Distinct chromosomal “niches” in the genome of *Saccharomyces cerevisiae* provide the background for genomic innovation and shape the fate of gene duplicates.

Athanasia Stavropoulou[1,2], Emilios Tassios[1,2], Maria Kalyva[1], Michalis Georgoulopoulos[3], Nikolaos Vakirlis[2], Ioannis Iliopoulos[1] and Christoforos Nikolaou[2,3]*

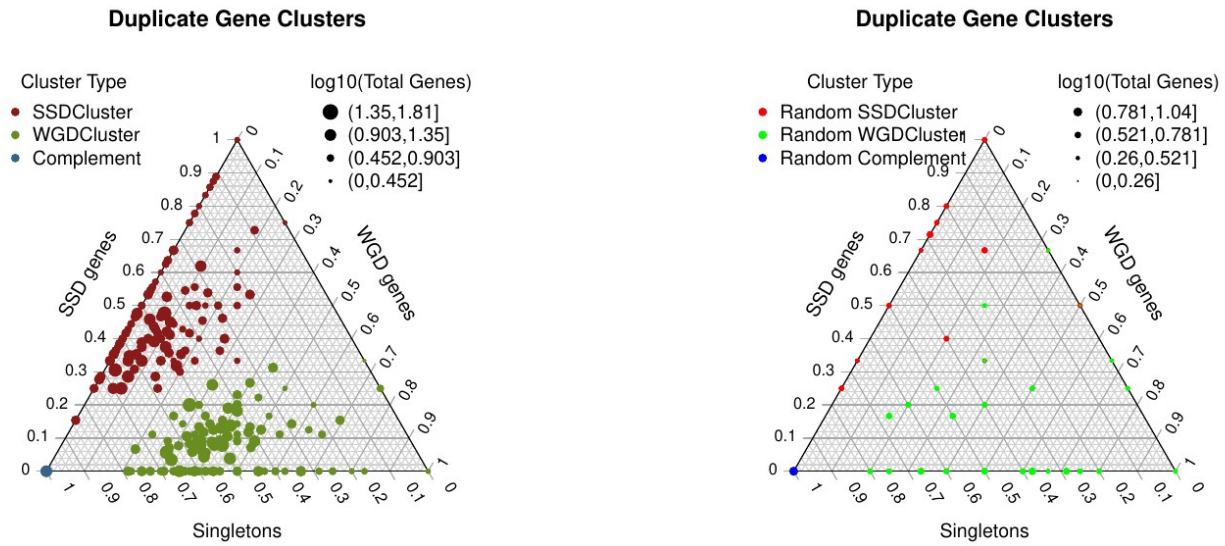
[1]*Medical School, University of Crete, Heraklion, 70013, Greece*

[2]*Computational Genomics Group, Biomedical Sciences Research Center “Alexander Fleming”, Athens, 16672, Greece*

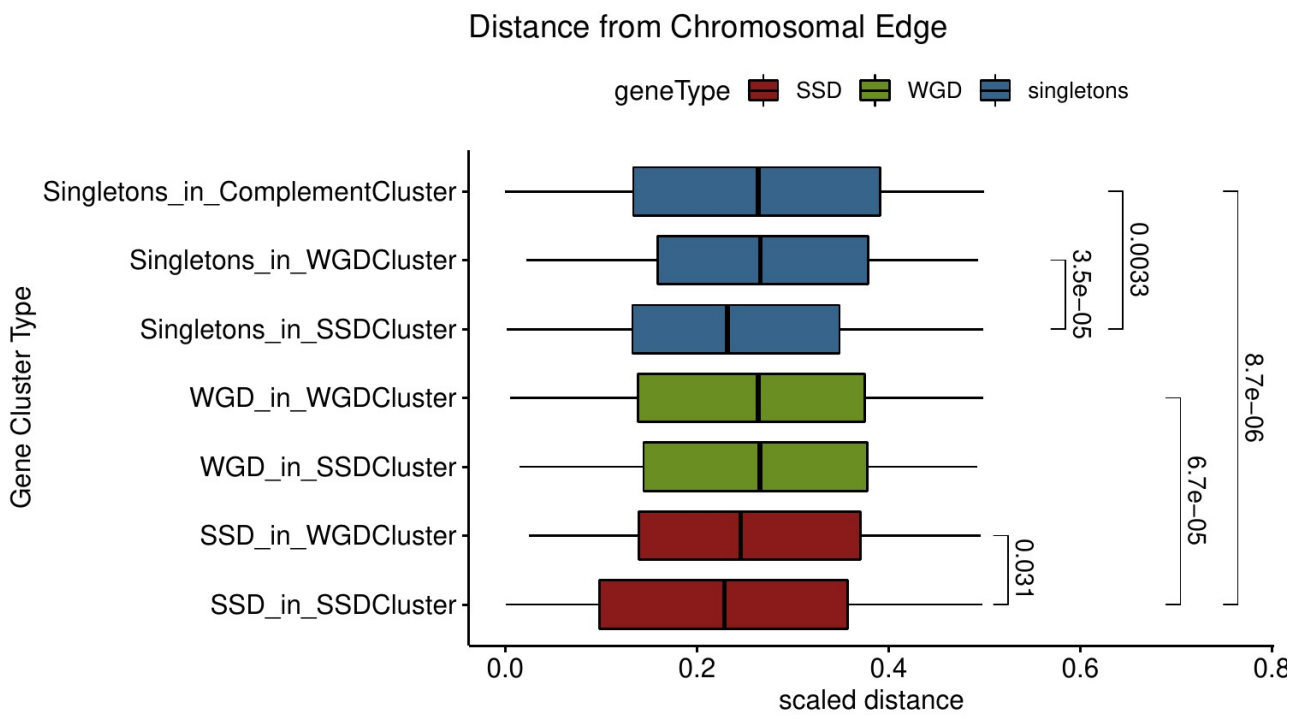
[3]*Hellenic Open University, Patras, 26335, Greece*

** Correspondence: cnikolaou@fleming.gr*

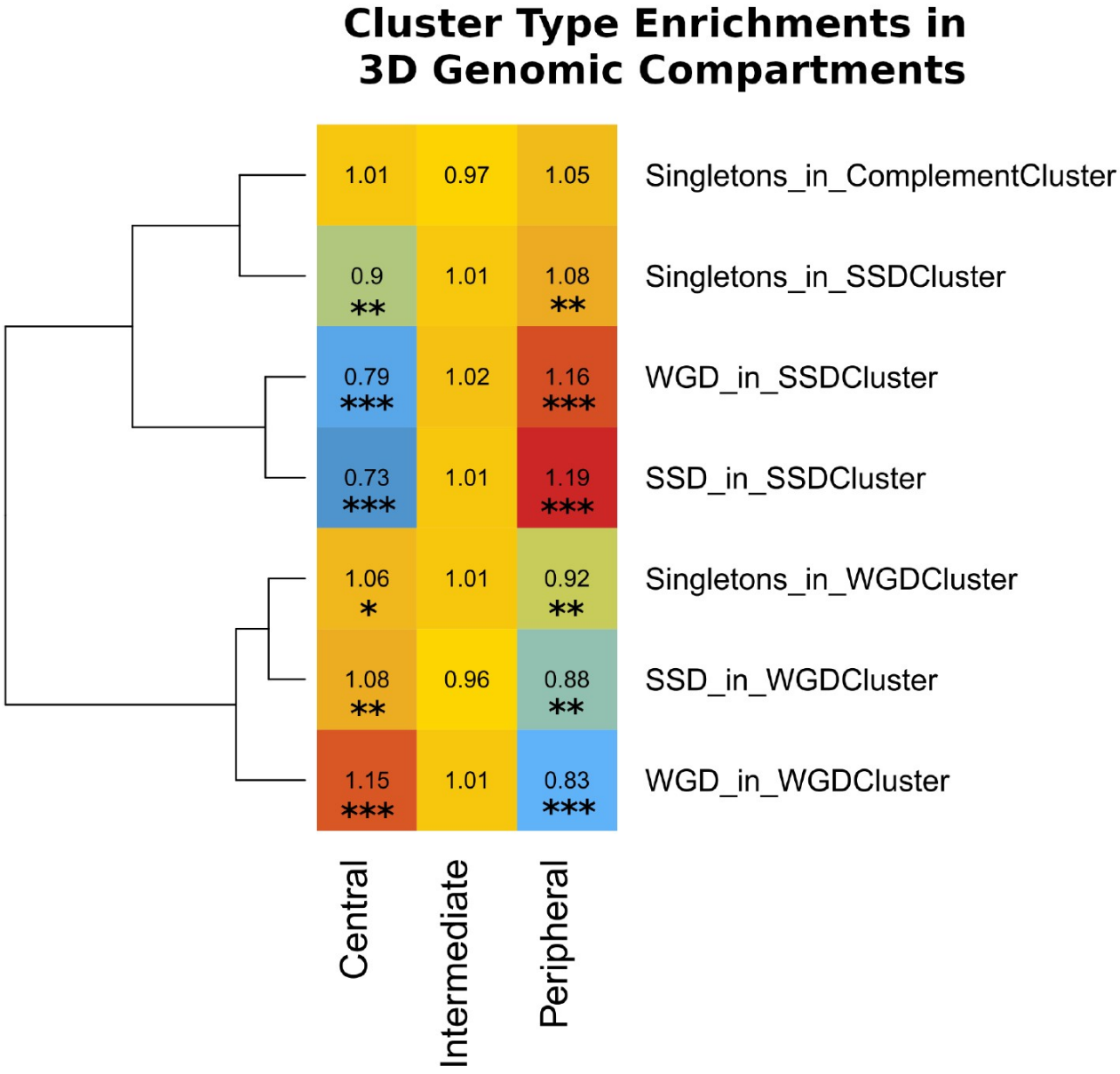
Supplementary Figures



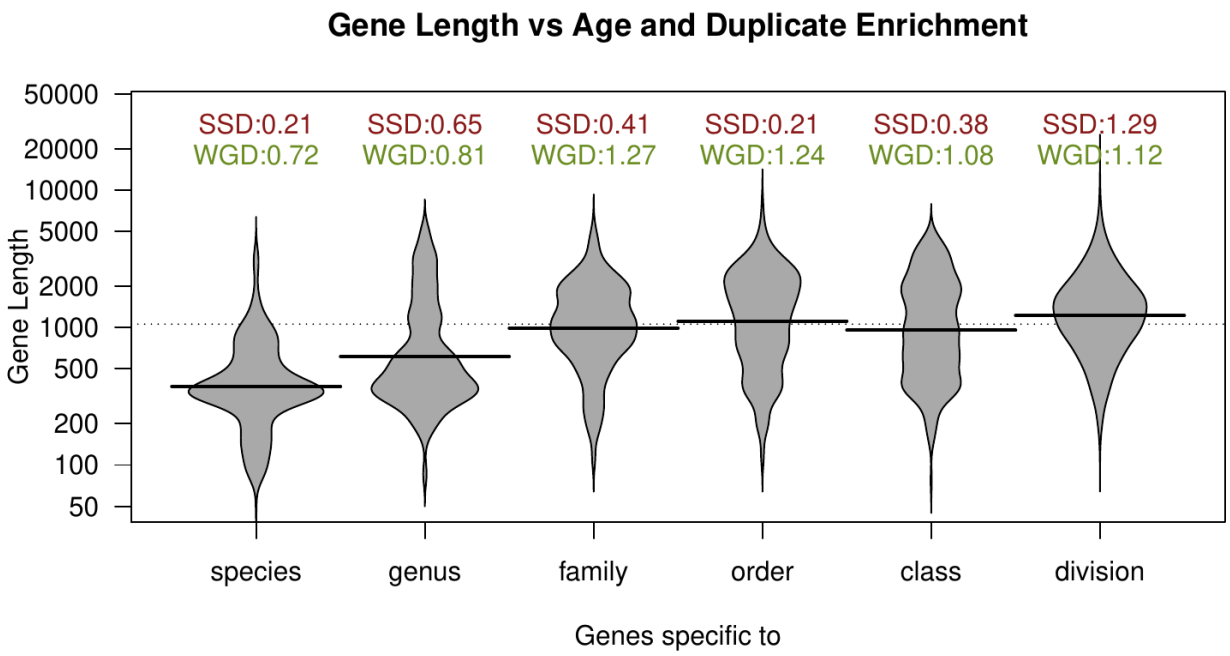
Supplementary Figure 1. Ternary plot showing the gene content of duplicate and complement clusters. The scale on each side of the triangle is relative and corresponds to the gene content in one of SSD, WGD and singleton gene categories. The colour of the dots corresponds to the cluster type. As all Complement Clusters contain exclusively singleton genes they all converge to one point at the lower left vertex of the plot. Size of the points corresponds to the size of the cluster in total number of genes. Left: Actual clusters used in this study (see also Figure 1E), with SSD and WGD clusters occupying different parts of the plot, suggestive of their relative purity in duplicate genes of a specific type. Right: The same plot for random clusters obtained through a random permutation of duplicate genes in the yeast genome. The smaller number of points is due to many clusters having identical gene content. Notice the smaller size of the clusters in terms of gene number.



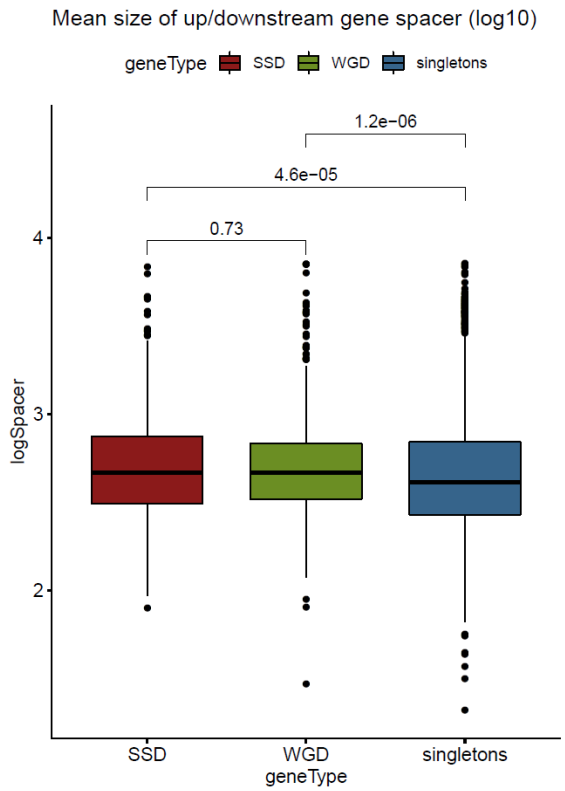
Supplementary Figure 2. Distribution of scaled distances of gene clusters from the most proximal chromosomal edge. Significant differences are denoted with p-values of a Mann Whitney test. P-values are adjusted for multiple comparisons.



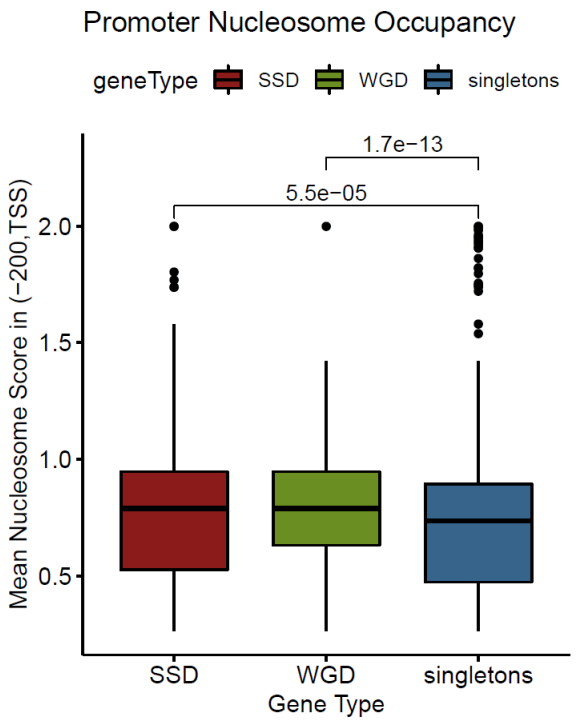
Supplementary Figure 3. Relative enrichments of gene clusters in the three partitions of the yeast genome according to the polar distance from an assumed genome center (see Methods in main text for details). Colors are representative of enrichment with cold colours (blue/green) corresponding to low values, while warm ones (orange, red) correspond to high values. Enrichments were calculated as observed over expected ratios of gene numbers, thus having a baseline value of 1. Significance was assessed with permutation tests (*, p-value<=0.001, **, p-value<=0.01; *, p-value <=0.05).**



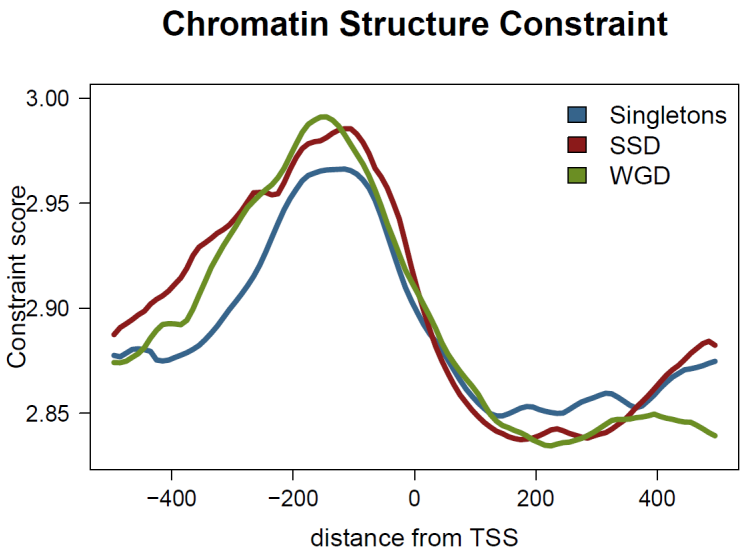
Supplementary Figure 4. Gene length for sets of yeast genes that are specific to different age groups, accompanied by the relative enrichments of duplicate genes in each group (baseline enrichment is 1).



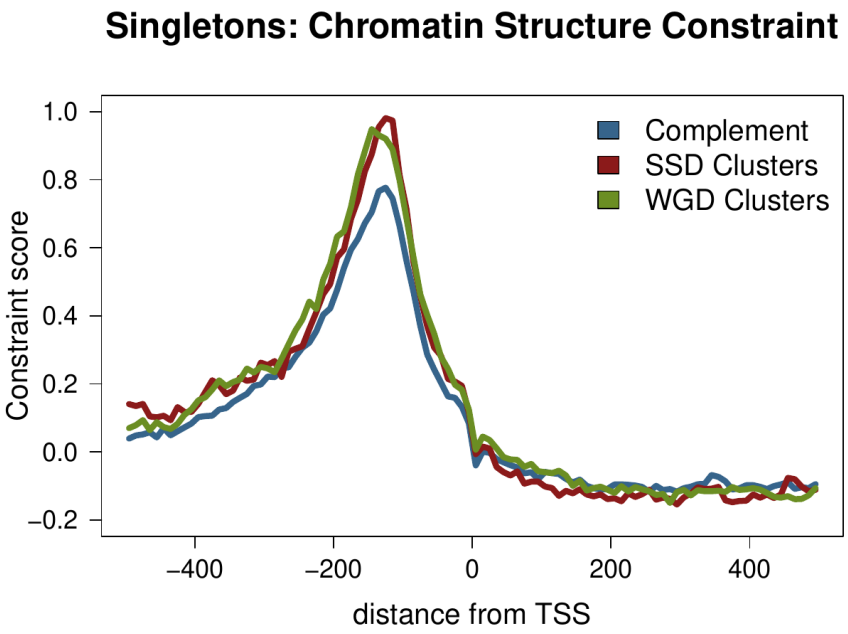
Supplementary Figure 5. Distribution of the mean size of gene upstream and gene downstream distances of SSD, WGD and singleton genes form their flanking genes. Values over brackets denote adjusted p-values of a Mann Whitney test.



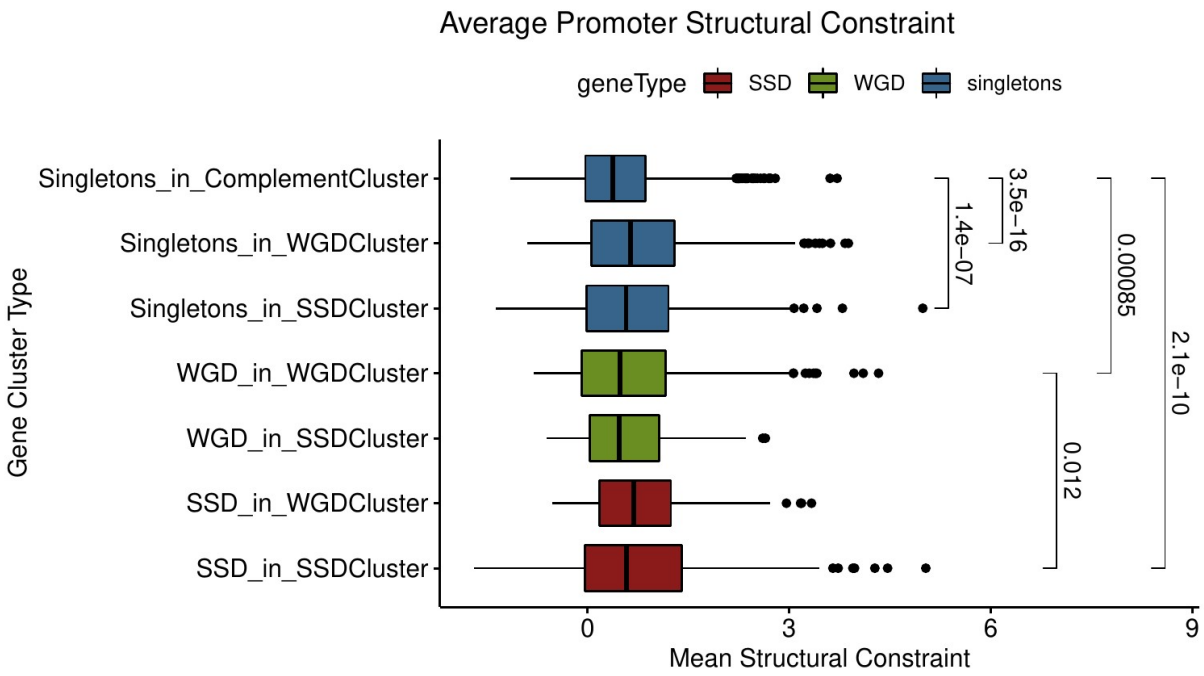
Supplementary Figure 6. Distribution of mean nucleosome occupancy scores for a 200bp region immediately upstream of the TSS for SSD, WGD and 1000 random singleton genes. Values over brackets denote adjusted p-values of a Mann Whitney test.



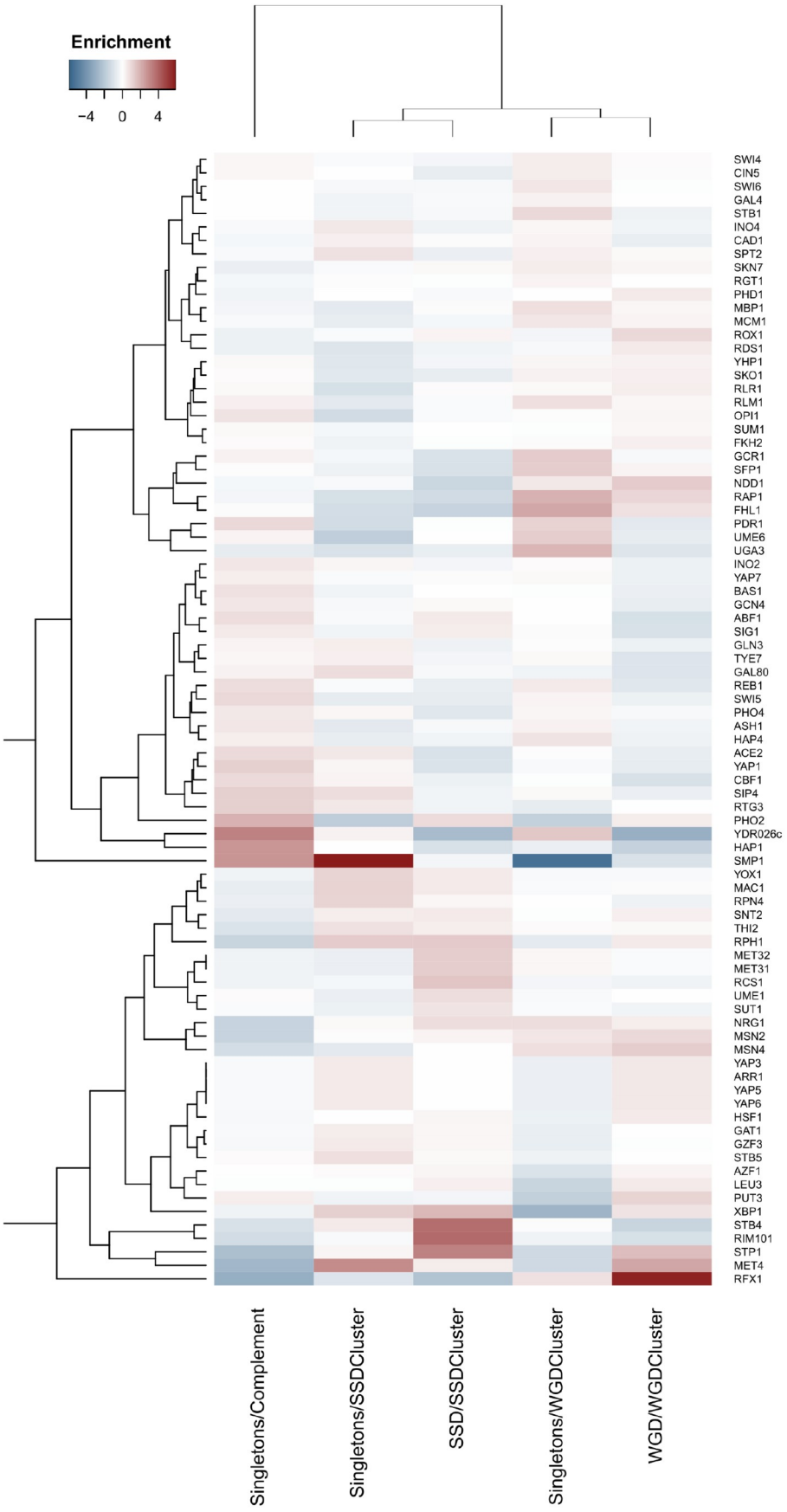
Supplementary Figure 7. Chromatin structure constraint measured as mean SymCurv Robustness (Nikolaou et al, Epigenetics and Chromatin, 2010) along a region spanning 500bps either side of TSS for singleton, SSD and WGD genes.



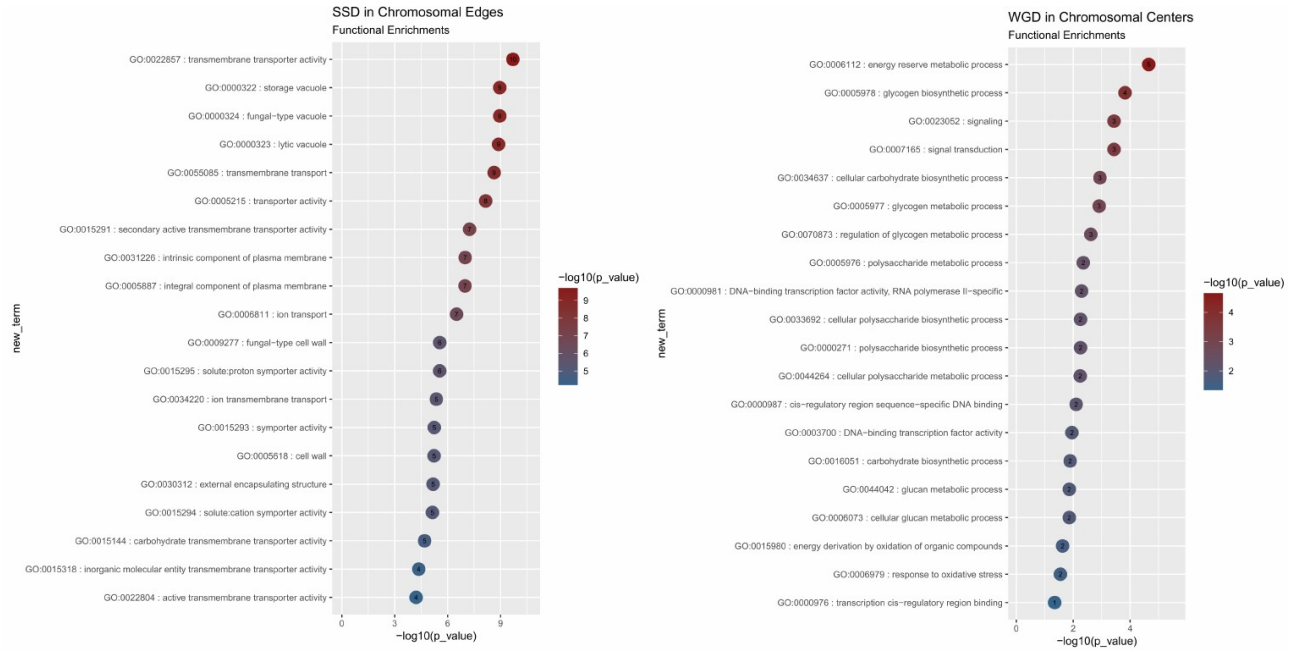
Supplementary Figure 8. Chromatin structure constraint score as measured with the "mutation score" (Routhier et al, Genome Research, 2020) along a region spanning 500bps either side of TSS for singleton genes residing in the genome complement, SSD and WGD clusters.



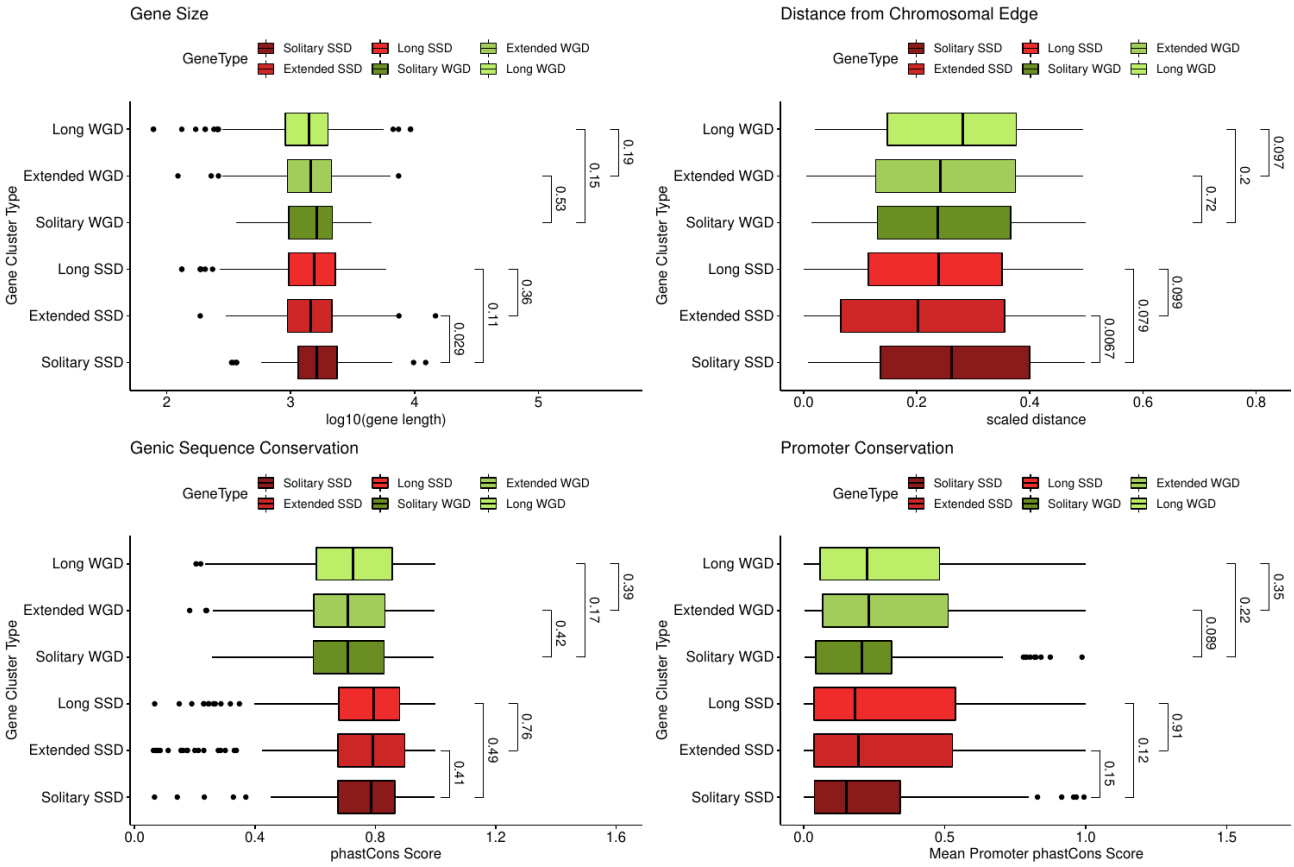
Supplementary Figure 9. Distribution of mean promoter structural constraint measured as mutation score conservation in the proximal promoter region (200bp to 50bp upsteam to TSS) measured as mutation score (Routhier et al, Genome Research, 2020). Values next to brackets denote p-values of a Mann Whitney test. P-values are adjusted for multiple comparisons.


Enrichment
−4
0
4
SWI4
CIN5
SWI6
GAL4
STB1
INO4
CAD1
SPT2
SKN7
RGT1
PHD1
MBP1
MCM1
ROX1
RDS1
YHP1
SKO1
RLR1
RLM1
OPI1
SUM1
FKH2
GCR1
SFP1
NDD1
RAP1
FHL1
PDR1
UME6
UGA3
INO2
YAP7
BAS1
GCN4
ABF1
SIG1
GLN3
TYE7
GAL80
REB1
SWI5
PHO4
ASH1
HAP4
ACE2
YAP1
CBF1
SIP4
RTG3
PHO2
YDR026c
HAP1
SMP1
YOX1
MAC1
RPN4
SNT2
THI2
RPH1
MET32
MET31
RCS1
UME1
SUT1
NRG1
MSN2
MSN4
YAP3
ARR1
YAP5
YAP6
HSF1
GAT1
GZF3
STB5
AZF1
LEU3
PUT3
XBP1
STB4
RIM101
STP1
MET4
RFX1
Singletons/Complement
Singletons/SSDCluster
SSD/SSDCluster
Singletons/WGDCluster
WGD/WGDCluster

Supplementary Figure 10. Enrichment heatmap of transcriptional regulator binding sites in different gene categories. Enrichments calculated as log2(observed/expected ratios).



Supplementary Figure 11. Functional enrichment plots representing the top most enriched GO terms for genes found in a) SSD genes in SSD clusters that lie close to the chromosomal edges (<20% of the chromosomal arm length) and b) WGD in WGD clusters found near the centromeres (>80% of the chromosomal arm length away from the edge).



Supplementary Figure 12. Comparisons for four major, duplicate-specific properties, discussed in this study (gene size, distance from chromosomal edge, coding sequence conservation, promoter sequence conservation) between SSD and WGD pertaining to single-gene clusters (Solitary, n=1 gene), intermediate-sized clusters (Extended, 1<n<=5 genes) and long clusters (n>5 genes). Overall there are very few significant differences, suggesting that the properties of duplicate genes are generally independent of the size of the cluster they are found in.