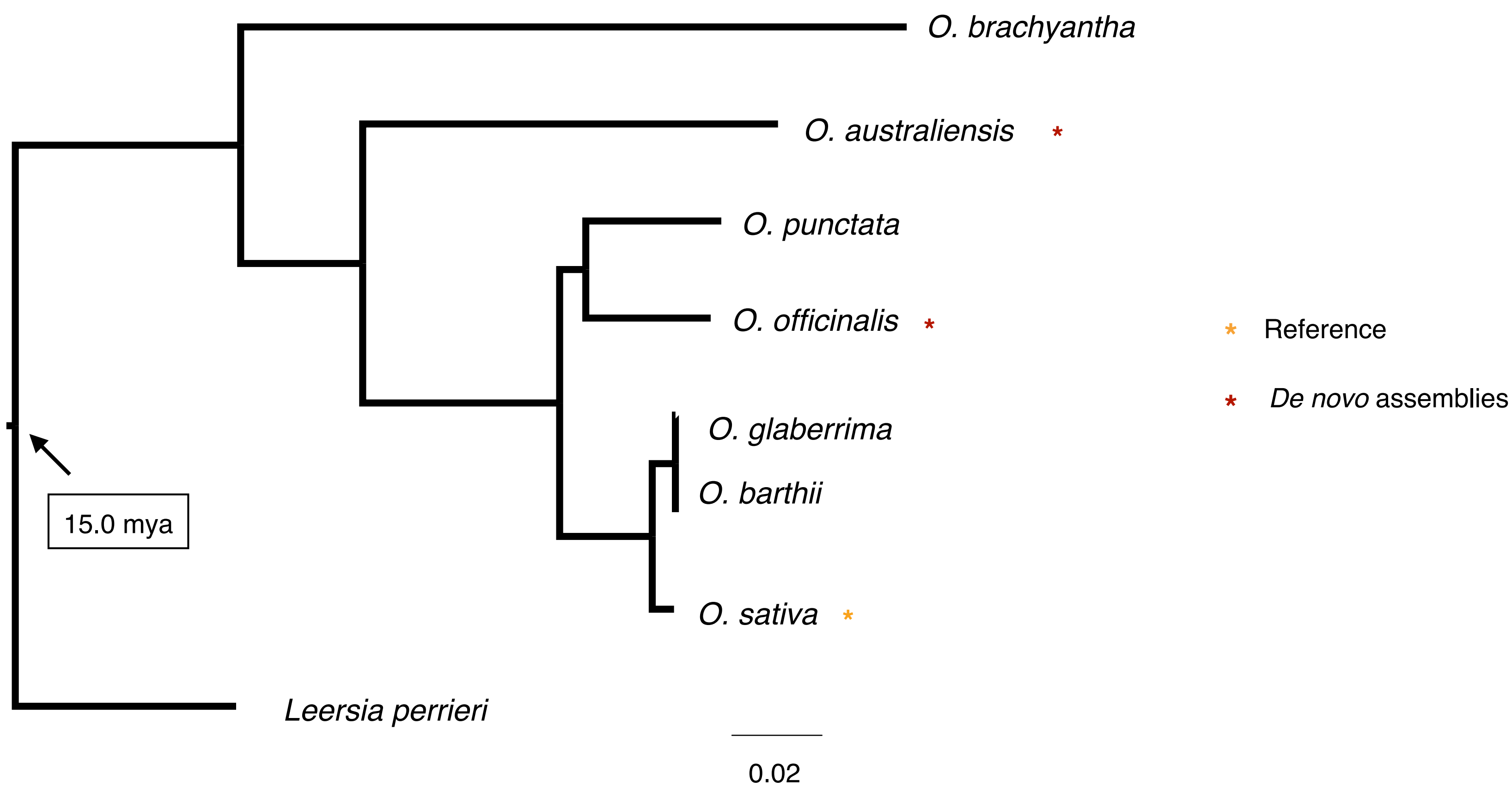
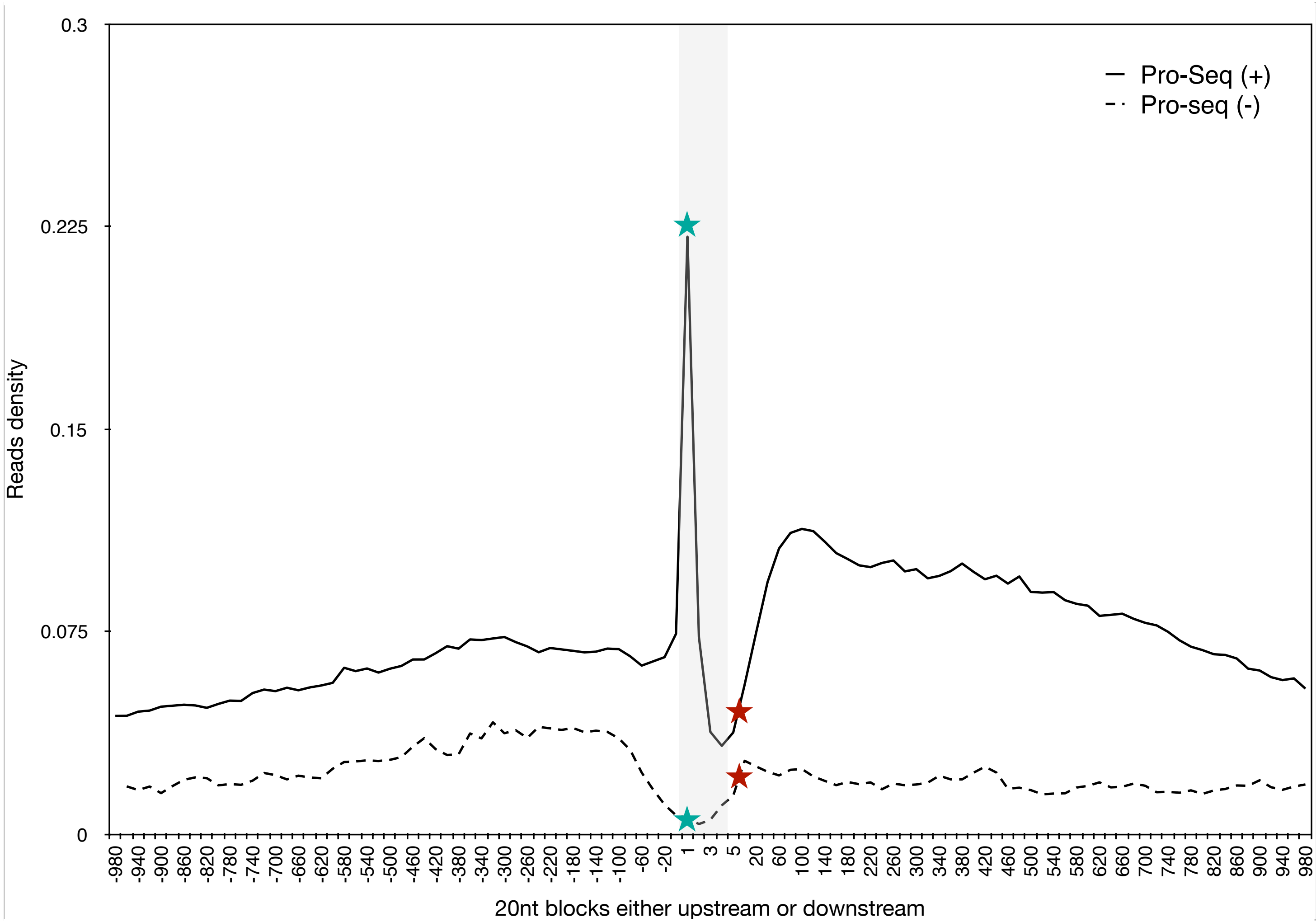
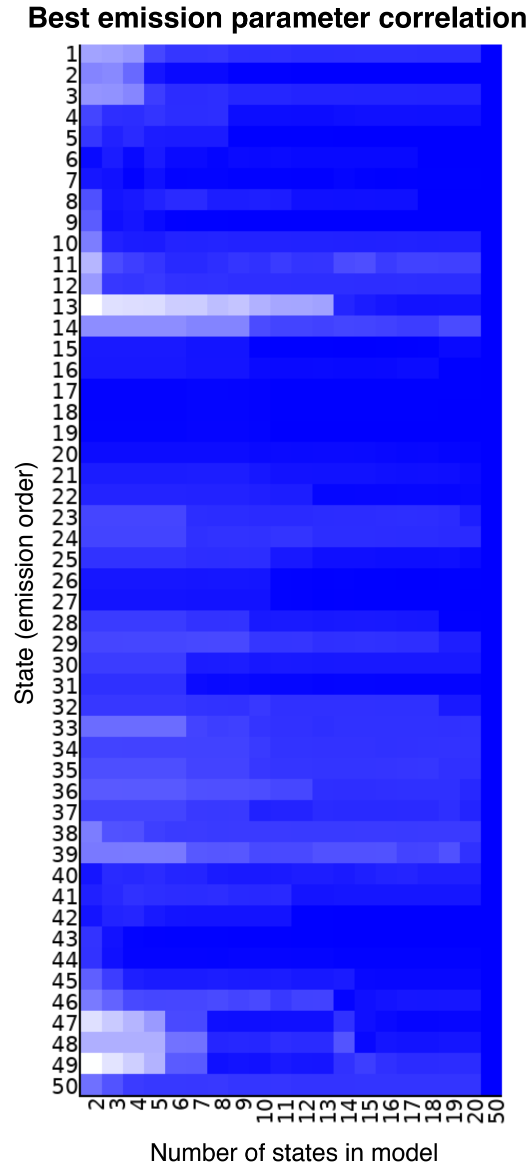


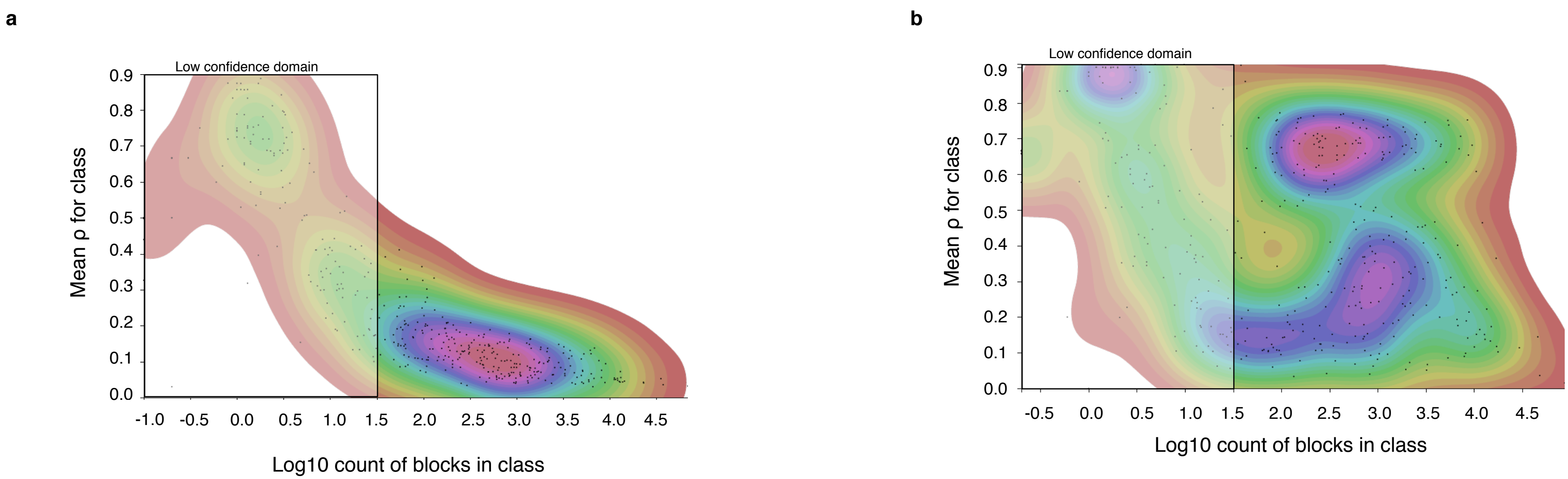
Supplementary Fig. 1. Phylogeny of the 8-way alignment of rice genomes. Different subsets of genomes were used for the INSIGHT analysis (*O. punctata*, *O. australiensis*, *O. officinalis*, *O. longistaminata*, *O. sativa*, *O. nivara*, *O. glaberrima*, *O. barthii*, *O. glumaepatula*, *O. ruffipogon*, *O. brachyantha*) and the classical analysis of conserved noncoding sequences where genomes with a lower tendency for introgression with the sativa reference were preferred (*Leersia perrieri*, *O. brachyantha*, *O. australiensis*, *O. officinalis*, *O. punctata*, *O. barthii*, *O. sativa**, *O. glaberrima*).



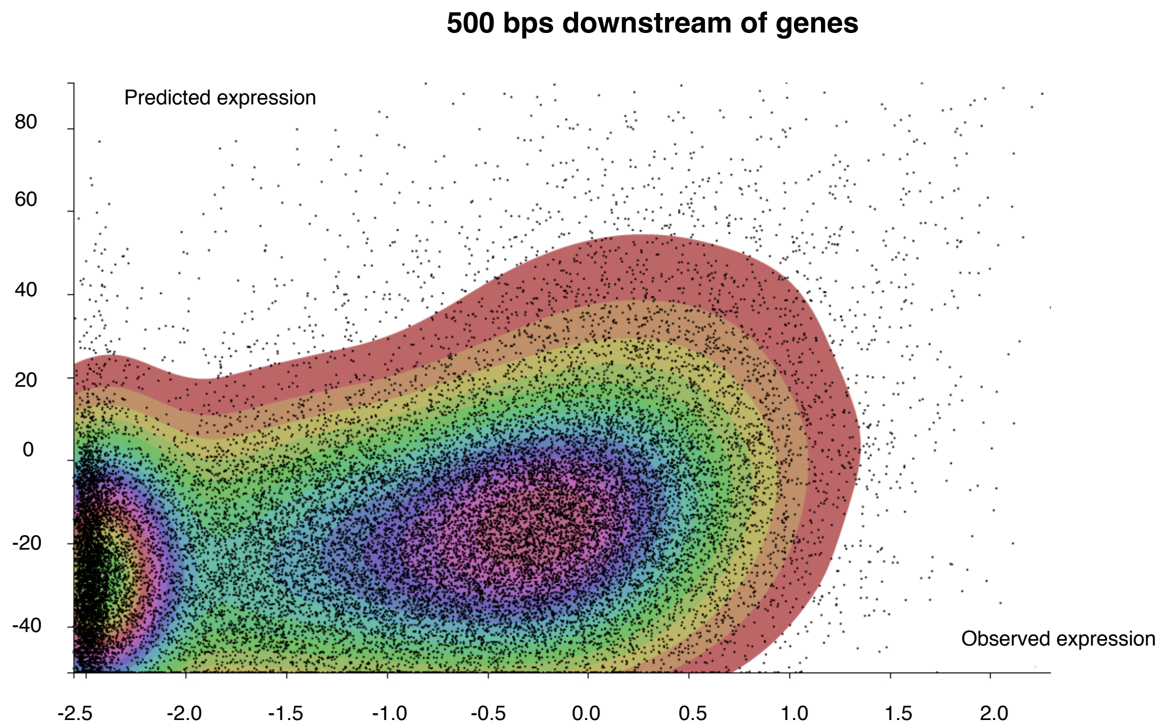
Supplementary Fig. 2. PRO-seq read distribution within and around protein-coding genes. Density plot of PRO-seq read signal around genes (grey box) (1kb upstream and 1kb downstream). Reads were aligned in both sense and antisense directions relative to the direction of gene transcription. Prominent promoter-proximal pausing (green stars), as well as accumulation of RNA polymerases at the 3' end of the genes (red stars) is evident.



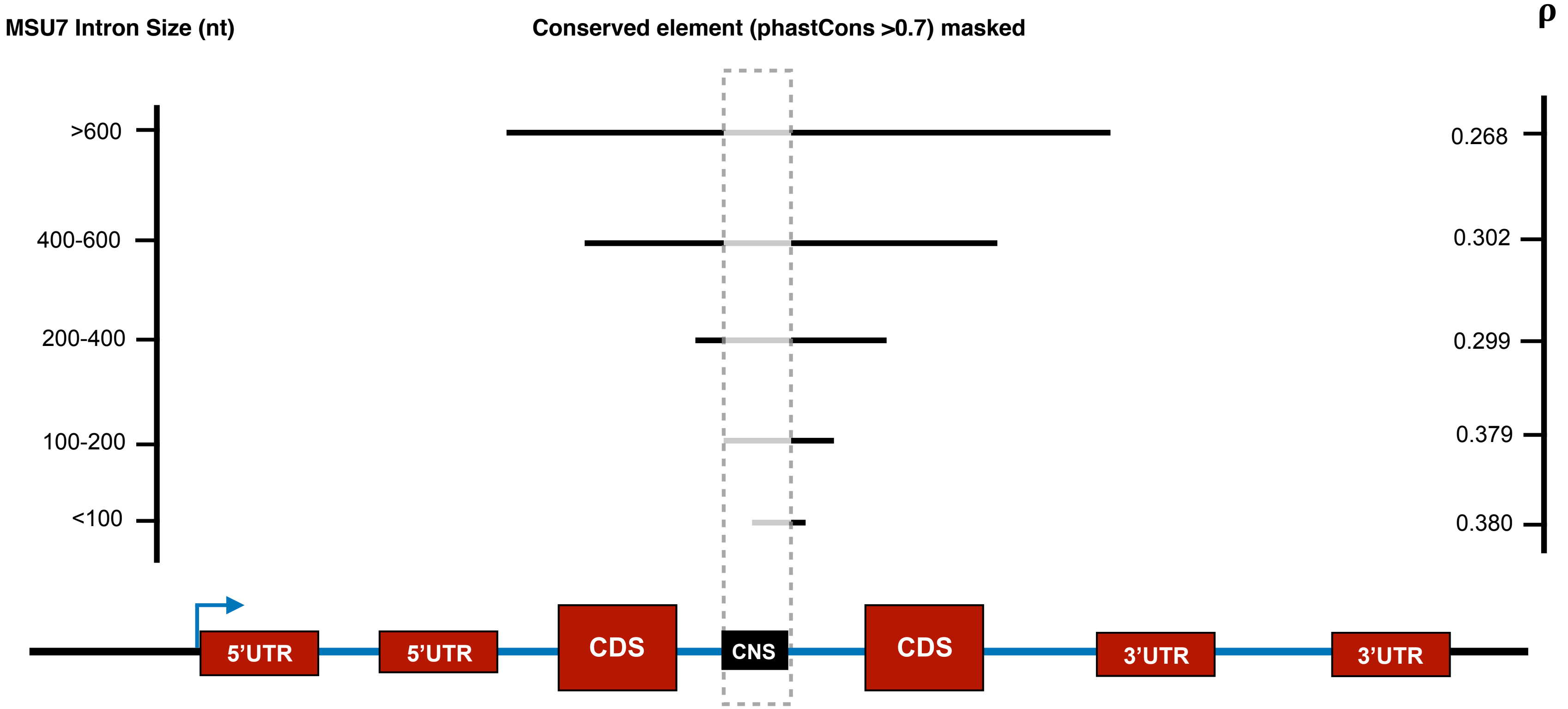
Supplementary Fig. 3. Emission parameter correlation comparison heat map. ChromHMM compare models shows convergence in emission correlations of models with fewer states relative to a 50 state model. Rows corresponds to a state from a 50 state model, and columns models. The intensity of a cell indicates the maximum emission parameter correlation of any state in the model of the column with the state of the row from the 50 state model. See Fig. 2c for the 20 emission parameters used in the FitCons analysis.



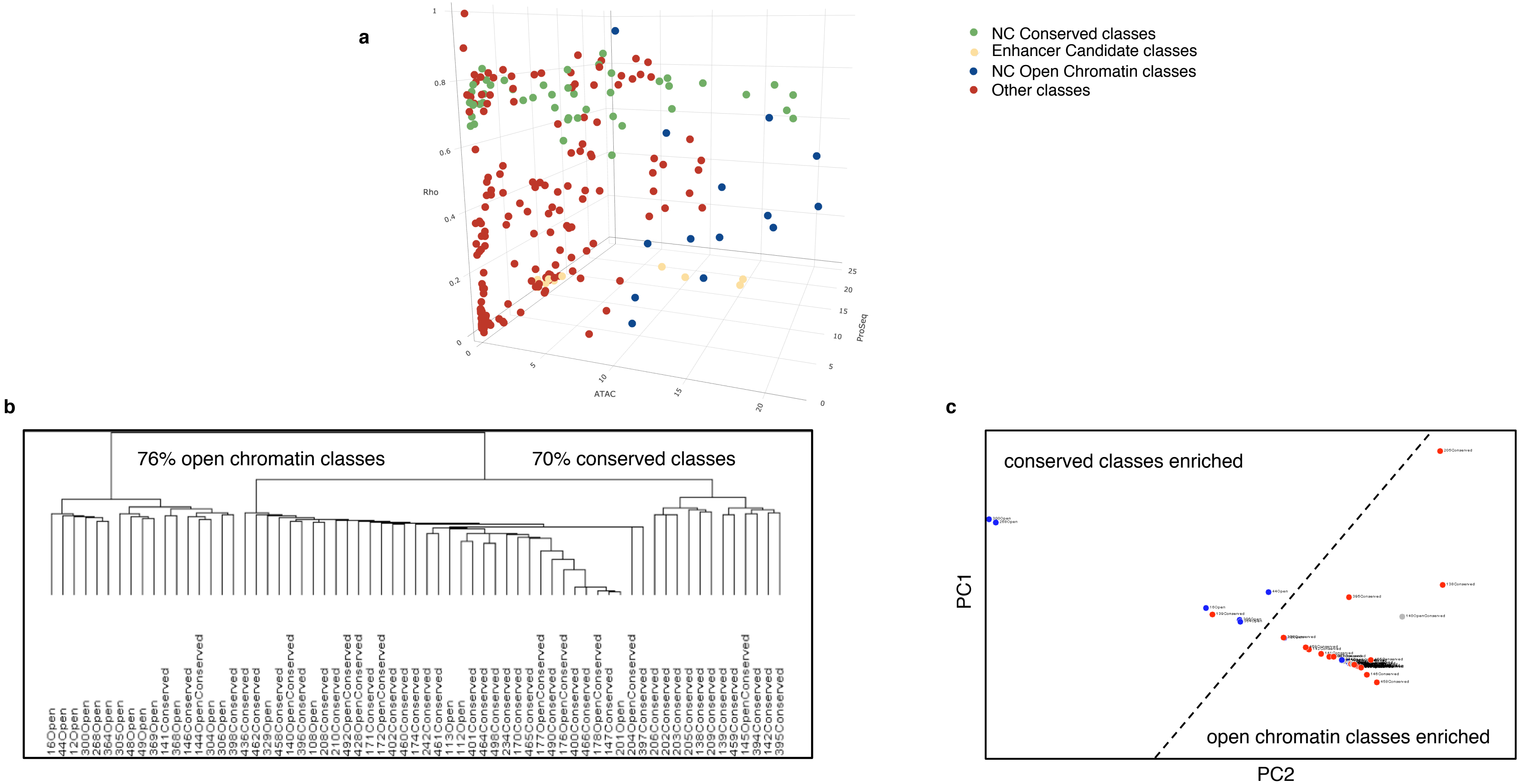
Supplementary Fig. 4. ρ scores validation between randomized fitCons classes and actual fitCons classes. The distribution of ρ scores across classes was compared with ρ scores for similarly sized classes transposed to random chromosomes to assess whether classes appeared to be more coherent than would be expected from a randomized model. **a,** ρ scores from randomized fitCons locations (average of 10 replicates) **b**, ρ from actual fitCons locations. The ρ distribution of classes with randomized genomic locations tended towards zero as block count increased, differing markedly from the observed trend in which a subset of high block count classes retained a high ρ.



Supplementary Fig. 5. Log expression predicted from fitCons class distribution 500bp downstream of genes against log gene expression (see Methods).



Supplementary Fig. 6. Intron ρ scores and effect of nearby coding sequences (CDSs). The ρ score of introns, when masked for detected conserved non coding sequences (CNSs) is indicated (right) for introns of various sizes (left, in nucleotides length). We detect a bias of higher ρ scores for introns that are smaller in size. ρ scores generally decreases as intron length increases and the decrease in ρ is more than would be expected from a dispersal of the same number of elements into a larger space, suggesting background selection. We detect a drop in ρ scores in introns longer than 200nt, suggesting strong linkage in the 1-199nt distance range.



Supplementary Fig. 7. Three categories of noncoding (NC) fitCons classes highlighted in this study. **a,** 3D plot (see Supplementary web data http://purugganan-genomebrowser.bio.nyu.edu/greenInsight/3dscatter/ for actual model) showing the spatial organization of the 246 classes along three axes y=ρ, x=ATAC enrichment, and z=PRO-seq enrichment. The three categories of classes are a subset of the 246 genome classes and are color-coded, while the remaining classes are labeled in red. Conserved classes are evident at high-ρ (green), Open Chromatin classes are evident with a high ATAC enrichment (blue), while Enhancer Candidates (yellow) show PRO-seq enrichment and generally open chromatin but with little enrichment for ρ values. Note that some classes with both ATAC enrichment and conservation were collapsed for the figure to one of the two classes **b**, Enriched 6-8bp motifs identified across all classes by Homer were used to hierarchically cluster the classes (DChip, classes n = 246, IUPAC consensus motifs 6-8bp (covariates) n = 5,549 , individual motif instances n = 8,357,163). The motif enrichments suggest a primary complexity-linked bifurcation that can partition the majority of low ρ/accessible associated site classes from high ρ/conserved classes. This is evident although less clearly by PCA (DChip, numbers as in b) (**c**).

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)	Best Match/Details
1		1e-52	-1.210e+02	11.43%	6.72%	33.0bp (37.7bp)	HAP3(CCAATHAP3)/col-HAP3-DAP-Seq(GSE60143)/Homer(0.766)
2		1e-47	-1.093e+02	1.05%	0.08%	21.2bp (49.8bp)	CDC5(MYB)/Arabidopsis thaliana/AthaMap(0.701)
3		1e-44	-1.035e+02	12.72%	8.07%	34.6bp (35.3bp)	WRKY62/MA1091.1/Jaspar(0.668)
4		1e-44	-1.032e+02	12.34%	7.77%	32.1bp (37.2bp)	bZIP52(bZIP)/colamp-bZIP52-DAP-Seq(GSE60143)/Homer(0.694)
5		1e-39	-9.068e+01	1.28%	0.21%	18.7bp (29.9bp)	TBP3(MYBrelated)/col-TBP3-DAP-Seq(GSE60143)/Homer(0.656)
6		1e-38	-8.908e+01	12.83%	8.47%	32.0bp (36.9bp)	LEC2/MA0581.1/Jaspar(0.702)
7		1e-38	-8.864e+01	5.06%	2.45%	29.0bp (37.3bp)	bZIP44(bZIP)/colamp-bZIP44-DAP-Seq(GSE60143)/Homer(0.805)
8		1e-35	-8.213e+01	4.59%	2.20%	40.2bp (37.7bp)	BPC1(BBRBPC)/colamp-BPC1-DAP-Seq(GSE60143)/Homer(0.774)
9		1e-28	-6.668e+01	7.83%	4.88%	34.5bp (33.5bp)	AT5G61620(MYBrelated)/colamp-AT5G61620-DAP-Seq(GSE60143)/Homer(0.743)
10		1e-25	-5.849e+01	8.41%	5.52%	32.0bp (25.9bp)	ABI5(bZIP)/col-ABI5-DAP-Seq(GSE60143)/Homer(0.737)
11		1e-24	-5.714e+01	3.65%	1.85%	26.4bp (32.0bp)	BHLH34/MA0962.1/Jaspar(0.829)
12		1e-21	-4.918e+01	1.53%	0.54%	33.4bp (36.4bp)	POL010.1_DCE_S_III/Jaspar(0.601)
13		1e-18	-4.339e+01	0.61%	0.10%	16.1bp (30.2bp)	E2FA(E2FDP)/colamp-E2FA-DAP-Seq(GSE60143)/Homer(0.721)
14		1e-15	-3.497e+01	1.55%	0.67%	25.5bp (39.9bp)	ASHR1(ND)/col-ASHR1-DAP-Seq(GSE60143)/Homer(0.778)
15 *		1e-11	-2.632e+01	8.98%	6.95%	34.7bp (37.6bp)	ATHB34(ZFHD)/colamp-ATHB34-DAP-Seq(GSE60143)/Homer(0.876)
16 *		1e-10	-2.322e+01	0.84%	0.33%	43.0bp (31.6bp)	TBP(- other)/several species/AthaMap(0.825)
17 *		1e-8	-2.070e+01	1.08%	0.52%	30.6bp (31.1bp)	GATA4(C2C2gata)/col-GATA4-DAP-Seq(GSE60143)/Homer(0.741)
18 *		1e-8	-1.855e+01	0.39%	0.11%	17.6bp (34.8bp)	At3g12730(G2like)/colamp-At3g12730-DAP-Seq(GSE60143)/Homer(0.801)
19 *		1e-7	-1.732e+01	0.69%	0.29%	35.6bp (30.9bp)	AtGRF6(GRF)/col-AtGRF6-DAP-Seq(GSE60143)/Homer(0.666)
20 *		1e-6	-1.399e+01	2.74%	1.95%	39.3bp (34.3bp)	ATHB24(ZFHD)/colamp-ATHB24-DAP-Seq(GSE60143)/Homer(0.896)

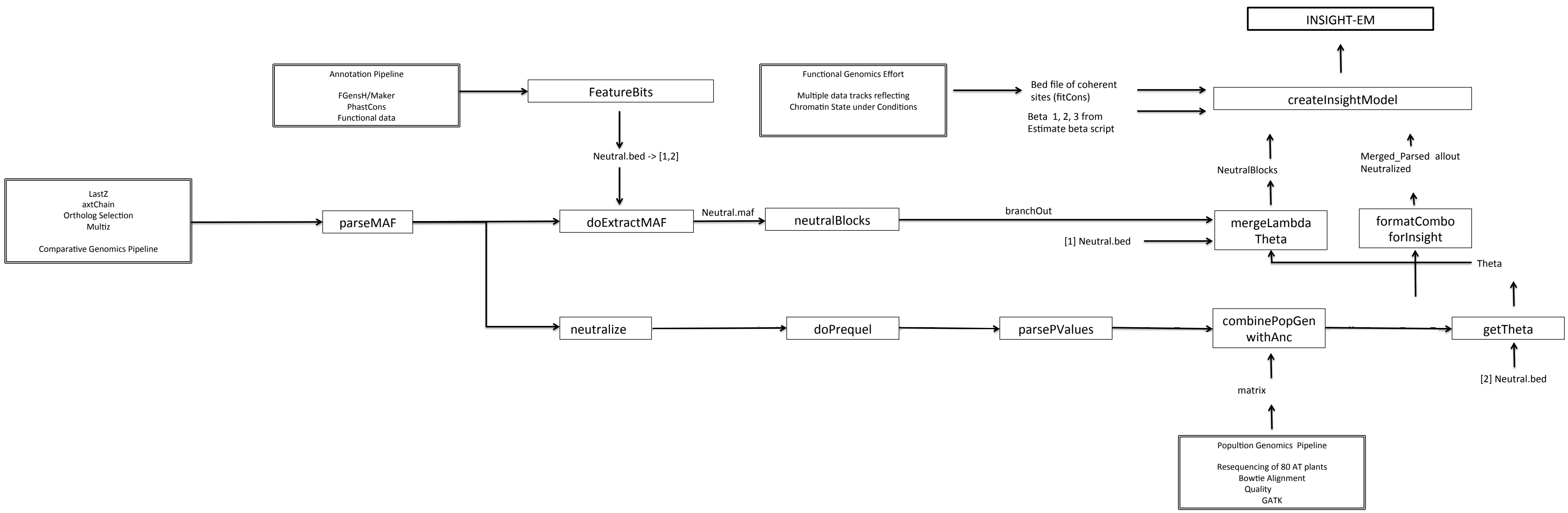
Supplementary Fig. 8. Motif analysis of a subset of 10,000 Conserved class sites (length = 0.35 Mb). Top 20 de novo motifs generated by the software Homer in which p-values and enrichment are modeled relative to a modified cumulative hypergeometric distribution (for more information see: http://homer.ucsd.edu/homer/motif/). Red stars indicate possible false positive motifs.

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)	Best Match/Details
1	CGGCGGCG	1e-447	-1.030e+03	15.47%	2.78%	35.1bp (39.5bp)	Os05g0497200/MA1034.1/Jaspar(0.933)
2	GAGGAG	1e-254	-5.854e+02	17.16%	5.91%	37.3bp (35.9bp)	TF3A(C2H2)/col-TF3A-DAP-Seq(GSE60143)/Homer(0.892)
3	TCTCCCTC	1e-232	-5.365e+02	25.06%	11.64%	39.6bp (30.0bp)	FRS9(ND)/col-FRS9-DAP-Seq(GSE60143)/Homer(0.785)
4	CTAGGGTT	1e-143	-3.297e+02	2.49%	0.07%	33.3bp (34.6bp)	P0510F09.23/MA1030.1/Jaspar(0.857)
5	CGAGTCGA	1e-103	-2.373e+02	7.33%	2.52%	35.2bp (42.3bp)	Knotted(Homeobox)/Corn-KN1-ChIP-Seq(GSE39161)/Homer(0.713)
6	CTGCAGCT	1e-90	-2.079e+02	2.38%	0.26%	37.8bp (29.0bp)	POL010.1_DCE_S_III/Jaspar(0.635)
7	CCACCACC	1e-72	-1.673e+02	2.68%	0.49%	46.0bp (38.0bp)	AT3G57600(AP2EREBP)/col-AT3G57600-DAP-Seq(GSE60143)/Homer(0.803)
8	TAGCTCGC	1e-59	-1.366e+02	2.72%	0.64%	32.8bp (46.8bp)	POL010.1_DCE_S_III/Jaspar(0.682)
9	CAAATT	1e-50	-1.169e+02	10.59%	6.12%	37.0bp (40.1bp)	AT1G20910(ARID)/col-AT1G20910-DAP-Seq(GSE60143)/Homer(0.763)
10	GCTTAGCT	1e-50	-1.165e+02	2.65%	0.71%	37.8bp (39.2bp)	CDC5(MYB)/Arabidopsis thaliana/AthaMap(0.678)
11	CGATCGAT	1e-48	-1.125e+02	1.85%	0.36%	41.4bp (41.9bp)	GATA20(C2C2gata)/colamp-GATA20-DAP-Seq(GSE60143)/Homer(0.743)
12	TTCTTC	1e-46	-1.070e+02	6.13%	2.96%	36.3bp (34.6bp)	PEND/MA0127.1/Jaspar(0.750)
13	TCCGGCGA	1e-45	-1.058e+02	1.59%	0.27%	34.2bp (60.0bp)	AT3G58630(Trihelix)/col-AT3G58630-DAP-Seq(GSE60143)/Homer(0.939)
14	TAGTAGTA	1e-37	-8.602e+01	0.99%	0.11%	30.7bp (46.1bp)	At3g60580(C2H2)/col-At3g60580-DAP-Seq(GSE60143)/Homer(0.716)
15	TGGCTTC	1e-35	-8.167e+01	4.44%	2.09%	43.5bp (42.9bp)	POL008.1_DCE_S_I/Jaspar(0.779)
16	CCCCCAA	1e-34	-7.998e+01	2.49%	0.86%	36.8bp (30.0bp)	MYB62(MYB)/colamp-MYB62-DAP-Seq(GSE60143)/Homer(0.637)
17	AAACAAA	1e-33	-7.664e+01	4.73%	2.36%	36.8bp (28.9bp)	REM19(REM)/colamp-REM19-DAP-Seq(GSE60143)/Homer(0.752)
18	TGTGTGTG	1e-21	-5.017e+01	4.45%	2.54%	38.6bp (46.9bp)	FAR1(FAR1)/col-FAR1-DAP-Seq(GSE60143)/Homer(0.712)
19	TGAAGCTT	1e-16	-3.741e+01	0.68%	0.15%	42.8bp (30.7bp)	HSF6(HSF)/col-HSF6-DAP-Seq(GSE60143)/Homer(0.769)
20 *	GATTGGAT	1e-11	-2.550e+01	0.29%	0.04%	40.5bp (23.8bp)	POL004.1_CCAAT-box/Jaspar(0.759)

Supplementary Fig. 9. Motif analysis of a subset of 10,000 Open Chromatin class sites (length = 0.36 Mb). Top 20 de novo motifs generated by the software Homer in which p-values and enrichment are modeled relative to a modified cumulative hypergeometric distribution (for more information see: http://homer.ucsd.edu/homer/motif/). Red stars indicate possible false positive motifs.

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)	Best Match/Details
1		1e-124	-2.876e+02	58.50%	19.19%	182.2bp (221.2bp)	REM19(REM)/colamp-REM19-DAP-Seq(GSE60143)/Homer(0.962)
2		1e-90	-2.081e+02	64.60%	29.56%	183.1bp (180.2bp)	Unknown4/Arabidopsis-Promoters/Homer(0.784)
3		1e-79	-1.822e+02	21.40%	2.70%	232.6bp (184.4bp)	Os05g0497200/MA1034.1/Jaspar(0.945)
4		1e-78	-1.798e+02	50.60%	20.02%	184.9bp (184.5bp)	FRS9(ND)/col-FRS9-DAP-Seq(GSE60143)/Homer(0.792)
5		1e-47	-1.100e+02	73.50%	48.18%	188.0bp (178.3bp)	HDG1(Homeobox)/col100-HDG1-DAP-Seq(GSE60143)/Homer(0.797)
6		1e-46	-1.079e+02	58.60%	33.39%	200.4bp (178.4bp)	At3g09600(MYBrelated)/colamp-At3g09600-DAP-Seq(GSE60143)/Homer(0.695)
7		1e-39	-9.131e+01	26.10%	9.24%	184.8bp (178.7bp)	OsI_08196/MA1050.1/Jaspar(0.941)
8		1e-39	-9.127e+01	35.60%	15.86%	204.8bp (185.5bp)	TF3A(C2H2)/col-TF3A-DAP-Seq(GSE60143)/Homer(0.892)
9		1e-34	-7.976e+01	33.20%	15.15%	208.8bp (176.0bp)	bZIP69(bZIP)/col-bZIP69-DAP-Seq(GSE60143)/Homer(0.699)
10		1e-34	-7.920e+01	21.60%	7.28%	171.9bp (178.2bp)	bZIP28(bZIP)/col-bZIP28-DAP-Seq(GSE60143)/Homer(0.918)
11		1e-29	-6.752e+01	43.50%	24.83%	202.6bp (182.4bp)	ESE3(AP2EREBP)/col-ESE3-DAP-Seq(GSE60143)/Homer(0.793)
12		1e-29	-6.737e+01	22.30%	8.60%	167.9bp (168.1bp)	AT1G76870(Trihelix)/col-AT1G76870-DAP-Seq(GSE60143)/Homer(0.788)
13		1e-23	-5.422e+01	16.10%	5.68%	168.9bp (165.2bp)	ALFIN1(HD-PHD)/Medicago sativa/AthaMap(0.700)
14		1e-23	-5.399e+01	43.60%	26.81%	202.1bp (176.9bp)	FUS3(ABI3VP1)/col-FUS3-DAP-Seq(GSE60143)/Homer(0.796)
15		1e-22	-5.256e+01	6.40%	0.82%	158.5bp (130.7bp)	TRP2(MYBrelated)/colamp-TRP2-DAP-Seq(GSE60143)/Homer(0.647)
16		1e-22	-5.208e+01	40.20%	24.12%	190.4bp (177.8bp)	At5g47390(MYBrelated)/col-At5g47390-DAP-Seq(GSE60143)/Homer(0.720)
17		1e-19	-4.498e+01	46.90%	31.23%	185.3bp (172.8bp)	TGA9(bZIP)/colamp-TGA9-DAP-Seq(GSE60143)/Homer(0.728)
18		1e-16	-3.911e+01	19.30%	9.20%	205.4bp (164.2bp)	POL008.1_DCE_S_I/Jaspar(0.650)
19		1e-12	-2.958e+01	39.30%	27.24%	186.6bp (173.7bp)	POL007.1_BREd/Jaspar(0.699)
20 *		1e-11	-2.640e+01	8.40%	3.10%	191.4bp (184.2bp)	FHY3(FAR1)/Arabidopsis-FHY3-ChIP-Seq(GSE30711)/Homer(0.730)

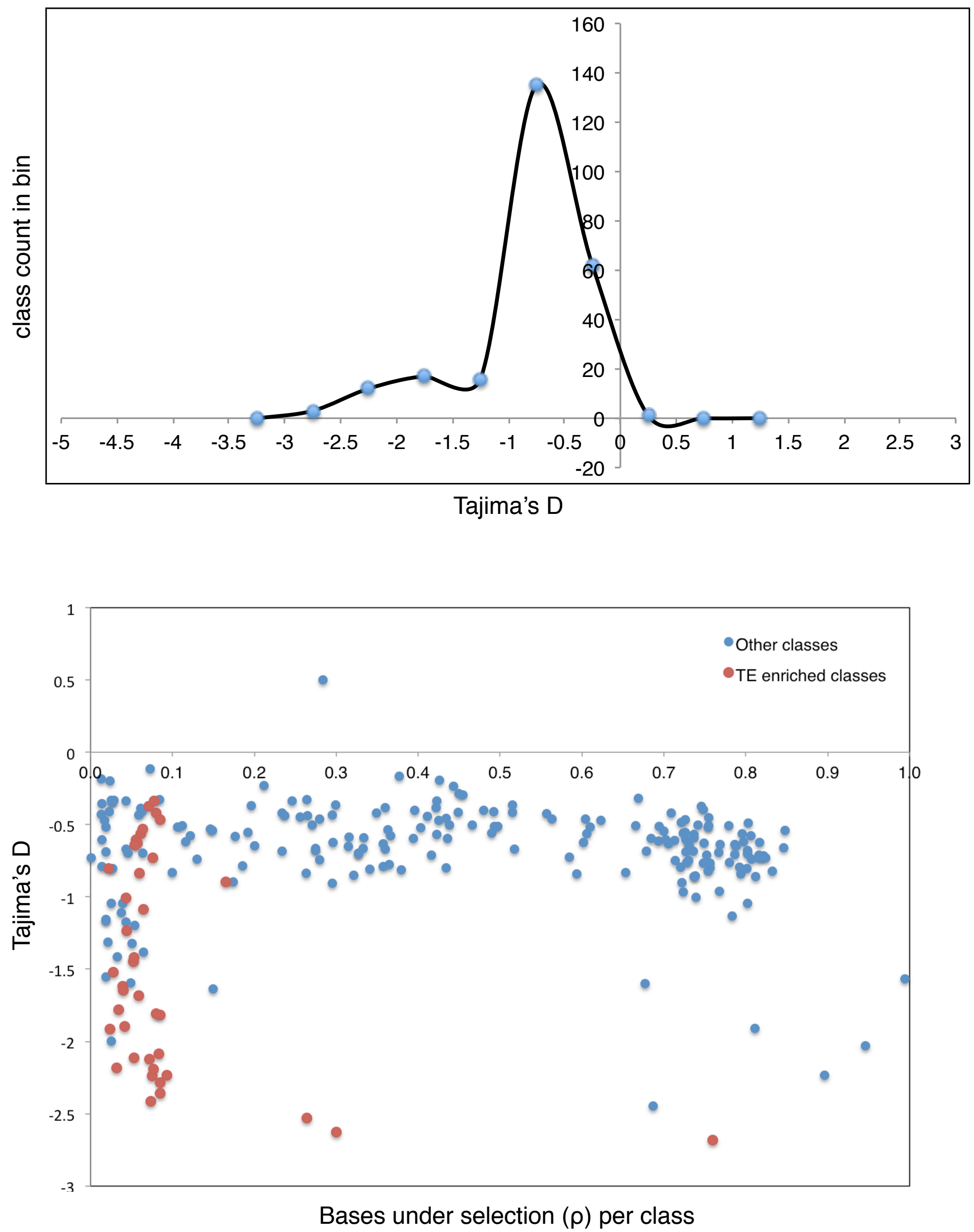
Supplementary Fig. 10. Motif analysis of the Enhancer Candidates classes (sites n = 1,000, length = 0.48Mb). Top 20 de novo motifs generated by the software Homer in which p-values and enrichment are modeled relative to a modified cumulative hypergeometric distribution (for more information see: http://homer.ucsd.edu/homer/motif/). Red stars indicate possible false positive motifs.



Supplementary Fig. 11. Outline of the greenINSIGHT pipeline. Each single-line box represents a command/script. See Methods for more details.

a

b



Supplementary Fig. 12. Tajima's D for all fitCons classes. a, Distribution of Tajima's D across fitcons classes. **b**, Tajima's D versus ρ.