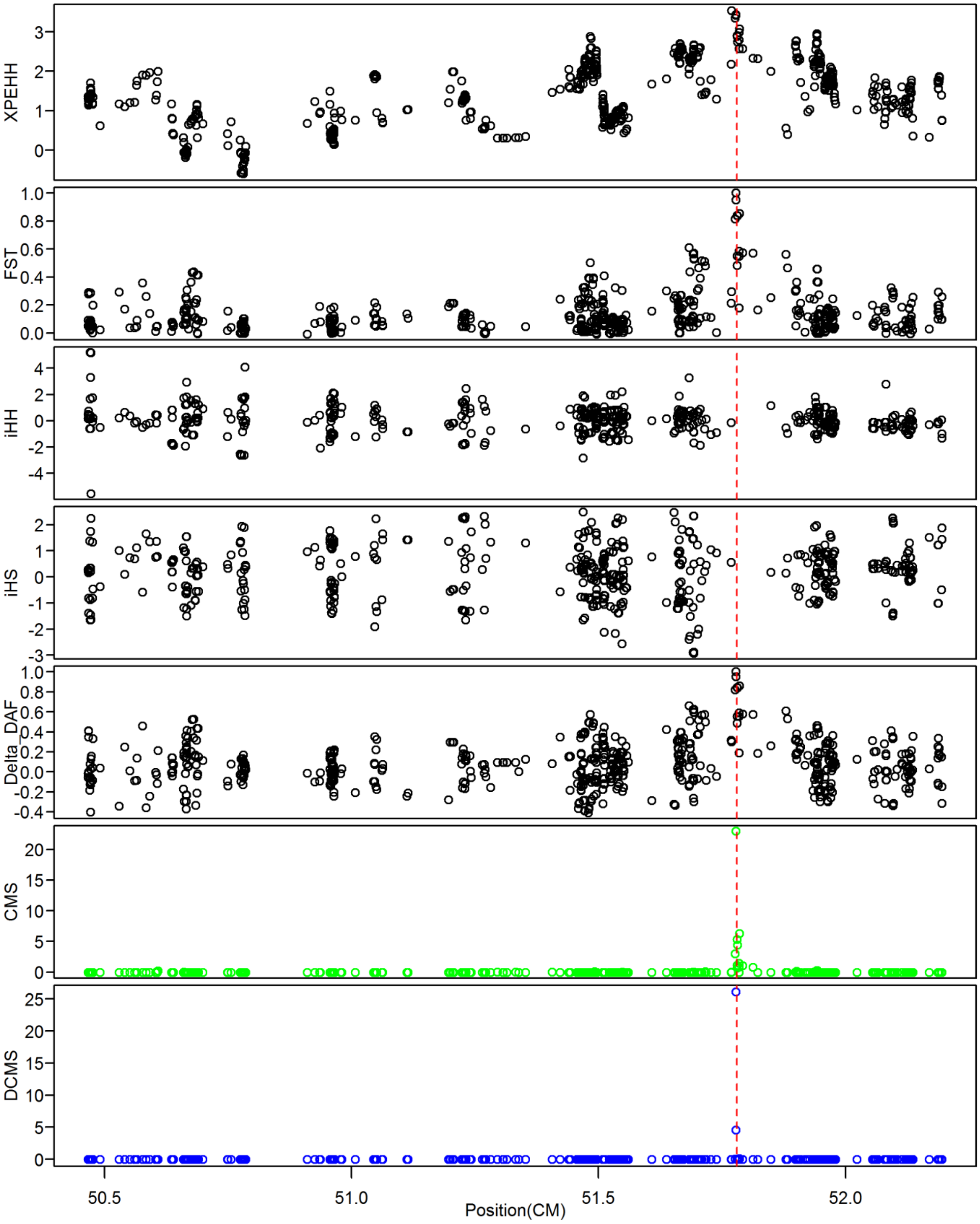
(Additional File 4)

Figure S4: **Power of eight different selection signature test statistics and the novel combining strategy when varying four different parameters: (A) Marker interval distance; (B) Frequency of the selected allele; (C) Sample Size; (D) Selection coefficient.** In this case, a selected population was used as reference population compared to another selected population in the between-population methods (Sel_1 vs. Sel_2).



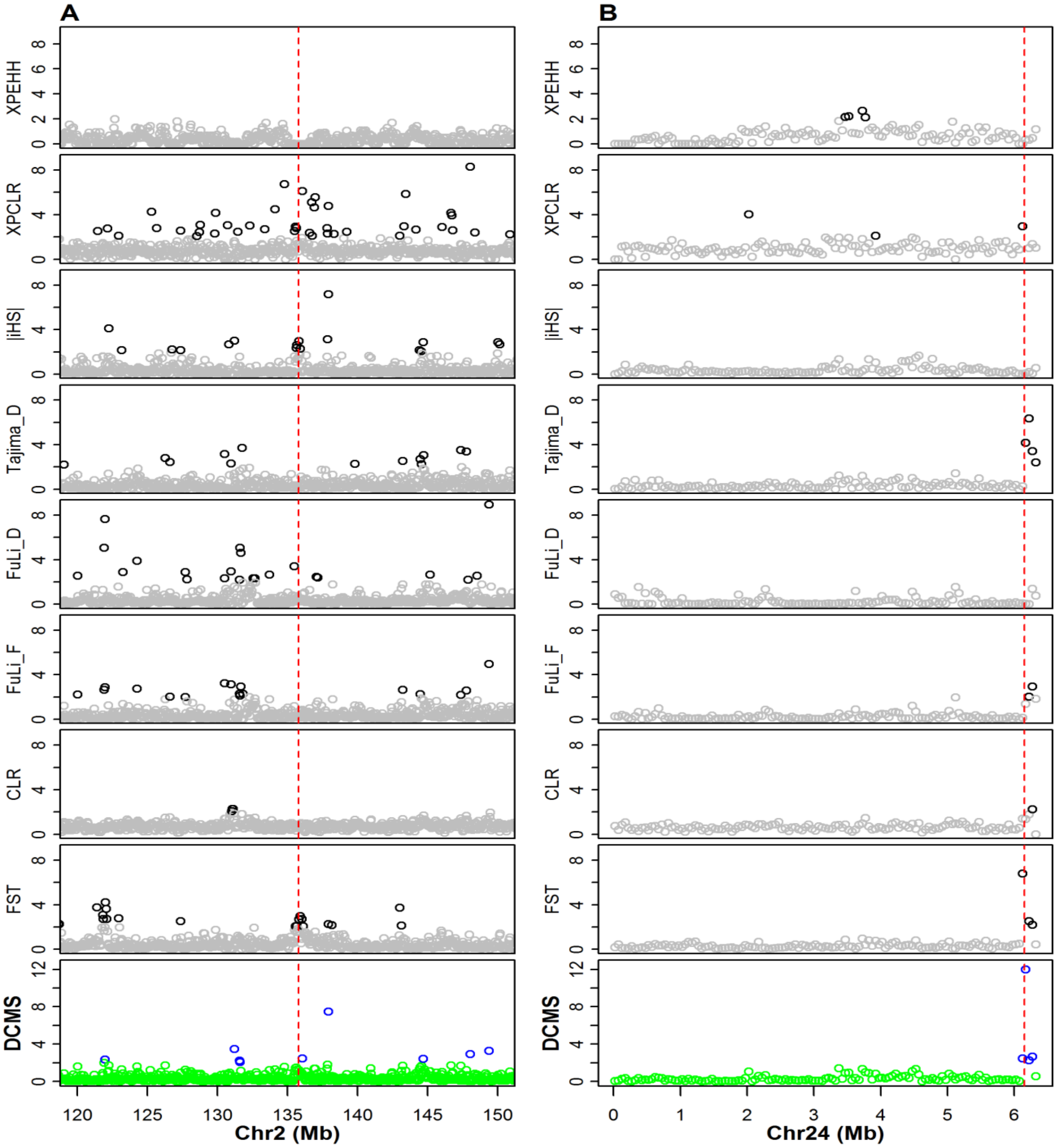
(Additional File 5)

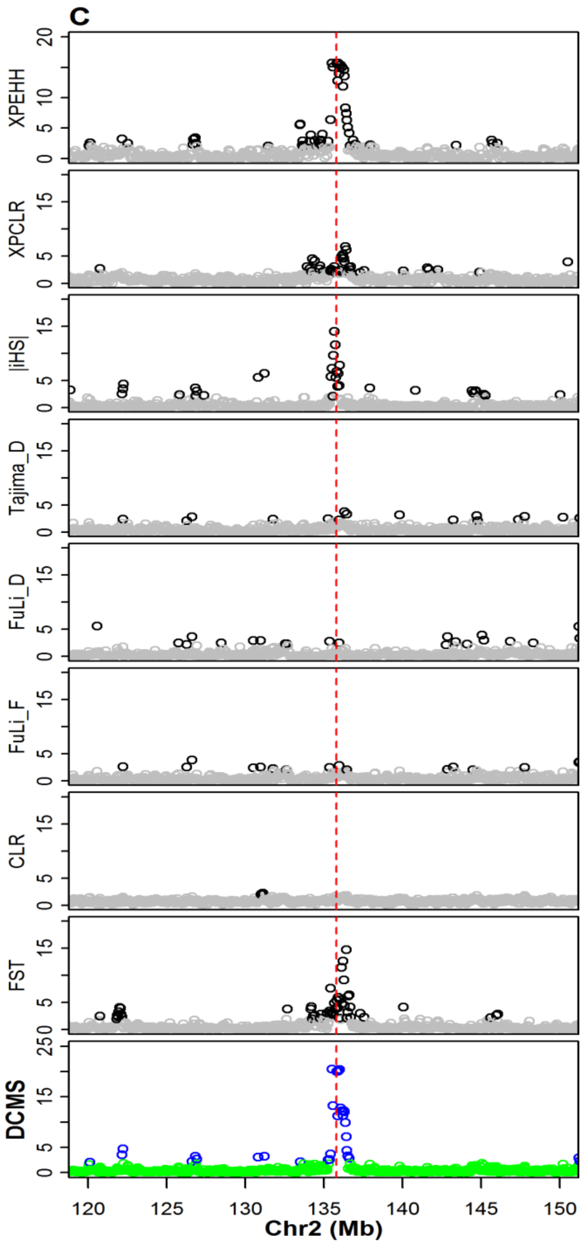
Figure S5: Localizing selection at MATP. Scores of five individual tests, CMS and DCMS for a region containing MATP.



(Additional File 6)

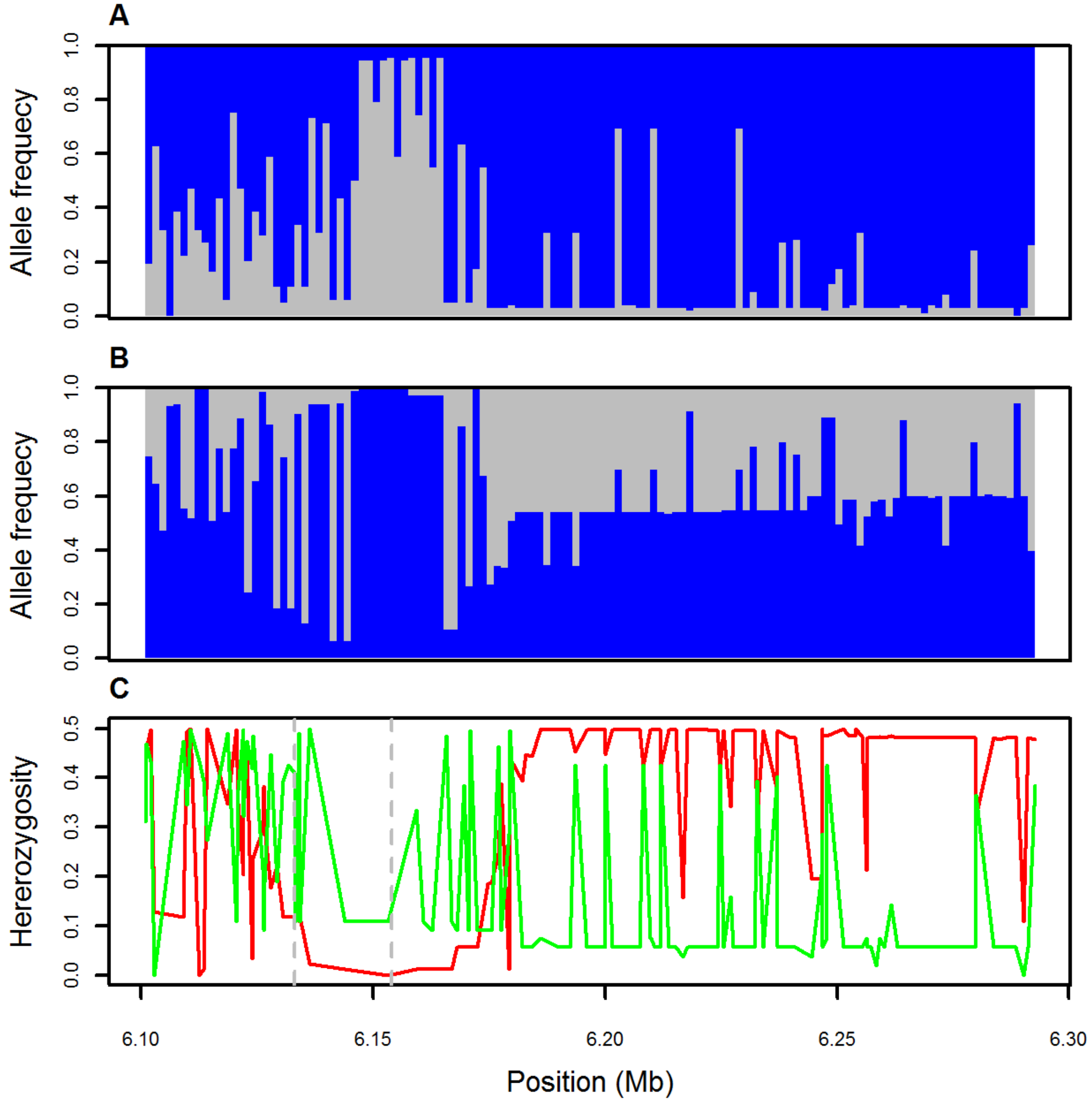
Figure S6: **Selection signature detected by DCMS in (A) Chromosome 2 in Human HapMap data in the analysis of the ASW population vs. the CEU population (B) Chromosome 24 in the comparison of white skin vs. yellow skin populations (C) Chromosome 2 in Human HapMap data in the analysis of the MKK population vs. the ASW population.** The Y axis reflects the –log (P-values). The red dashed line in (A,C) marks the location of the LCT gene in the human genome, and the red dashed line in (B) marks the location of the BCO2 gene in the chicken genome. The deep colored symbols represent the p-value of statistical scores for each statistic less than 1%.




C
XPEHH
XPCLR
|iHS|
Tajima_D
FuLi_D
FuLi_F
CLR
FST
DCMS
0
5
10
15
20
25
120
125
130
135
140
145
150
Chr2 (Mb)

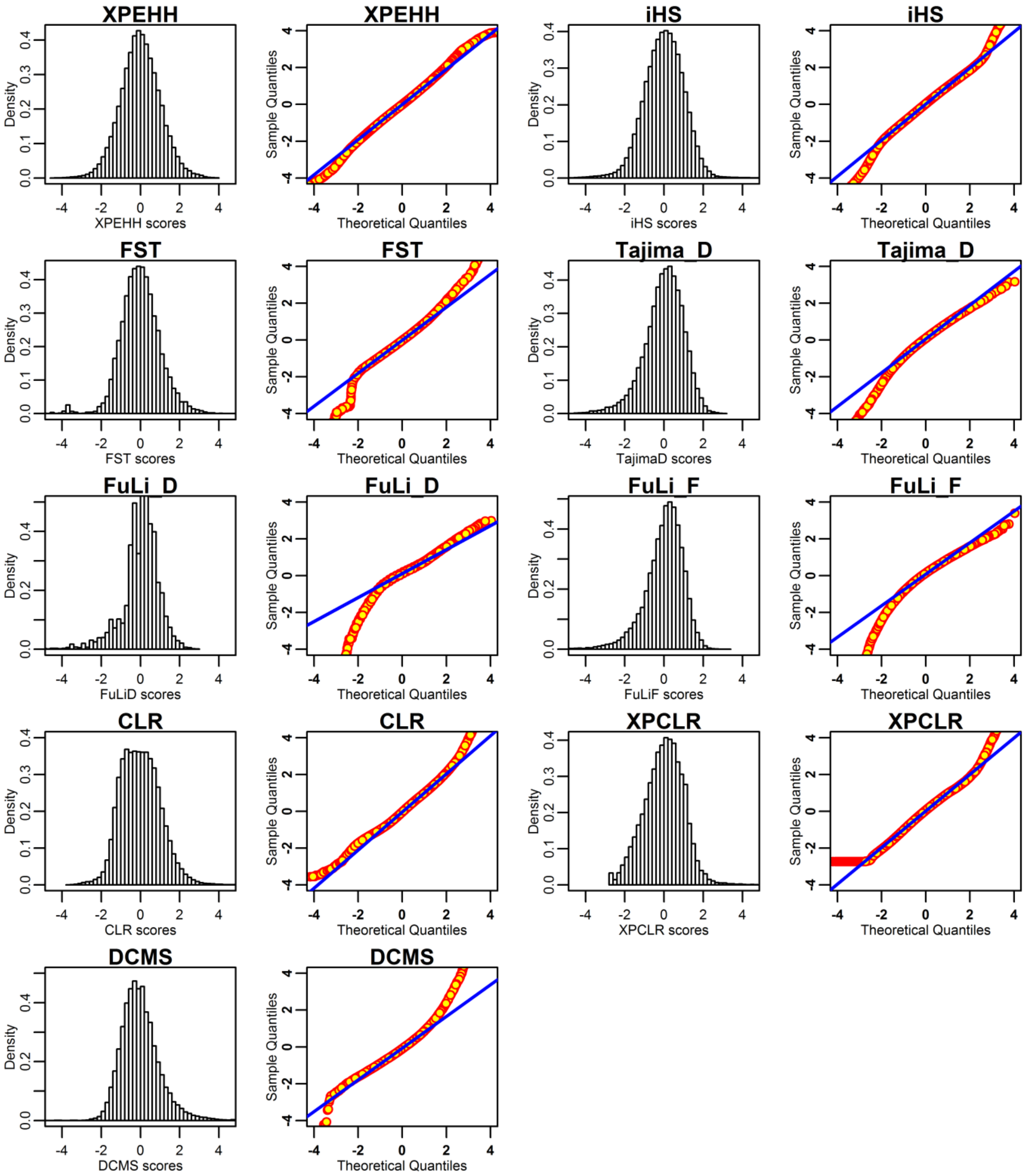
(Additional File 7)

Figure S7: The visualization of the Heterozygosity and Allele Frequency of two chicken populations. (A) The allele frequency in the region of 6.10–6.30 Mb on 24 Chromosome in the white skin population. (B) The allele frequency in the region of 6.10–6.30 Mb on 24 Chromosome in the yellow skin population. (C) Heterozygosity of the two populations. The red and green lines represent the yellow and white skin population, respectively.



(Additional File 8)

Figure S8: The histogram and the quantile-quantile (Q-Q) plots of statistical scores calculated by all methods in yellow skin populations. F_{ST} and XPCLR were normalized by sqrt transformation. CLR and DCMS were normalized by log transformation. Finally, all statistics were normalized by a z-transformation.



(Additional File 9)

Table S1: The power and false positive rate (FPR) in the scenario with maker interval d=62.5 kb, allele frequency p=0.8, selection coefficient s=0.02 and sample size N=50. The empirical significance threshold value was separately defined as 1 percent of the rank of all scores in all selection replicates for each method. Correspondingly, the false positive rate equaled to the power in neutral simulation scenario.

	XPEHH	**XPCLR**	**|iHS|**	**CLR**	**Tajima D**	**FuLi D**	**FuLi F**	**F_{ST}**	**DCMS**
power	0.12	0.21	0.16	0.19	0.08	0.08	0.11	0.28	0.16
FPR	0.44	0.93	0.91	0.40	0.83	0.73	0.82	0.81	0.56

(Additional File 10)

Table S2: The resolution of eight methods and the novel combining strategy.

	scenario	CLR	Tajima_D	XPEHH	iHS	XPCLR	F_{ST}	FuLi_D	FuLi_F	DCMS
Interval distance (kb)	62.5	-	-	0.056	0.125	0.025	0.085	-	-	_
	12.5	0.112	0.144	0.121	0.143	-	0.098	0.075	0.160	0.111
	2.5	0.093	0.061	0.125	0.138	-	0.069	0.103	0.093	0.118
	0.5	0.083	0.052	0.120	0.130	0.063	0.065	0.068	0.065	0.112
	0.1	0.079	0.075	0.113	0.130	0.097	0.069	0.095	0.094	0.108
Frequency	0.2	0.103	0.152	-	0.141	-	-	-	-	0.136
	0.4	0.129	-	0.101	0.117	-	-	-	-	0.116
	0.6	0.118	0.225	0.117	0.125	-	0.118	0.155	0.168	0.121
	0.8	0.093	0.061	0.125	0.138	-	0.069	0.103	0.093	0.119
	1.0	0.102	0.100	0.140	0.186	-	0.098	0.096	0.080	0.123
Sample size (chromosome)	10	0.117	0.107	0.125	0.145	0.201	0.106	0.107	0.119	0.118
	30	0.101	0.056	0.128	0.139	-	0.115	0.112	0.090	0.121
	50	0.084	0.156	0.126	0.136	-	0.078	0.122	0.115	0.120
	70	0.096	0.135	0.126	0.125	-	0.120	0.130	0.130	0.120
	90	0.084	0.053	0.119	0.116	-	0.072	0.090	0.090	0.113
Selection coefficient	0.005	0.074	0.051	0.082	0.082	-	-	0.056	0.056	0.086
	0.01	0.069	0.103	0.091	0.122	-	0.075	0.144	0.140	0.094
	0.02	0.089	0.075	0.126	0.143	0.125	0.117	0.048	0.043	0.121
	0.04	0.112	0.090	0.139	0.139	-	0.214	0.117	0.100	0.128
	0.08	0.129	0.112	0.143	0.133	-	0.142	0.133	0.141	0.127

Note: The scores represent the mean squared error of the estimated position in different scenarios, respectively. '-' suggested that the corresponding method has no power in the scenario.

(Additional File 11)

Table S3: Genome-wide DCMS scores.

(See Table S3.xlsx)

(Additional File 12)

Table S4: Summary of whole genome potential selection regions in chicken population (Mb).

	Yellow Skin Population		White Skin Population	
Chr.	Number of regions	Length	Number of regions	Length
1	228	11.40	192	9.60
2	214	10.70	211	10.55
3	140	7.00	138	6.90
4	85	4.25	129	6.45
5	86	4.30	50	2.50
6	41	2.05	38	1.90
7	22	1.10	42	2.10
8	22	1.10	25	1.25
9	17	0.85	22	1.10
10	12	0.60	9	0.45
11	20	1.00	12	0.60
12	13	0.65	15	0.75
13	5	0.25	8	0.40
14	8	0.40	5	0.25
15	11	0.55	10	0.50
17	6	0.30	20	1.00
18	13	0.65	10	0.50
19	9	0.45	11	0.55
20	8	0.40	5	0.25
21	5	0.25	1	0.05
22	15	0.75	24	1.20
23	8	0.40	4	0.20
24	8	0.40	7	0.35
25	3	0.15	4	0.20
26	2	0.10	8	0.40
27	10	0.50	6	0.30
28	2	0.10	7	0.35
Total	1013	50.65	1013	50.65

(Additional File 13)

Table S5: Candidate regions identified by the novel combining strategy analysis in two chicken populations.

(See Table S5.xlsx)

(Additional File 14)

Table S6: The absolute values of correlation coefficient of the eight statistical methods in the CEU population (upper triangular) and ASW population (lower triangular), respectively.

	XPEHH	**XPCLR**	**|iHS|**	**CLR**	**Tajima D**	**FuLi D**	**FuLi F**	**F_{ST}**
XPEHH		0.14	0.09	0.08	0.24	0.02	0.18	0.44
XPCLR	0.09		0.00	0.11	0.20	0.06	0.18	0.22
|iHS|	0.03	0.01		0.04	0.01	0.21	0.10	0.00
CLR	0.06	0.08	0.02		0.31	0.14	0.31	0.07
Tajima D	0.20	0.18	0.07	0.28		0.27	0.86	0.20
FuLi D	0.04	0.02	0.03	0.19	0.26		0.68	0.04
FuLi F	0.12	0.14	0.03	0.31	0.86	0.71		0.17
F_{ST}	0.10	0.01	0.10	0.03	0.07	0.01	0.05	

Note: the correlation coefficients were calculated using those statistics which deleted all loci located at the top 5% quantile in any of the employed statistics.