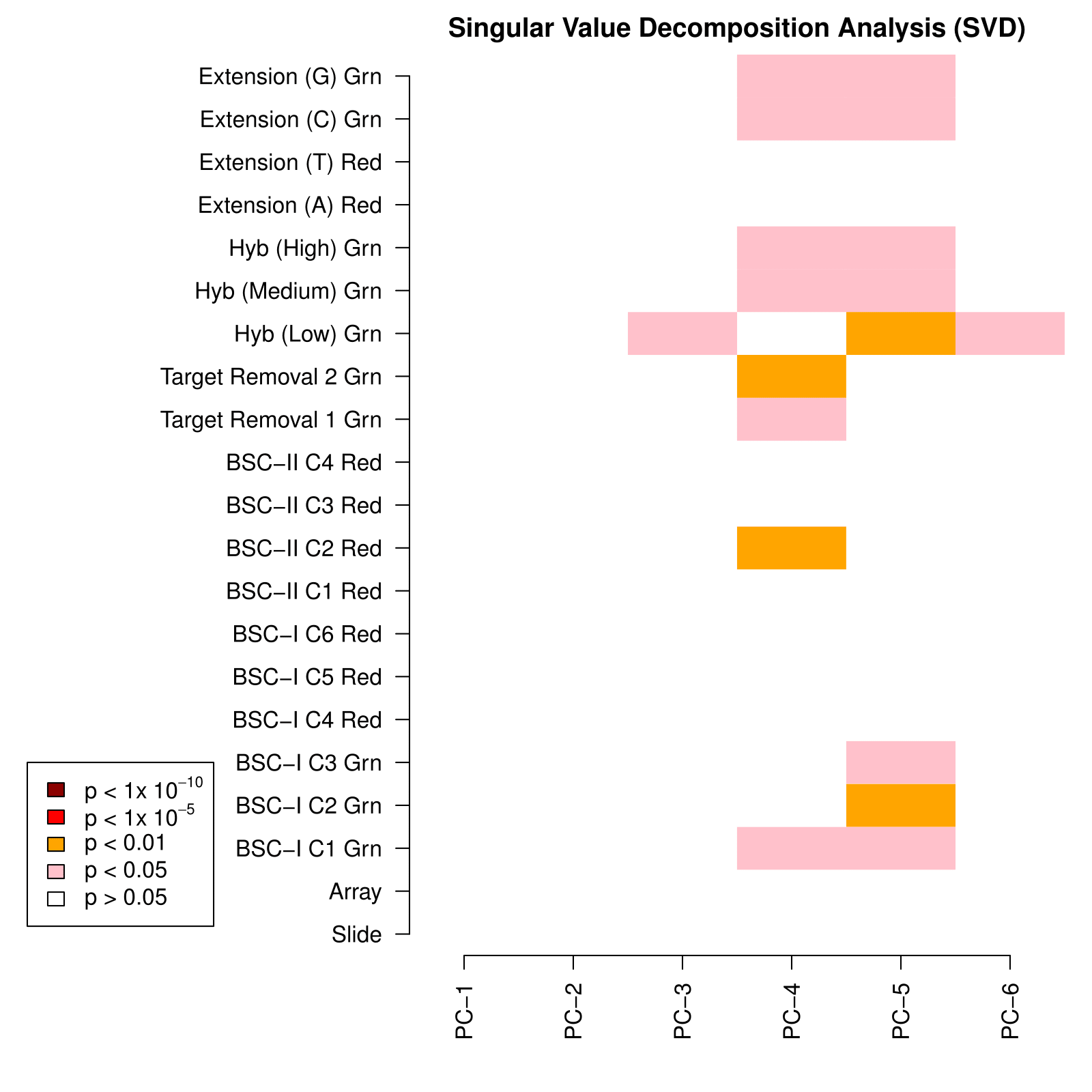
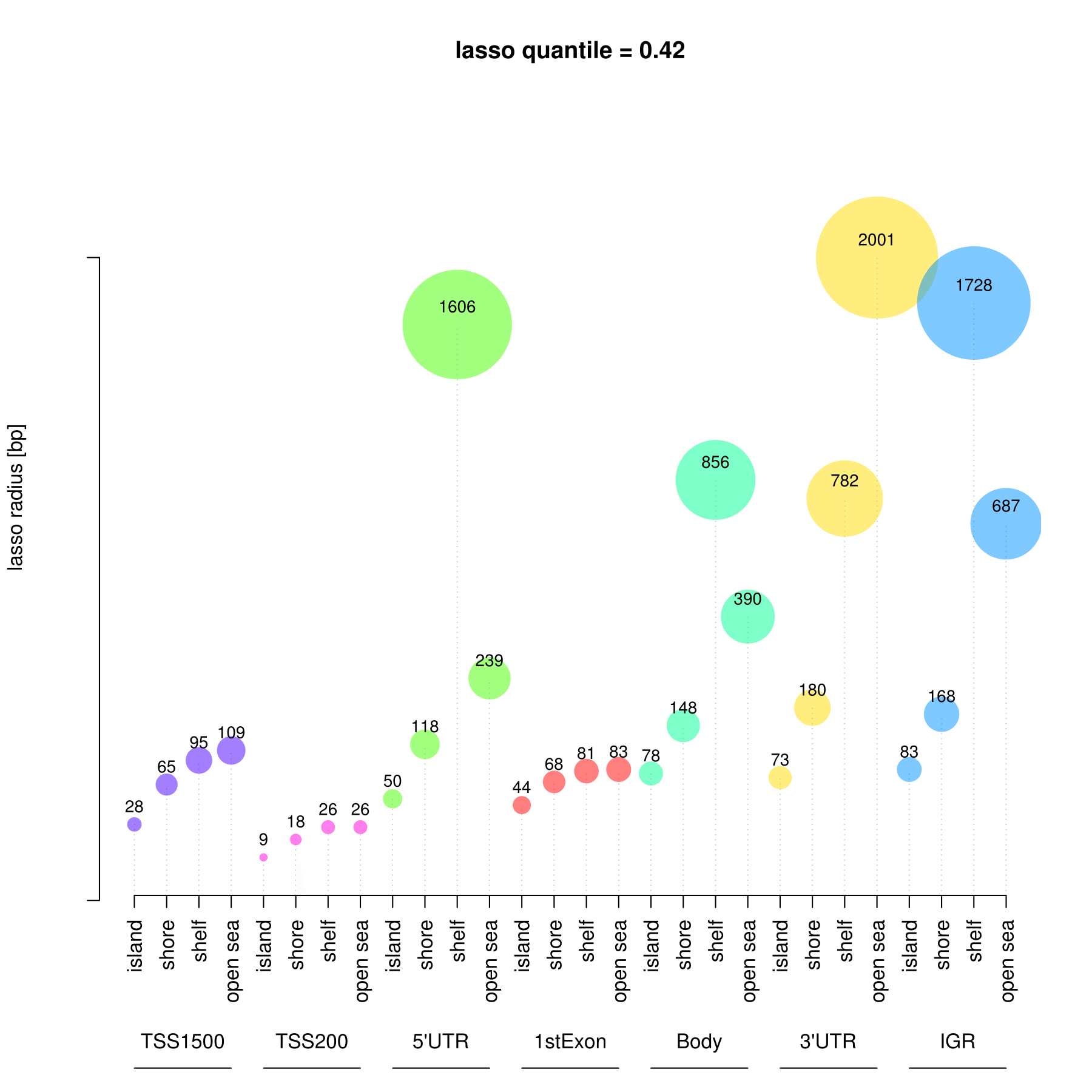
# Supplementary Material

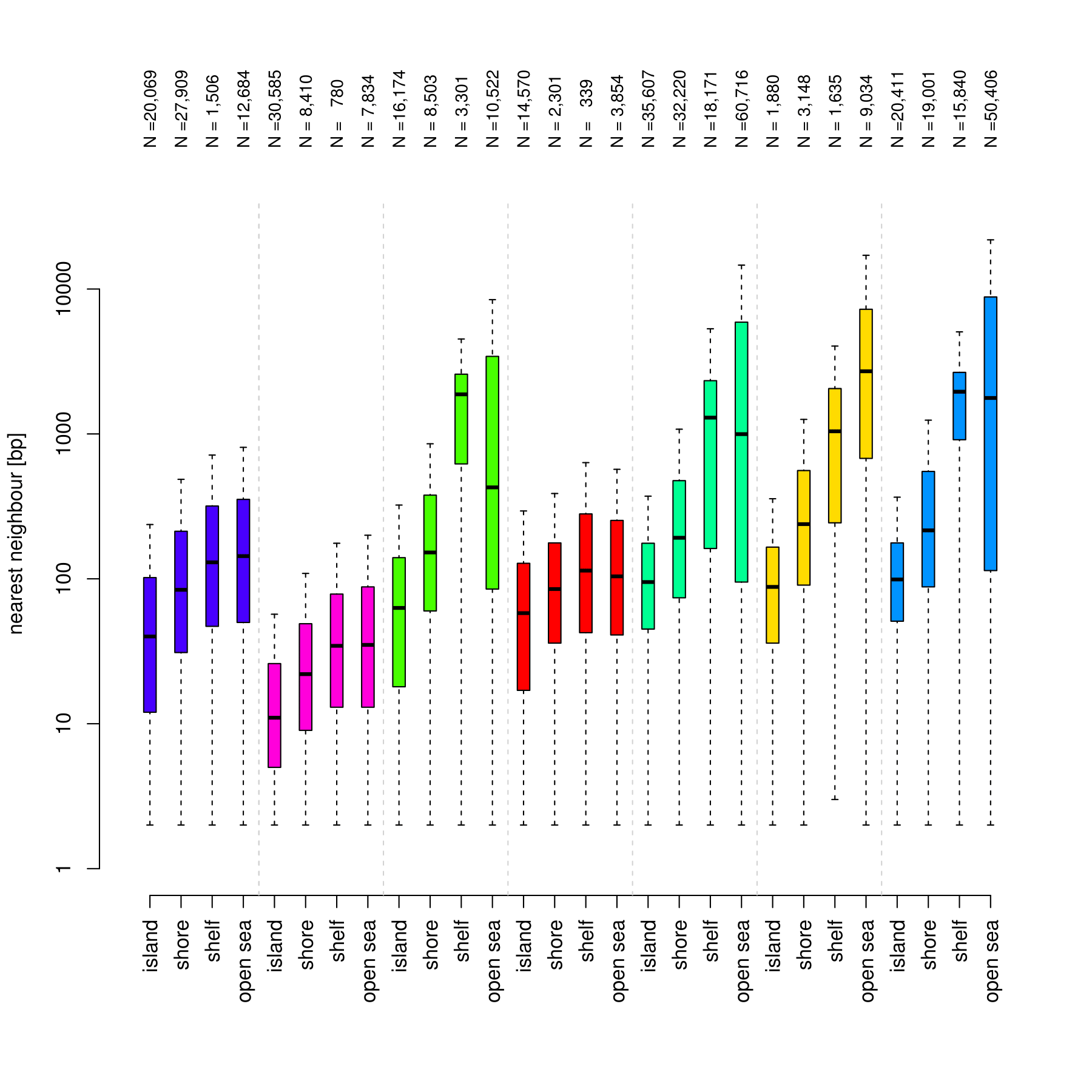
**Supplementary Figure 1:** Singular value decomposition (SVD). An example output from implementing champ.SVD to reveal confounding variables within the matrix of beta values. This figure illustrates that there were negligible technical confounders associated with this subset of TCGA COAD samples.



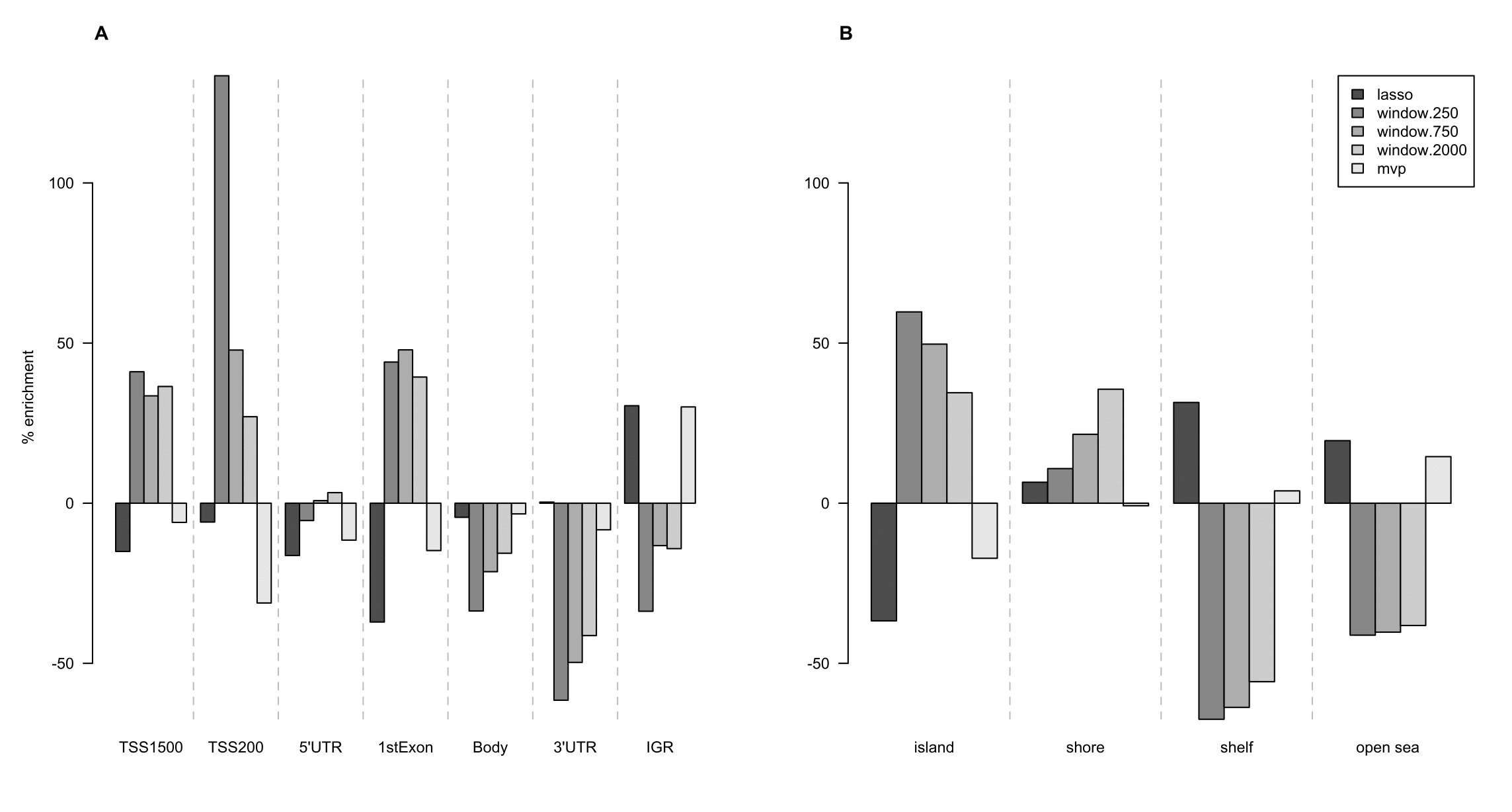
**Supplementary Figure 2:** Example output figure from Probe Lasso implemented in R using ChAMP illustrating the sizes of probe-lassos centred on probes in the various genetic/epigenetic features.

****

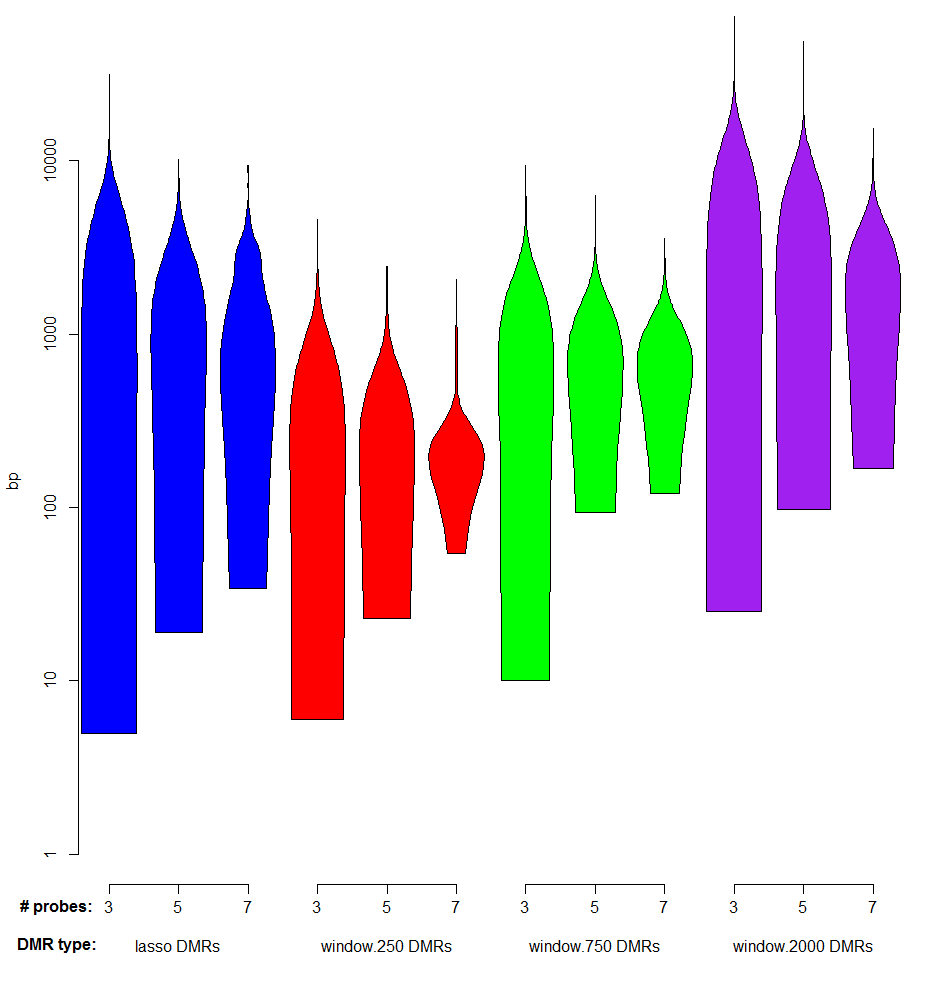
**Supplementary Figure 3**: Example output figure from Probe Lasso implemented in R using ChAMP illustrating the range of probe-spacing in the underlying dataset.

****

**Supplementary Figure 4**: Enrichment of gene relations (A) and CGI relations (B) for the lasso DMRs (dark grey) and window250 DMRs (mid greys) using stringency set at >= 3 significant probes captured. Enrichment is also presented using the distribution of significant MVPs only (light grey).

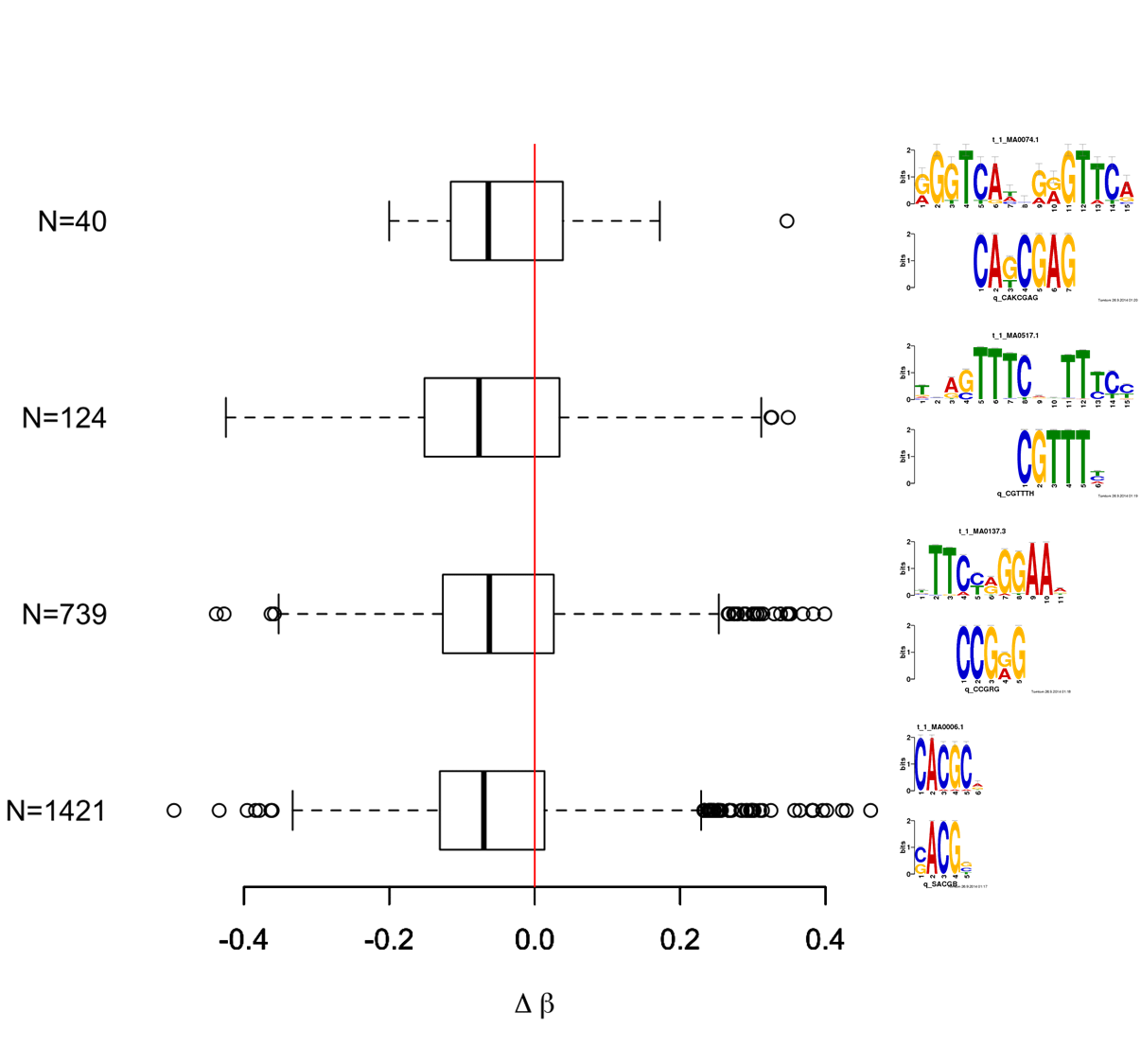


**Supplementary Figure 5:** Distributions of the sizes of DMR cores (space between first and last probes in a DMR) that result from the Probe Lasso (blue) and Sliding fixed window250 (red), window750 (green) and window2000 approaches. The results are split by stringency, with the effect being a gradual reduction in DMR size.

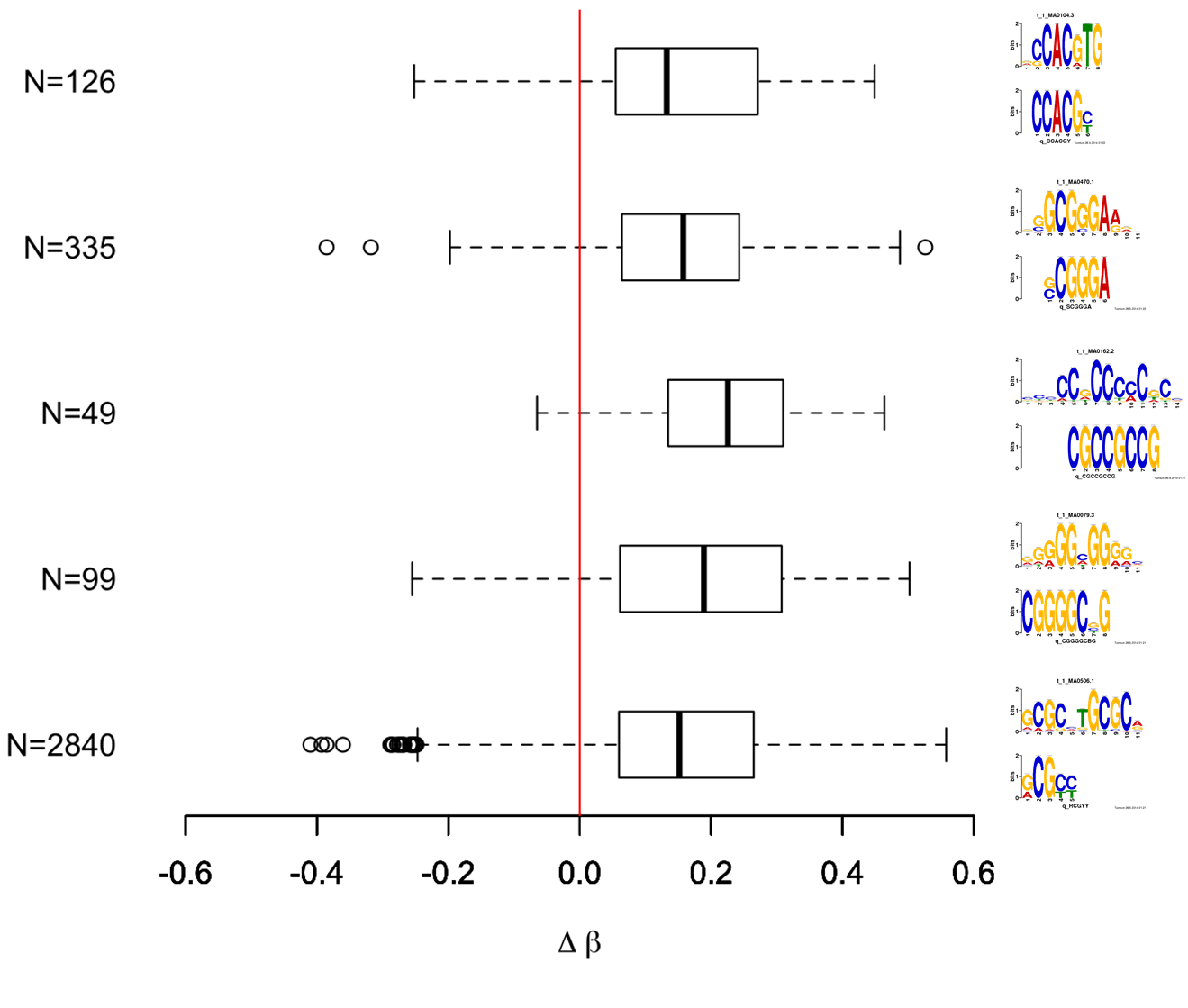


**Supplementary Figure 6:** Box plots to indicate the delta-beta between TCGA colorectal samples and healthy control colon for significantly associated probes that overlap with motifs enriched in unique DMRs. Negative delta beta means hypomethylation in colon cancer and vice-versa. Probes at enriched motifs in unique Probe Lasso DMRs were substantially hypomethylated (A) while probes at enriched motifs in unique window250 DMRs were considerably hypermethylated (B). The red vertical line indicates no methylation difference between cohorts.

**(A)**

****

**(B)**

****

**Supplementary Table 1:** Identification of putatively DNA binding proteins associated with probes in DMRs unique to either Probe Lasso or window250 algorithms. Superscripts refer to the motif presented in Supplementary Figure 6 (1-5 = bottom-top).

|  |  |  |
| --- | --- | --- |
| **motif match** | **Probe Lasso** | **Window250** |
| Arnt | yes1 | yes5 |
| Arnt::Ahr | yes1 | yes5 |
| CTCF |  | yes1 |
| E2F1 |  | yes2,4 |
| E2F3 |  | yes2,3,4 |
| E2F4 |  | yes2,4 |
| E2F6 |  | yes2,4 |
| EGR1 |  | yes2,3,5 |
| EGR2 | yes1 | yes5 |
| EHF | yes4 |  |
| ELK4 |  | yes4 |
| GABPA |  | yes4 |
| HIF1A::ARNT | yes1 |  |
| INSM1 |  | yes3 |
| Klf1 |  | yes5 |
| Klf4 |  | yes2,5 |
| KLF5 |  | yes2,3 |
| Mafb | yes1 | yes3 |
| Myc |  | yes5 |
| MYC::MAX |  | yes5 |
| Mycn | yes1 | yes5 |
| NHLH1 |  | yes3 |
| NRF1 |  | yes1,3 |
| Pax2 | yes1 |  |
| PAX5 | yes1 |  |
| Pax6 | yes1 |  |
| PPARG | yes1 |  |
| REST |  | yes1 |
| RFX5 | yes4 |  |
| RXR::RAR\_DR5 | yes4 |  |
| RXRA::VDR | yes4 |  |
| SP1 |  | yes2,3 |
| SP2 |  | yes2,3,4 |
| STAT1 | yes2 | yes4 |
| STAT2::STAT1 | yes3 |  |
| STAT3 | yes2 | yes4 |
| Stat4 | yes2,3 | yes4 |
| TAL1::GATA1 | yes4 |  |
| USF1 |  | yes5 |
| YY1 |  | yes3 |
| Zfx | yes2 | yes2 |

**Supplementary Table 2:** Gene ontology predictions for motifs overlapping with significantly associated probes in DMRs unique to Probe Lasso and window250 algorithms. Superscripts refer to the motif presented in Supplementary Figure 6 (1-5 = bottom-top). MF = Molecular function; CC = Cellular component; BP = Biological process.

|  |  |  |  |
| --- | --- | --- | --- |
| **category** | **GO prediction** | **Probe Lasso** | **Window250** |
| MF | zinc ion binding | yes2 |  |
| MF | transcription factor activity | yes1 |  |
| MF | transcription activator activity | yes2 | yes3,5 |
| MF | sequence-specific DNA binding | yes1 |  |
| MF | RNA binding | yes3 | yes4 |
| MF | protein serine/threonine kinase activity | yes4 |  |
| MF | magnesium ion binding |  | yes1,3,5 |
| MF | DNA binding | yes3 |  |
| MF | ATP binding | yes2,4 | yes1,2,3,4,5 |
| CC | transcription factor complex | yes2 | yes2,4 |
| CC | spliceosomal complex | yes3 |  |
| CC | nucleus | yes1 |  |
| CC | nucleolus | yes3,4 | yes1,4,5 |
| CC | microtubule |  | yes3 |
| CC | cytosol | yes2 | yes2,3,4,5 |
| BP | small GTPase mediated signal transduction |  | yes2 |
| BP | response to organic cyclic substance | yes1 |  |
| BP | regulation of transcription |  | yes1 |
| BP | protein folding |  | yes1 |
| BP | nuclear mRNA splicing, via spliceosome | yes3,4 |  |
| BP | negative regulation of cell migration | yes1 |  |
| BP | modification-dependent protein catabolic process |  | yes2 |
| BP | cell division | yes4 |  |