Supplementary Information

Figure S1. Illustration of capture uniformity for different multiplex levels. The logs (base 10) of the relative sequencing depth of each target locus were calculated, sorted and plotted. The capture uniformity varies within 2-4 orders of magnitude for 296-plex, 795-plex, 1,293-plex and 12,472-plex, with 98.1%, 97.4%, 97.8% and 95.1% of loci falling within in a 100-fold range, and 95.0%, 90.4%, 90.7% and 70.3% of loci falling within in a 10-fold range.

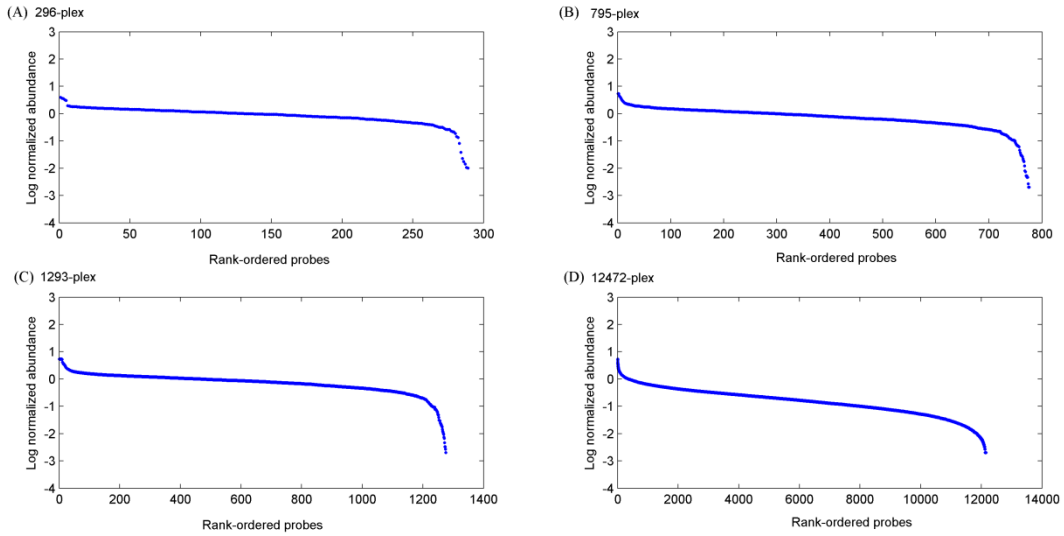


Figure S2. Dot-plots showing that locus sequencing depths (y-axis) are uncorrelated with the GC content of target regions (x-axis) for all multiplex levels.

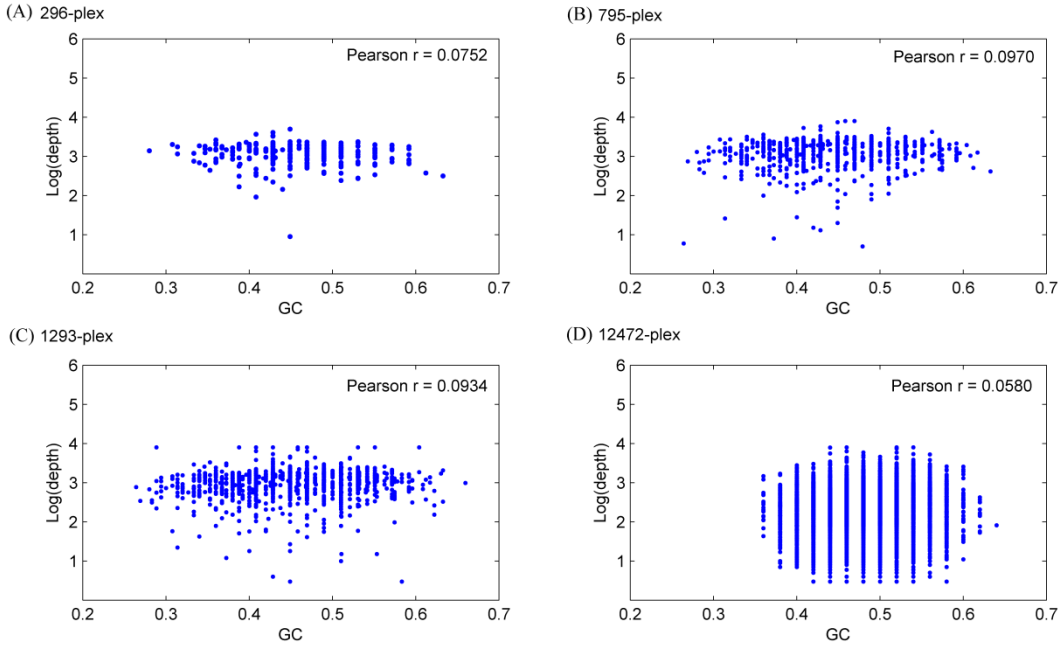


Figure S3. The distributions of sequencing depths for the exonic, intronic and UTR regions targeted by the 12472-plex, showing the largely comparable distributions among the three regions.

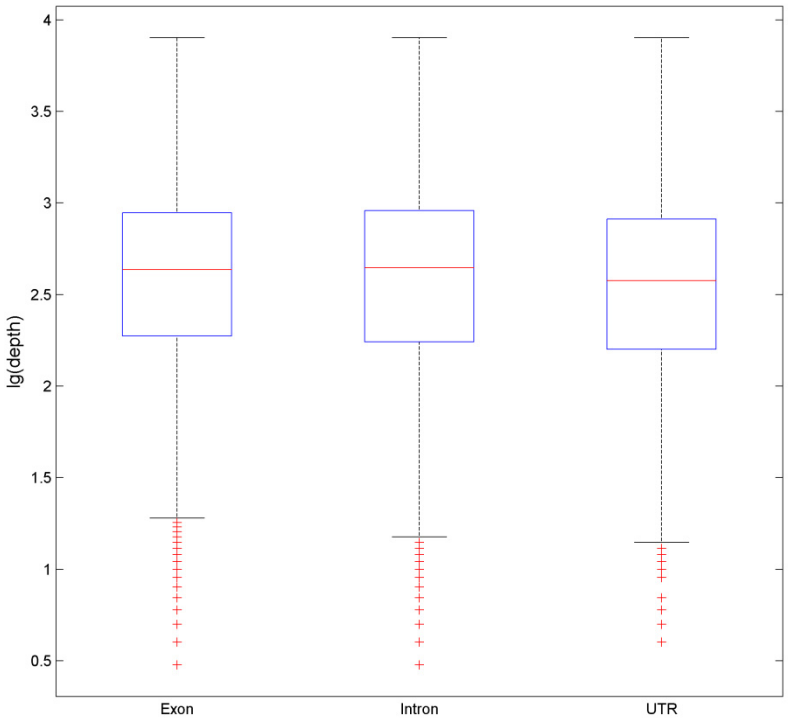


Figure S4. The correlation heatmap showing that the sequencing coverage of commonly detected loci is highly correlated across technical replicates (r= 0.97 to 0.99) and multiplex levels (r= 0.95 to 0.99).

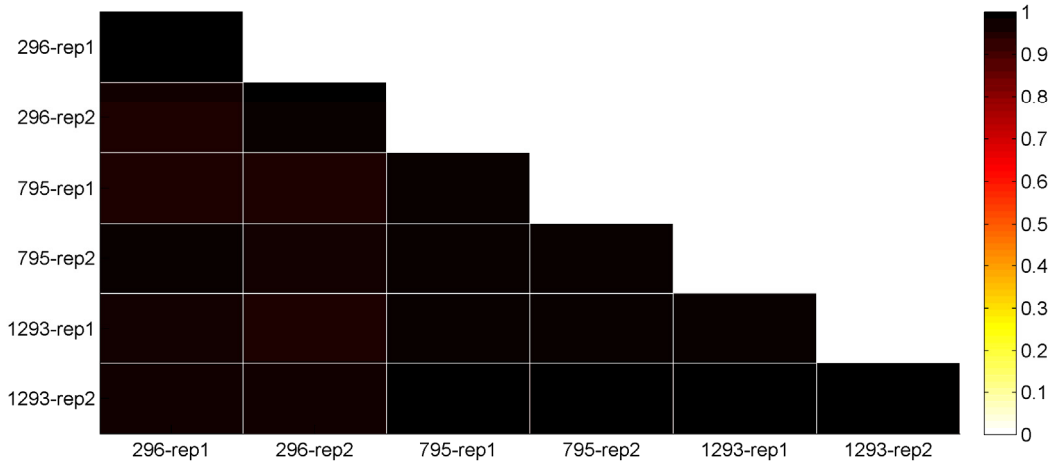


Figure S5. The dot-plot comparison of allelic sampling between HD-Marker and WGS. The performance of allelic sampling by HD-Marker is largely comparable to that of WGS, with 47.0% (HD-Marker) and 49.4% (WGS) of heterozygous loci falling within the allele frequency interval of 0.4-0.6 and 73.4% (HD-Marker) and 79.6% (WGS) of heterozygous loci falling within the allele frequency interval of 0.3-0.7. Note, for each locus, allele frequency was calculated for the allele that is present in the reference genome.

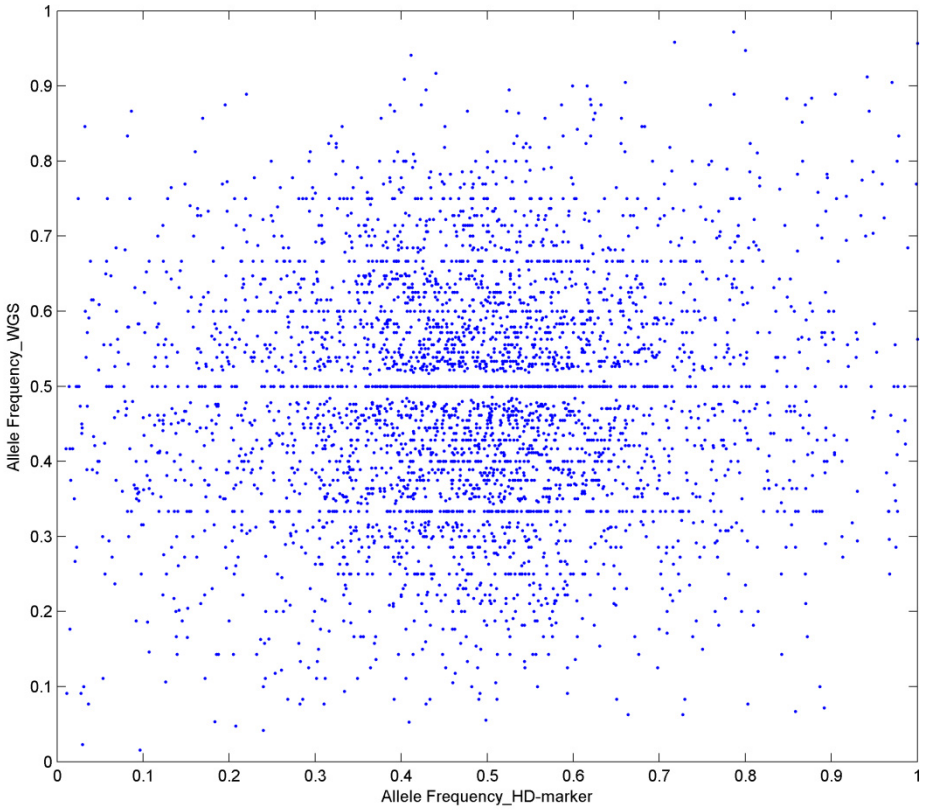


Figure S6. Three-dimensional dot-plot showing the distribution of allelic sampling within and across multiplex levels. Each dot represents a SNP locus and for a given locus, the frequency of minor allele is shown.

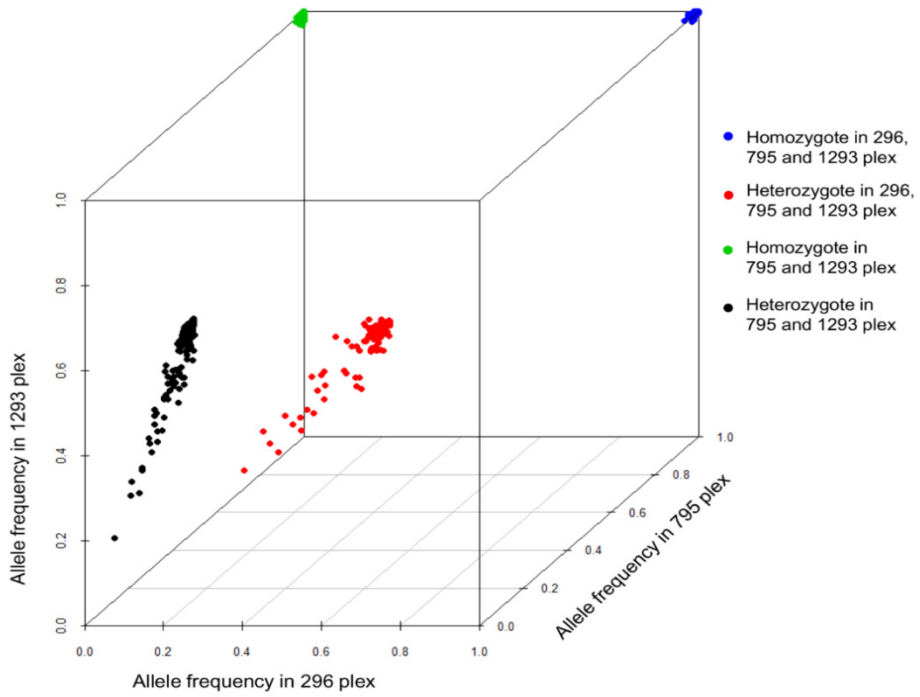
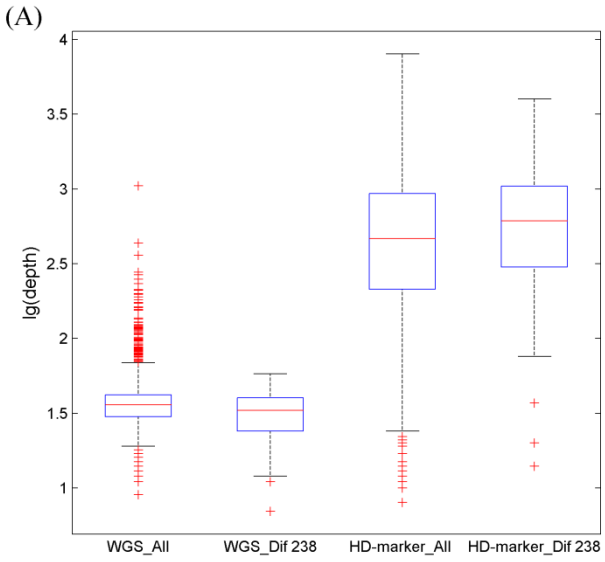


Figure S7. The sequencing depth distributions of the inconsistently genotyped loci that were genotyped as homozygotes by WGS but as heterozygotes by HD-Marker. (A) The overall sequencing coverage of the 238 inconsistently genotyped loci is comparable to the average of the whole dataset for WGS or HD-Marker; (B) For the 26 loci whose alleles detected by WGS correspond to the minor alleles detected by HD-Marker, the WGS sequencing coverage of these loci is notably lower than that for the whole dataset. It is very likely that the real genotypes of these 26 loci are heterozygotes, with only one allele captured by WGS due to the low sequencing coverage of these loci (i.e., biased allele sampling).



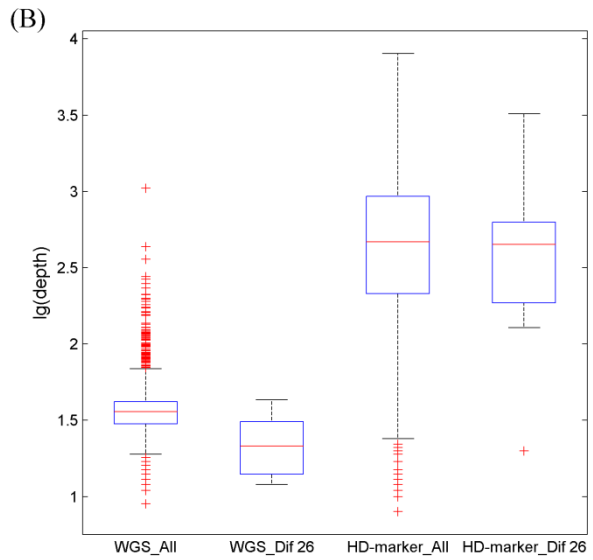


Figure S8. Evaluation of allelic sampling by HD-Marker for the inconsistently genotyped loci that were genotyped as homozygotes by WGS but as heterozygotes by HD-Marker. For a large majority (212) of the 238 inconsistently genotyped loci, the alleles detected by WGS are also major alleles detected by HD-Marker, and the two alleles detected by HD-Marker remarkably deviate from the expected 1:1 ratio. The additional minor alleles detected by HD-Marker possibly represent rare variants (presumably resulting from somatic mutations), and the inconsistent genotyping by the two methods could be related to their large difference in sequencing coverage (~35× for WGS and ~470× for HD-Marker) with differential power for detection of rare variants.

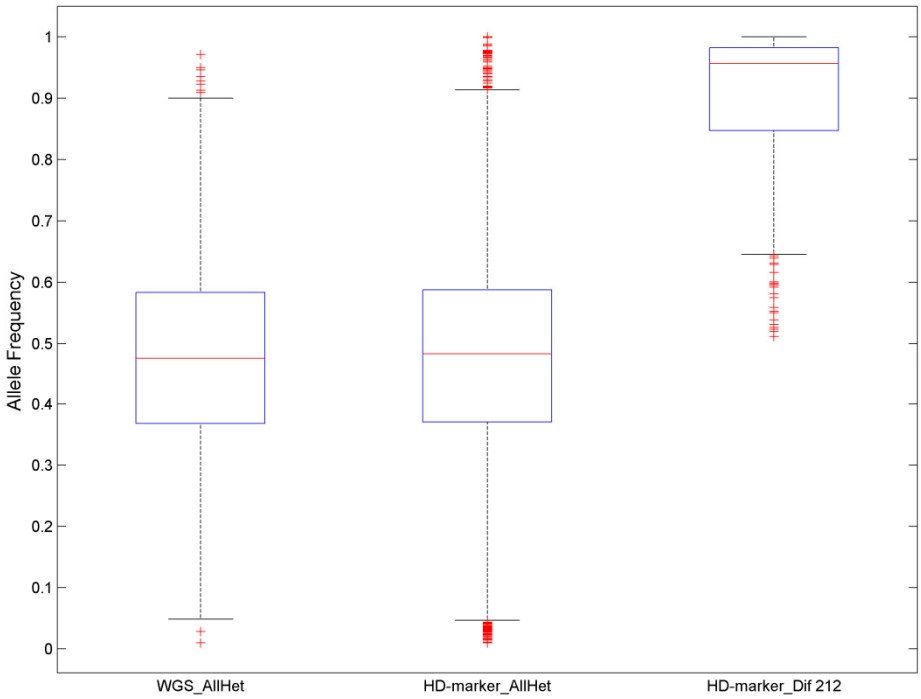


Table S3. Distribution of 12,472 target SNPs in the genic regions of *P. yessoensis*.

Genic regions	No. of target SNPs	Percentage (%)
Exon	7478	59.96
Intron	3454	27.69
5'_UTR	480	3.85
3'_UTR	1060	8.50
Total	12472	100.00

Table S4. Overview of the sequencing depth of the detected loci at four multiplex levels.

Depth	296-plex				795-plex				1293-plex				12472-plex			
	Rep1		Rep2		Rep1		Rep2		Rep1		Rep2		Rep1		Rep2	
	Detected loci	Percentage (%)	Detected loci	Percentage (%)	Detected loci	Percentage (%)	Detected loci	Percentage (%)	Detected loci	Percentage (%)	Detected loci	Percentage (%)	Detected loci	Percentage (%)	Detected loci	Percentage (%)
≥1	290	100.00	288	100	782	100	784	100	1279	100	1279	100	12245	100	12246	100
≥2	290	100	288	100	778	99.49	783	99.87	1277	99.84	1278	99.92	12229	99.87	12224	99.82
≥3	289	99.66	288	100	776	99.23	782	99.74	1275	99.69	1277	99.84	12209	99.71	12211	99.71
≥4	289	99.66	288	100	774	98.98	779	99.36	1274	99.61	1277	99.84	12203	99.66	12200	99.62
≥5	289	99.66	288	100	773	98.85	777	99.11	1273	99.53	1273	99.53	12195	99.59	12188	99.53
≥6	289	99.66	288	100	773	98.85	776	98.98	1271	99.37	1273	99.53	12182	99.49	12176	99.43
≥7	289	99.66	287	99.65	773	98.85	776	98.98	1271	99.37	1269	99.22	12167	99.36	12167	99.35
≥8	289	99.66	287	99.65	771	98.59	776	98.98	1270	99.30	1269	99.22	12152	99.24	12148	99.20
≥9	289	99.66	287	99.65	769	98.34	776	98.98	1270	99.30	1269	99.22	12139	99.13	12127	99.03
≥10	289	99.66	287	99.65	769	98.34	776	98.98	1270	99.30	1269	99.22	12125	99.02	12113	98.91
≥50	284	97.93	283	98.26	759	97.06	765	97.58	1252	97.89	1255	98.12	11513	94.02	11486	93.79

Table S5. Genotyping performance of common target SNPs across multiplex levels based on Illumina data.

	For 296 common loci					For 795 common loci			
	296-plex	795-plex	1,293-plex	Common (percentage[a])	Consistent (percentage[b])	795-plex	1,293-plex	Common (percentage[a])	Consistent (percentage[b])
Detected	288	286	288	286 (99.31%)	/	781	778	778 (99.62%)	/
Calling	283	279	279	273 (96.47%)	275 (100.00%)	757	743	741 (97.89%)	739 (99.73%)

[a] Percentage was calculated by dividing the number of loci that are commonly detected or called across multiplex levels by the number of loci that are detected or called in the lowest multiplex levels (296 or 795).

[b] Percentage was calculated by dividing the number of consistently-genotyped loci by the number of commonly-called loci across multiplex levels.

Table S6. Summary of allele frequency distribution at four multiplex levels.

		296-plex	795-plex	1293-plex	12472-plex
Heterozygote	Mean	0.440	0.437	0.427	0.348
	Median	0.472	0.474	0.472	0.386
	Std Dev	0.090	0.101	0.114	0.129
Homozygote	Mean	0.999	0.999	0.999	0.999
	Median	1.000	1.000	1.000	1.000
	Std Dev	0.001	0.001	0.003	0.005

Table S7. Summary of the locus detection and genotyping performance for five multi-gene families covered by the 12472-plex.

Gene families	No.of genes in genome	12472-plex		Replicate	Loci detection			Genotype calling			Concordance between replicates			Validation by resequencing		
		No. of Genes	Targeted SNPs		No. of locus	Rate (%)	Ave. rate (%)	No. of locus	Rate (%)	Ave. rate (%)	Common calling	Consistent genotyping	Consistent rate (%)	Common calling	Same genotype	Validation rate (%)
CYP450	93	54	54	Rep1	54	100%	99.1%	53	98.1%	99.1%	53	53	100%	53	53	100%
				Rep2	53	98.1%		53	100%							
Homeobox	119	64	64	Rep1	64	100%	100%	64	100%	100%	64	63	98%	60	59	98.3%
				Rep2	64	100%		64	100%							
ANK	52	32	32	Rep1	32	100%	100%	32	100%	100%	32	32	100%	32	31	96.9%
				Rep2	32	100%		32	100%							
HSP70	67	36	36	Rep1	35	97.2%	97.2%	35	100%	100%	35	35	100%	35	34	97.1%
				Rep2	35	97.2%		35	100%							
CHST	86	30	30	Rep1	30	100%	100%	30	100%	100%	30	30	100%	29	29	100%
				Rep2	30	100%		30	100%							

Abbreviations: CYP450, *Cytochrome P450*; *Homeobox*, *Homeobox protein*; *ANK*, *Ankyrin repeat domain-containing protein*; *Hsp70*, *Heat shock 70kDa protein*; *CHST*, *Carbohydrate sulfotransferase*

Table S8. SOLiD data processing and alignment to target regions.

Multiplex level	Technical replicate	Read processing				Aligned to target regions		
		Raw reads (M)	HQ reads (M)	Efficiency (%)	Ave. efficiency (%)	Aligned reads (M)	Efficiency* (%)	Ave. efficiency (%)
296	Rep1	0.70	0.48	68.57	72.57	0.35	72.92	75.23
	Rep2	0.64	0.49	76.56		0.38	77.55	
795	Rep1	2.57	1.70	66.15	70.99	1.26	74.12	75.80
	Rep2	2.40	1.82	75.83		1.41	77.47	
1,293	Rep1	6.31	3.77	59.75	65.44	2.71	71.88	73.60
	Rep2	5.47	3.89	71.12		2.93	75.32	

*Mapping efficiency was calculated by dividing the number of aligned reads by the total number of HQ reads.

Table S9. Summary of locus detection, genotype calling and concordance between replicates based on SOLiD data.

Multiplex level	Replicate	Loci detection			Genotype calling			Concordance between replicates		
		No. of locus	Rate (%)	Ave. rate (%)	No. of locus	Rate (%)	Ave. rate (%)	Common calling (RMSE/R-squared)	Consistent genotyping (RMSE/R-squared)	Consistent rate (%)
296	Rep1	286	96.62	96.62	251	87.76	91.09	246 (0.05/0.95)	245 (0.05/0.95)	99.59
	Rep2	286	96.62		270	94.41				
795	Rep1	769	96.73	97.11	690	89.73	91.19	676 (0.04/0.96)	670 (0.04/0.96)	99.11
	Rep2	775	97.48		718	92.65				
1,293	Rep1	1,263	97.53	98.00	1,112	88.04	89.43	1,090 (0.03/0.96)	1,077 (0.03/0.97)	98.81
	Rep2	1,275	98.46		1,158	90.82				

Table S10. SOLiD genotype validation by Sanger-based amplicon sequencing.

Sanger-based genotypes	HD-Marker SNP genotypes (SOLiD platform)								
	296-plex[a]			795-plex[a]			1,293-plex[a]		
	Same	Different	Validation rate (%)	Same	Different	Validation rate (%)	Same	Different	Validation rate (%)
Homozygote	69	0	100	84	2	97.67	98	2	98.00
Heterozygote	51	3	94.44	50	3	94.34	60	2	96.77
Total	120	3	97.56	134	5	96.40	142	4	97.53

[a], 123, 139 and 146 SNP loci from 296-plex, 795-plex and 1,293-plex respectively were subject to genotype validation by Sanger-based amplicon sequencing.

Table S11. Genotyping consistency between the Illumina and SOLiD platform for three multiplex levels.

Multiplex level	Commonly detected loci	Consistently genotyped loci (percentage)
296-plex	241	231 (95.85%)
795-plex	653	621 (95.10%)
1,293-plex	1,061	1,014 (95.57%)

Table S14. Microsatellite data processing and alignment to target regions.

Multiplex level	Technical replicate	Read processing				Aligned to target regions		
		Raw reads (M)	HQ reads (M)	Efficiency (%)	Ave. efficiency (%)	Aligned reads (M)	Efficiency* (%)	Ave. efficiency (%)
50	Rep1	0.28	0.26	91.87	93.18	0.25	98.29	97.77
	Rep2	0.28	0.27	94.50		0.26	97.24	

*Mapping efficiency was calculated by dividing the number of aligned reads by the total number of HQ reads.

Table S15. Statistics of microsatellite genotyping and validation by capillary electrophoresis analysis (CEA).

Multiplex level	Replicate	Loci detection			Genotype calling			Concordance between replicates			Validation by CEA		
		No. of locus	Rate (%)	Ave. rate (%)	No. of locus	Rate (%)	Ave rate (%)	Common calling	Consistent genotyping	Consistent rate (%)	Genotyped by CEA	Consistent genotyping	Consistent rate(%)
50	Rep1	48	96.00	92.00	39	81.25	83.81	31	28	90.32	10	10	100.00
	Rep2	44	88.00		38	86.36							

Table S16. Targeted genotyping of indels that are combined with three levels of SNP plexes.

Multiplex level	Replicate	Loci detection			Genotype calling			Concordance between replicates		
		No. of locus	Rate (%)	Ave. rate (%)	No. of locus	Rate (%)	Ave. rate (%)	Common calling	Consistent genotyping	Consistent rate (%)
15indels+296SNPplex	Rep1	14	93.33	93.33	13	92.86	92.86	13	13	100.00
	Rep2	14	93.33		13	92.86				
15indels+795SNPplex	Rep1	15	100.00	100.00	14	93.33	93.33	14	14	100.00
	Rep2	15	100.00		14	93.33				
15indels+1293SNPplex	Rep1	14	93.33	93.00	13	92.86	92.86	13	13	100.00
	Rep2	14	93.33		13	92.86				

Table S18. Summary of probe structures and primer/adaptor sequences used in the HD-Marker library preparation.

	Sequence (5'-3')	Notes
Probes		
Column-synthesized probe (LSP1)	CTCTCTATGGGCAGTCGGTGATXXXXXXXXXXXXXXXXXXXXX	Xs refer to locus-specific seq
Column-synthesized probe (LSP2)	YYYYYYYYYYYYYYYYYYYYYTTCGCCTTGGCCGTACAGCAG	Ys refer to locus-specific seq
Array-synthesized probe	AGGACCGGATCAACTYYYYYYYYYYYYYYYYYYYYYYAGATCGGAAGAGCAC ACGTCTGAGTCTCGCCTACACGACGCTCTTCCGATCTXXXXXXXXXXXXXXXXX XXXXXXCATTGCGTGAACCGA	
Probe amplification		
Oligo_F	TGCCTAGGACCGGATCAACT	
Oligo_R	GAGCTTCGGTTCACGCAATG	5' biotin labeling
gDNA labeling for microsatellites		
MseI-Adaptor-L	GACGATGAGTCCTGAG	
MseI-Adaptor-S	TACTCAGGACTCAT	3' Amino C6 blocking
EcoRI-Adaptor-L	CTCGTAGACTGCGTACC	
EcoRI-Adaptor-S	AATTGGTACGCAGTCTAC	3' Amino C6 blocking
MseI Primer	GATGAGTCCTGAGTAA	5' Biotin labeling
EcoRI Primer	GACTGCGTACCAATTC	5' Biotin labeling
Library amplification for 296, 795, 1293-plex		
1st-UP1	CTCTCTATGGGCAGTCGGTGAT	
1st-UP2	CTGCTGTACGGCCAAGGCGAA	
SLD-2nd-Primer (SOLiD)	CCACTACGCCTCCGCTTTCCTCTCTATGGGCAGTCGGTGAT	
SLD-2nd-Barcode (SOLiD)	CTGCCCCGGGTTCCTCATTCTCTNNNNNNNNNNCTGCTGTACGGCCAAGGCG	10nt barcode is highlighted
Slx-2nd-Primer (Illumina)	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGA TCTCTCTCTATGGGCAGTCGGTGAT	
Slx-2nd-Barcode (Illumina)	CAAGCAGAAGACGGCATACGAGATNNNNNNGTGACTGGAGTTCAGACGTGTGC TCTTCCGATCTCTGCTGTACGGCCAAGGCGAA	6nt barcode is highlighted
Library amplification for 12472-plex		
1st-UP1	ACACTCTTTCCCTACACGACGCTCTT	
1st-UP2	GTGACTGGAGTTCAGACGTGTGCTCTT	
Slx-2nd-Primer (Illumina)	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCT	
Slx-2nd-Barcode (Illumina)	CAAGCAGAAGACGGCATACGAGATNNNNNNGTGACTGGAGTTCAGACGTGT	6nt barcode is highlighted

Table S20. PCR primer sequences used for microsatellite genotype validation.

Microsatellite ID	Forward primer (5'-3')*	Reverse primer (5'-3')
CFXY007	CACGACGTTGTAAAACGACTTCTCTGAGGATACGGGAGC	GTTTATTGTACGCGGCGTC
CFDD098	CACGACGTTGTAAAACGACGCTGAGAAAGACATTCGTTAG	AGTCGTGTCTATCGTCTGC
CFIT07070	CACGACGTTGTAAAACGACCCACACATATAAGTAAGGTGAA	AACTTGGATTATATTGATGACC
CFJD002	CACGACGTTGTAAAACGACGAGCTGCCTTCATCAGGTG	CTTTACATCGTGGACATCATC
CFIT08826	CACGACGTTGTAAAACGACGATGTAGACACTGCACCAGAAG	CAGTACACTACAACTACATCTA
CFDD078	CACGACGTTGTAAAACGACACTCTGGTGCATAACCAATTC	TGGTAGGAGACCGTTTCAAG
CFAD157	CACGACGTTGTAAAACGACATATGATCTGGGGAGAAGTTGC	CACCATCTTTCCATCCCATAC
CFIT07492	CACGACGTTGTAAAACGACATGTGACCTGCACGGAATATCC	TACCCTGTATGGCCCACGGATA
CFBD193	CACGACGTTGTAAAACGACGGATGTAACCGATAAATATGGC	TGAAGGCGATGTTGGTGCTC
CFE24	CACGACGTTGTAAAACGACGTGCTTCTCAGTAGTATAGGG	GGTATGCAGAATAGGACCG

*, The universal portion of forward primers corresponds to the binding region of M13 primer (5'HEX-CACGACGTTGTAAAACGAC).