**Penalized Classification for Optimal Statistical Selection of markers from high-throughput genotyping: Application in Sheep Breeds**

G. Sottile 1, M.T. Sardina 2, S. Mastrangelo 2, R. Di Gerlando 2, M. Tolone 2, M. Chiodi 1 and B. Portolano 2

1 Dipartimento Scienze Economiche, Aziendali e Statistiche, University of Palermo, Palermo, Italy.

2 Dipartimento Scienze Agrarie e Forestali, University of Palermo, Palermo, Italy

Corresponding author: Gianluca Sottile. Email: gianluca.sottile@unipa.it

Short title: Penalized Classification for Optimal Selection

**Table S1**: *List of the 48 final selected SNPs in the different breeds. Information for each SNP was present in dbSNP database (*[*http://www.ncbi.nlm.nih.gov/snp*](http://www.ncbi.nlm.nih.gov/snp)*). OAR=Ovis aries chromosome number; NULL= no available rs.*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **SNP name** | **OAR** | **Position (bp)** | | **rs** | **Alleles** | **Gene Symbol** | | **Region** |
| OAR1\_174856446.1 | 1 | 161912812 | rs400156588 | | A/G | |  | Intergenic |
| OAR1\_177088394.1 | 1 | 163956292 | rs426113629 | | G/T | |  | Intergenic |
| s50912.1 | 1 | 196095847 | rs405031407 | | G/T | | TPRG1 | Intron |
| s24380.1 | 2 | 52042919 | rs419504162 | | A/G | |  | Intergenic |
| s20468.1 | 2 | 52463944 | rs426272889 | | C/T | | FAM221B | Intron |
| OAR2\_56828090.1 | 2 | 53054335 | rs412458465 | | A/G | | UNC13B | Intron |
| OAR2\_60563723.1 | 2 | 56331862 | rs412994477 | | A/G | |  | Intergenic |
| OAR2\_83949840.1 | 2 | 78848917 | rs411630259 | | A/G | |  | Intergenic |
| OAR2\_87942973.1 | 2 | 82769620 | rs406304097 | | A/G | | ZDHHC21 | Intron |
| s23687.1 | 2 | 83342823 | rs406393355 | | C/T | |  | Intergenic |
| OAR2\_89592915.1 | 2 | 84307737 | rs418811505 | | A/G | |  | Intergenic |
| OAR2\_89949804.1 | 2 | 84669442 | rs425292738 | | C/T | | BNC2 | Intron |
| OAR2\_91422909.1 | 2 | 85985270 | rs423495572 | | A/G | | ADAMTSL1 | Intron |
| OAR2\_102309617.1 | 2 | 95163145 | rs411085381 | | C/T | | MOB3B | Intron |
| OAR2\_208630788.1 | 2 | 197047501 | rs411588608 | | C/T | | DNAH7 | Intron |
| OAR3\_12358231.1 | 3 | 11714427 | rs421150057 | | A/G | |  | Intergenic |
| OAR3\_51035614.1 | 3 | 47779836 | rs420911481 | | A/G | |  | Intergenic |
| s32906.1 | 3 | 108054039 | rs410037867 | | C/T | |  | Intergenic |
| OAR3\_127397300.1 | 3 | 119370074 | rs417995525 | | A/G | |  | Intergenic |
| OAR3\_130762263.1 | 3 | 122515163 | rs410833230 | | C/T | |  | Intergenic |
| OAR3\_135788851.1 | 3 | 127150812 | rs428614821 | | A/G | |  | Intergenic |
| OAR3\_156245923.1 | 3 | 146260228 | rs416557154 | | C/T | | MUC19 | Intron |
| OAR3\_156253529.1 | 3 | 146267771 | rs427947538 | | C/G | | MUC19 | Exon (missense) |
| s69653.1 | 3 | 153976304 | rs414901427 | | C/T | |  | Intergenic |
| OAR5\_6540581.1 | 5 | 6225640 | rs410850715 | | G/T | | CALR3 | Intron |
| OAR5\_69570106.1 | 5 | 63221217 | rs404844842 | | C/T | |  | Intergenic |
| s50174.1 | 5 | 90032387 | rs419585475 | | C/T | |  | Intergenic |
| OAR6\_40409402.1 | 6 | 36121839 | rs409468158 | | C/T | | PYURF | Intron |
| OAR6\_40724811\_X.1 | 6 | NULL | NULL | | NULL | | NULL | NULL |
| s19002.1 | 6 | 69634266 | rs405729829 | | A/G | |  | Intergenic |
| OAR9\_33155950.1 | 9 | 31548651 | rs419118491 | | A/G | |  | Intergenic |
| s71002.1 | 9 | 33327765 | rs428521816 | | C/T | | SNTG1 | Intron |
| OAR10\_1797787.1 | 10 | 3966297 | rs415988084 | | A/G | |  | Intergenic |
| OAR10\_29159858.1 | 10 | 29101583 | rs410054529 | | A/G | | FRY | Intron |
| OAR10\_29448537.1 | 10 | 29392142 | rs419203432 | | C/T | | LOC106991357 | Intron |
| OAR11\_40842530.1 | 11 | 38321708 | rs403627121 | | C/T | | LOC105616364 | Downstream |
| OAR12\_33879358.1 | 12 | 30053741 | rs417153016 | | C/G | | LOC101106609 | Intron |
| DU257726\_402.1 | 12 | 31160400 | rs55628371 | | A/G | |  | Intergenic |
| OAR12\_76492821.1 | 12 | 69546949 | rs404226659 | | C/T | | INTS7 | Intron |
| s11337.1 | 14 | 1433351 | rs409869687 | | A/G | |  | Intergenic |
| OAR14\_25284939.1 | 14 | 24378737 | rs408203843 | | A/G | | NLRC5 | Intron |
| OAR18\_13826286.1 | 18 | 13915292 | rs404159484 | | A/G | |  | Intergenic |
| OAR18\_65042450.1 | 18 | 60956686 | rs420671587 | | C/T | |  | Intergenic |
| OAR19\_33278780.1 | 19 | 31614145 | rs419333175 | | A/G | | MITF | Intron |
| OAR19\_33355170.1 | 19 | 31689556 | rs408878307 | | A/T | | MITF | Intron |
| OAR19\_33417302.1 | 19 | 31752386 | rs420681217 | | C/T | | MITF | Intron |
| s34077.1 | 19 | 41587674 | rs416892089 | | A/G | | FHIT | Intron |
| OAR25\_10278164.1 | 25 | 10202323 | rs414463797 | | A/G | | RYR2 | Intron |



**Figure S1** Plot of the first three principal components obtained using the panel of 48 068 Single Nucleotide Polymorphisms (coded as genotype). • = Valle del Belice (VdB), • = Comisana (Com), • = Pinzirita (Pin), • = Barbaresca (Bar), • = Sarda (Sar).



Figure S2 Cluster plot of the first two principal components obtained using k-means technique. Grey asterisks are the centroids for each cluster and the cross highlights the missclassified unit.