nature plants

ARTICLES

https://doi.org/10.1038/s41477-019-0589-3

An inferred fitness consequence map of the rice genome

Zoé Joly-Lopez[1], Adrian E. Platts[1,2], Brad Gulko[2], Jae Young Choi[1], Simon C. Groen[1], Xuehua Zhong[3], Adam Siepel[2] and Michael D. Purugganan[1,4*]

The extent to which sequence variation impacts plant fitness is poorly understood. High-resolution maps detailing the constraint acting on the genome, especially in regulatory sites, would be beneficial as functional annotation of noncoding sequences remains sparse. Here, we present a fitness consequence (fitCons) map for rice (*Oryza sativa*). We inferred fitCons scores (ρ) for 246 inferred genome classes derived from nine functional genomic and epigenomic datasets, including chromatin accessibility, messenger RNA/small RNA transcription, DNA methylation, histone modifications and engaged RNA polymerase activity. These were integrated with genome-wide polymorphism and divergence data from 1,477 rice accessions and 11 reference genome sequences in the Oryzeae. We found ρ to be multimodal, with ~9% of the rice genome falling into classes where more than half of the bases would probably have a fitness consequence if mutated. Around 2% of the rice genome showed evidence of weak negative selection, frequently at candidate regulatory sites, including a novel set of 1,000 potentially active enhancer elements. This fitCons map provides perspective on the evolutionary forces associated with genome diversity, aids in genome annotation and can guide crop breeding programs.

Determining the likely impact of sequence variation in genomes, particularly at noncoding sites, continues to be problematic, in part because functional annotation is generally sparse. Nonetheless, high-resolution maps of sequence constraint that reveal both functional coding and regulatory sites would benefit crop breeding and genetic engineering interventions. Several approaches to estimating selective constraint from sequence conservation across species have been developed, and more recently complemented by approaches that utilize large population genomic datasets[1,2]. While conservation-based approaches frequently provide high spatial resolution, and population-based approaches can be used to infer recent changes in constraint, neither class of approaches has until recently been able to simultaneously provide both perspectives.

One contemporary approach to determining recent selection on genome sequences at high resolution is inference of natural selection from interspersed genomically coherent elements (INSIGHT)[3], which infers the fraction of nucleotide sites under selection. This is accomplished by comparing patterns of within-species sequence polymorphism with between-species divergence across dispersed genomic sites, relative to nearby neutrally evolving sites[3]. The shorter evolutionary time scales associated with intraspecies variation make this approach more robust to evolutionary turnover, and its applicability to short sequence domains (for example, regulatory sites) makes it particularly powerful for surveying the fitness consequences of point mutations in noncoding DNA[4–7]. The use of locally matched rather than global neutral models make the INSIGHT approach similar to the McDonald–Kreitman test[4] in its robustness to confounding factors such as non-equilibrium demography, mutation rate variation and background selection.

With INSIGHT, the influence of natural selection at each site is summarized by ρ—the probability that a mutation at that site will affect fitness. Values of ρ closer to 1 suggest that a larger proportion of sites in a sequence class are under selection[3] compared with neutral regions (for which ρ is closer to 0). INSIGHT additionally quantifies other parameters, including the number of segregating polymorphic sites per kilobase pair under weak negative selection (P_w)[3].

Integrating INSIGHT with functional genomic data, and aggregating regions of the genome using joint patterns across functional genomic marks, allows the development of genomic fitness consequence (fitCons) maps that permit selection to be inferred at a high resolution across the genome[8,9]. By leveraging patterns of polymorphism within species, these maps measure natural selection at shorter time scales than traditional evolutionary conservation methods. Because polymorphic sites are sparse across many genomes, the fitCons approach uses functional genomic data to pool information across putatively functionally similar genomic sites, and therefore define discrete genomic classes for which we can infer ρ and other selection parameters (for example, P_w).

The fitCons approach, with INSIGHT at its core, was first developed for the human genome. Here, we report a fitCons map in a plant genome, using rice (*Oryza sativa*) as a model system. Rice is one of the most important domesticated food crops in the world and is the target of intense effort for crop improvement to advance food security and sustainable agriculture[10]. The rice fitCons map will contribute to a better understanding of selection in this key crop species, and potentially guide the identification of functional components of genome structure, for future breeding efforts.

Results

greenINSIGHT. To produce a rice fitCons map, we adapted the INSIGHT model for plants (greenINSIGHT; http://purugganan-genomebrowser.bio.nyu.edu/greenInsight/) by modifying elements that were human specific and accounting for some aspects of plant

[1]Center for Genomics and Systems Biology, Department of Biology, New York University, New York, NY, USA. [2]Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA. [3]Laboratory of Genetics and Wisconsin Institute for Discovery, University of Wisconsin-Madison, Madison, WI, USA. [4]Center for Genomics and Systems Biology, NYU Abu Dhabi Research Institute, NYU Abu Dhabi, Abu Dhabi, United Arab Emirates. *e-mail: mp132@nyu.edu

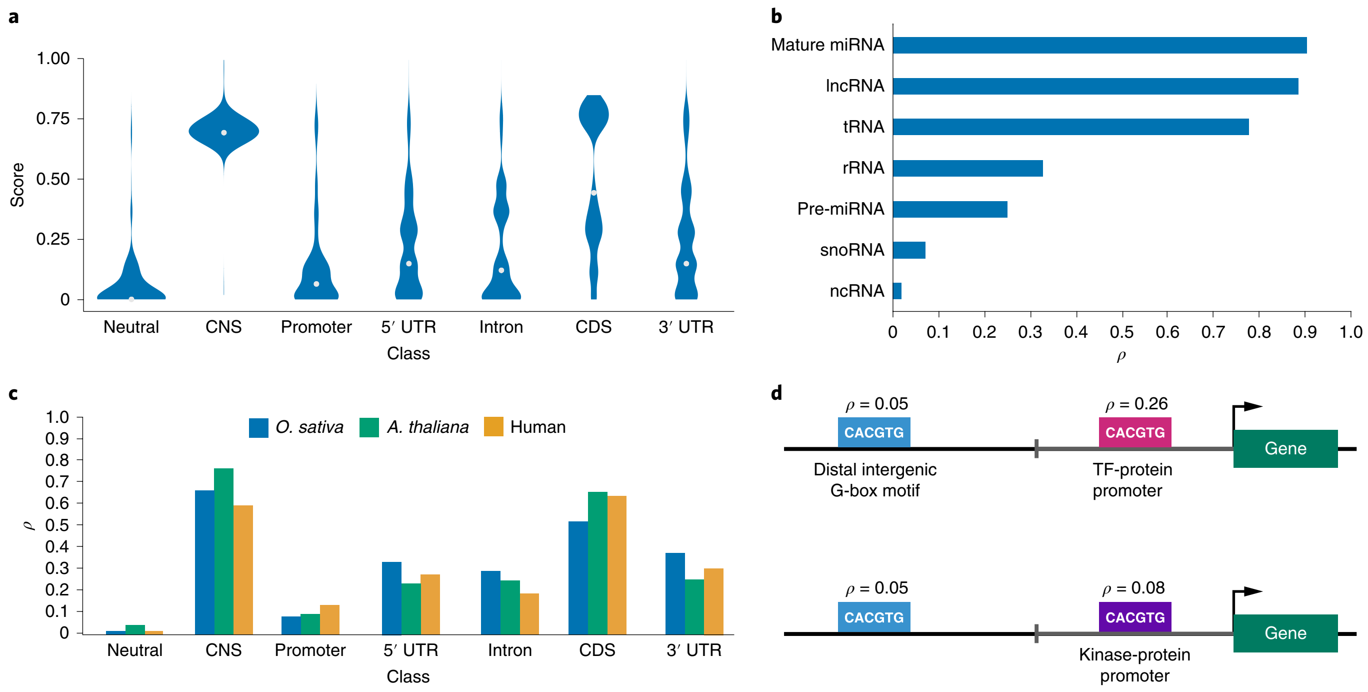


Fig. 1 | greenINSIGHT scores across different genomic annotations in rice. a, Violin plots of the ρ distribution within annotated gene and intergenic site classes. For each annotation type, the approximate size of the feature and minimum, mean and maximum values of ρ for the distribution are: 50.63 Mb, 9.04×10^{-4}, 0.052 and 0.994 (neutral sites); 3.54 Mb, 1.93×10^{-2}, 0.701 and 0.994 (CNS sites); 17.32 Mb, 9.04×10^{-4}, 0.132 and 0.896 (promoter sites); 14.79 Mb, 9.04×10^{-4}, 0.225 and 0.994 (5′ UTR sites); 73.20 Mb, 9.04×10^{-4}, 0.194 and 0.994 (intron sites); 35.91 Mb, 9.04×10^{-4}, 0.498 and 0.848 (CDS sites); and 21.91 Mb, 9.04×10^{-4}, 0.256 and 0.994 (3′ UTR sites). Annotations for neutral and CNS refer to distal neutral and distal CNS sites, respectively. White dots indicate the median ρ for each class. **b**, ρ values across seven types of annotated ncRNA. lncRNA, long ncRNA; rRNA, ribosomal RNA; snoRNA, small nucleolar RNA; tRNA, transfer RNA. **c**, ρ values across annotated gene and intergenic site classes in humans, *A. thaliana* and *O. sativa*. **d**, ρ values at different intergenic locations for the plant G-box motif (CACGTG). TF, transcription factor.

genome organization and recent introgression (see Methods). In essence, the greenINSIGHT pipeline (Supplementary Fig. 11) combines genomic alignments, from which an ancestral base probability distribution is generated, with polymorphism data from the focal species, to infer selection acting at sites of interest relative to a set of matched local neutral sites. Polymorphism data were aggregated from 1,477 *O. sativa* accessions in the 3K Rice Genome Project panel (National Center for Biotechnology Information (NCBI) BioProject accession PRJEB6180)[11] (Supplementary Table 1). Alignments were generated from *O. sativa* to the genomes of ten species in the *Oryza* genus, with the closest outgroup (*Oryza rufipogon*) being the most closely related species to domesticated Asian rice[12], and from which Asian rice was domesticated beginning ~9,000 years ago (Supplementary Fig. 1; see also Methods). We also developed an INSIGHT model for *Arabidopsis thaliana* using an 80-way subpopulation alignment of the *Arabidopsis* 1,001 Genomes dataset[13] (see Methods) and alignments generated previously[14]. The robustness of INSIGHT to complex demography is particularly relevant in rice because assortative mating in the different *O. sativa* variety groups (for example, Japonica and Indica) has led to relatively differentiated populations[3,15,16].

We explored the distribution of ρ reported by greenINSIGHT at a set of well-characterized genomic locations (Fig. 1a). As expected, intergenic regions depleted of open chromatin and distal to genes and conserved noncoding sequences (CNSs) (see Methods) had the lowest ρ (<0.03)[17]. Genomic regions annotated as coding sequences (CDSs) had a significantly higher median ρ, albeit with a broad distribution probably reflective of the considerable variation in constraint across protein domains.

The selection profiles on 5′ and 3′ untranslated regions (UTRs) were similarly complex, again suggestive of a broad range of functional sites. Promoter regions have ρ distributions similar to neutral sites overlaid with a small number of selected sites. Introns may experience low levels of direct selection, but possibly higher levels of intron length-dependent background selection[18,19] due to their linkage with CDSs (Fig. 1a; see also Supplementary Fig. 6). Distal CNSs had the highest ρ values, similar to those of the more constrained CDSs[14].

Sites that generate noncoding RNA (ncRNA) displayed a range of ρ, with sites generating long ncRNA, mature microRNA (miRNA) and transfer/ribosomal RNA showing evidence of more selection than sites generating pre-miRNA, small nucleolar RNA and other unclassified ncRNA (frequently, small interfering RNA) (Fig. 1b). Differences between ρ in humans and plants (*A. thaliana* and *O. sativa*) were mostly subtle, except in introns (Fig. 1c), where the much longer human[20,21] introns probably make background selection less of a confounding factor.

We sought to determine the sensitivity of greenINSIGHT to constraint in regulatory regions by estimating constraint on a well-documented plant transcription factor-binding motif—the G-box motif (CACGTG)[22,23]—in different genomic contexts. This motif is often functional in the promoters of genes targeted by abscisic acid signalling[24], and so unsurprisingly ρ was higher in the promoters of transcription factor genes than in the promoters of genes with more enzymatic roles (Fig. 1d). The proportion of constrained bases in this motif was, as expected, higher when found in promoters than it was in distal intergenic locations.

Genome partitioning to build a fitCons map of rice. To estimate ρ genome wide in rice, we used the fitCons approach to partition the genome into a set of classes[8] based on their shared functional characteristics. Functional genomic datasets were generated from

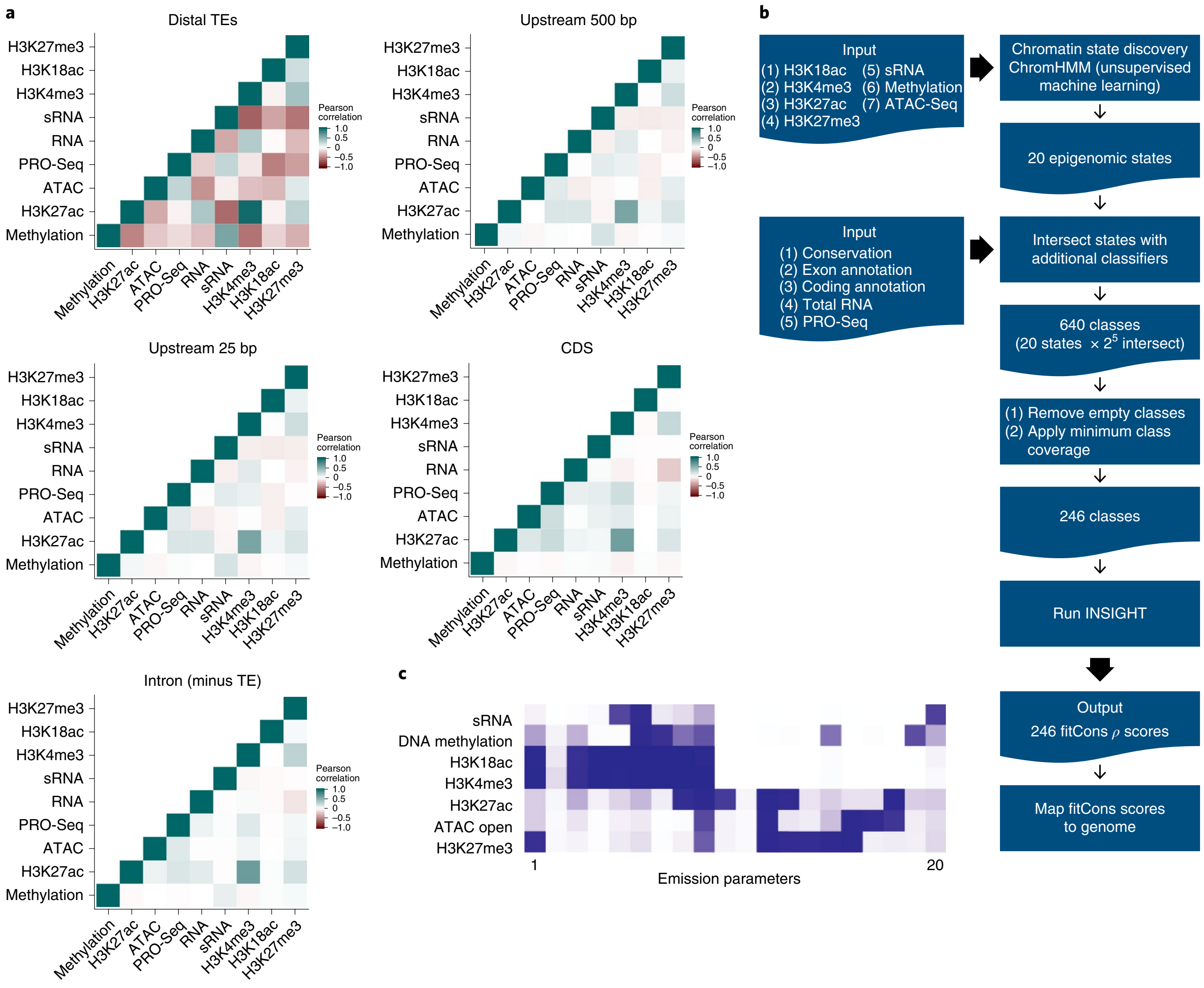


Fig. 2 | Partitioning and scoring the rice genome for selection (ρ). **a**, Pearson's correlation matrices for high-confidence, uniquely mapped reads for histone modification, DNA methylation, open chromatin and transcriptomic datasets in five genomic regions. The numbers of regions (n) for each type are as follows: 165,589 (>1 kb distal transposable elements (TEs)); 27,859 (upstream region <500 bp from the gene translation start site); 30,790 (upstream region <25 bp from the gene translation start site); 130,300 (CDS sites); and 139,269 (intron sites; excluding transposable elements). See also Supplementary Table 3. **b**, Conceptual overview of the analysis pipeline used to generate the fitCons map. See Fig. 3 and Supplementary Table 5 for details on the 246 fitCons scores. **c**, Emission parameters from the 20-state ChromHMM model for seven covariates of chromatin state.

the leaves of 3-week-old *O. sativa* tropical Japonica (cultivar Azucena) plants and included total RNA transcription[25], small RNA (sRNA) transcription[26], assay for transposase-accessible chromatin using sequencing (ATAC-Seq)[27,28], DNA methylation[29], precision nuclear run-on and sequencing (PRO-Seq), which maps transcriptionally engaged polymerase activity at base-pair (bp) resolution[30,31], and H3K27me3, H3K27ac, H3K18ac and H3K4me3 histone modifications (using ChIP-Seq[32,33]) (Supplementary Table 2).

Epigenomic data broadly supported observations of chromatin states reported elsewhere in plants[33–43] (Fig. 2a and Supplementary Table 3). Promoter regions of expressed genes were marked with open chromatin and decreased methylation, while the enhancer mark H3K27ac positively correlated with polymerase activity in distal and proximal gene regions (Fig. 2a). In addition, the distribution of transcriptionally engaged polymerases confirmed 5′ and 3′ polymerase pausing around genes, with the highest signal found in actively expressed genes around transcription start sites[36] (Supplementary Fig. 2). Active gene expression was negatively correlated with the presence of H3K27me3, while transposable elements showed positive correlations between sRNA transcription and DNA methylation, indicative of transposable element silencing. Most sRNA coverage appeared to arise from broadly distributed small interfering RNA, while DNA methylation was mostly found in CpG contexts, with CpG methylation slightly enriched in gene bodies relative to CHG/CHH methylation; however, their distribution was similar in transposable elements, as previously reported[44] (Supplementary Table 4).

The rice genome was partitioned into a set of coherent classes in a two-step process. First, the epigenomic datasets were used to infer a small set of chromatin states. Then, in a second step, these states were intersected with transcriptomic and annotation data (Fig. 2b) to generate a more nuanced global classification.

The first step employed a hidden Markov modeller ChromHMM[45] to binarize chromatin signals and produce a

Class ID	Percentage genome coverage	CDS	Introns	Promoter	End	CNS	ATAC	PRO-Seq	TEs	Class ρ
1	2.851	0.30	0.19	0.31	0.26	0	0.04	0.01	2.66	0.083
4	6.302	0.24	0.21	0.27	0.27	0	0.01	0	2.61	0.085
10	6.312	0.16	1.30	1.08	1.46	0	0.02	0.44	0.41	0.146
11	22.004	0.23	1.06	1.07	1.36	0	0.02	0.21	0.66	0.001
15	2.988	0.66	0.38	0.81	0.55	0	0.08	0.05	1.78	0.065
44	0.100	0.47	0.48	3.23	3.08	0	15.65	13.72	0.46	0.064
48	0.142	0.25	0.35	4.05	3.94	0	15.16	14.95	0.52	0.019
49	0.410	0.06	0.33	3.77	3.57	0	10.47	14.93	0.71	0.023
50	0.182	0.16	0.41	2.75	3.96	0	7.60	17.32	0.58	0.014
52	0.282	0.03	0.40	3.12	2.10	0	2.34	8.11	2.03	0.084
102	0.005	0.04	3.93	0.63	1.38	0	0.61	8.64	2.50	0.061
108	0.031	0.40	3.08	1.95	1.31	0	12.98	17.22	0.23	0.363
112	0.011	0.28	2.02	2.66	2.97	0	16.71	18.24	0.40	0.256
138	0.268	0.80	1.63	0.94	1.21	36.36	0.01	0.44	0.04	0.693
144	0.120	0.81	0.52	3.62	1.26	40.82	13.72	1.14	0.04	0.738
145	0.158	0.54	0.82	2.26	1.50	42.34	10.45	1.19	0.02	0.684
174	0.005	1.29	0.48	1.47	6.19	36.38	0	16.28	0.07	0.700
200	0.017	2.69	0.33	3.50	4.37	3.77	0.25	0	0.09	0.994
201	0.014	1.88	0.42	1.99	2.87	4.96	10.56	0.08	0.14	0.946
234	0.005	1.43	4.65	0.74	1.02	22.67	0	21.87	0.01	0.802
300	0.070	0.36	0.21	1.76	0.86	0	21.79	8.59	0.35	0.416
306	0.043	0.25	0.28	1.80	1.59	0	14.12	11.96	0.41	0.237
466	0.016	0.99	0.93	0.50	1.34	41.39	1.56	11.97	0.02	0.806
782	1.205	7.02	0.18	0.27	0.25	0	0.01	0.31	0.14	0.271
908	0.122	7.70	0.06	0.41	0.28	0	1.46	3.50	0.16	0.768
970	0.330	7.76	0.11	0.07	0.29	0	0	2.78	0.01	0.817
974	0.378	7.77	0.09	0.04	0.26	0	0	1.80	0.02	0.807
977	0.096	7.76	0.10	0.05	0.31	0	0.27	2.98	0.01	0.802
978	0.055	7.74	0.14	0.23	0.32	0	0.51	8.38	0.04	0.797
1,003	0.011	7.75	0.11	0.08	0.80	0	0	17.34	0.01	0.848
1,010	0.023	7.71	0.17	0.31	0.56	0	2.79	24.31	0.06	0.824

Fig. 3 | Properties of a subset of the 246 fitCons genome classes. Genomic coverage and enrichment, with respect to ten annotated regions of the genome, and ρ values are reported for each class. The Promoter is the sequence of <500 bp proximal to the gene site. The End is the sequence of <500 bp distal from the gene end. The colours represent the minimum (dark blue) to maximum (dark red) enrichment across classes relative to each annotation. TEs, transposable elements.

low-resolution map of chromatin states across the rice genome (Fig. 2b,c; see also Methods). The selection of the number of states was informed by the ChromHMM option CompareModels[46] to determine the correlation between the emissions of models having different numbers of states. Selecting a 50-state model as an overparameterized reference, we observed a rapid convergence towards this model's outputs after 15 states were incorporated, and the mean state correlation with the closest state in the 50-state model exceeded 0.9 once 20 states were included (Supplementary Fig. 3). We therefore selected a 20-state ChromHMM model (Fig. 2b,c). This number is higher than early estimates of the number of chromatin states in plants[47,48] but fewer than the 38 states previously inferred in rice from a broader set of histone marks[32]. The inferred number of chromatin states is anticipated to vary to some degree with the type of chromatin marks used as input, and to this end we selected marks based on testing for high levels of intra-replicate correlation but low levels of correlation between marks in *A. thaliana* public datasets.

In a second step, we intersected these chromatin states with further binarized annotation data and evidence for roles in transcription or transcription initiation. These data included reference genome annotation (coding and exon), phastCons scores, and RNA-Seq and PRO-Seq alignments (Fig. 2b; see also Methods). The intersection of the 20 states with these data generated a more ontologically complete and higher-resolution set of 640 possible genome classes[8], of which 246 were identified with an appreciable coverage of the rice genome (>20 kilobases (kb) of total sequence) (Fig. 2b and Supplementary Table 5).

With the rice genome partitioned into 246 genomic classes (fitCons classes), we estimated ρ for each class using greenINSIGHT, as previously described[8] (Fig. 3 and Supplementary Tables 5 and 6). The 246 class ρ scores were then distributed back to each nucleotide in each class, giving each nucleotide in the genome a fitCons score (Fig. 4a). A simple validation of class coherence was performed to ensure that the distribution of ρ as a function of class size was different from that expected under a random sampling model (Supplementary Fig. 4 and Supplementary Table 7; see also Methods).

Distribution of fitCons scores. The 246 rice fitCons classes we inferred were distributed in ~4.3 million blocks ranging in size from 1–600 bp, with most in blocks of 10–40 bp (Fig. 4a,b). We found a multimodal distribution of ρ over the 246 states, with peaks at $\rho =$ ~0.08, ~0.44 and ~0.76. Most of the genome comprises classes with low-to-moderate ρ (86.4% of the genome has $\rho < 0.4$), while higher ρ classes ($\rho > 0.5$) make up only 8.98% of the genome (Fig. 4c). The cumulative distribution of ρ in rice is consistent with a similar number of coding sites in a small genome space relative to humans, and more intermediate selection on noncoding functional sites, some of which may arise from background selection (Fig. 4d).

Relative to the genome's broader annotation, classes with low-to-moderate selection were primarily located in unannotated intergenic regions, or were enriched for transposable elements (Fig. 4c and Supplementary Table 5). In contrast, genome classes with higher ρ were enriched for CDSs and CNSs, with some overlapping regions with open chromatin, and/or actively transcribed sites. Classes that had intermediate values of ρ were enriched for a mixture of genomic annotations, and it was hard to identify any predominant constituents for many of these classes.

As expected, many classes with higher ρ were enriched for known functional elements. For example, classes 974, 977 and 1,003 (median: $\rho = 0.817$) were associated with CDSs, whereas classes 138–146 (median: $\rho = 0.694$) and 170–178 (median: $\rho = 0.7$) were associated with a set of inferred CNSs (Fig. 3 and Supplementary Table 5; see also Methods). Several classes were associated with different transposable element types; for instance, classes 15 ($\rho = 0.065$) and 4 ($\rho = 0.085$) were enriched for mutator-like transposable elements and gypsy long terminal repeat retrotransposons, respectively (Fig. 4e,f).

Several of these high-ρ classes appeared to be enriched for facultative regulatory sites, showing small but significant correlations with the expression of downstream genes (for example, class 17 ($r = 0.113$; two-tailed t-test; $P = 5.2 \times 10^{-70}$; $n = 24{,}296$) and class 145 ($r = 0.08$; $P = 4.5 \times 10^{-32}$; $n = 24{,}296$); Supplementary Table 8). Combining all classes, a multiple regression model trained on chromatin classes upstream (500 bp) of genes on chromosomes 2–12 had modest but significant predictive power for gene expression when tested on chromosome 1 genes ($r = 0.419$; two-tailed t-test; $P = 3.6 \times 10^{-148}$; $n = 3{,}502$) (Fig. 5). Predictive power arose primarily from epigenomic states around slightly distal promoter sequences, as masking the 50-bp core promoter did not significantly impact model power. Notable for 3′-to-5′ looping models of gene regulation[49], downstream gene regions were enriched for a different set of classes with equally high individual associations with gene expression (for example, class 49 ($r = 0.18$; two-tailed t-test; $P = 1.0 \times 10^{-177}$; $n = 24{,}296$) and class 50 ($r = 0.181$; $P = 7.2 \times 10^{-179}$; $n = 24{,}296$), but that had less power as a combined model (Supplementary Fig. 5).

As expected, the site-frequency spectrum of polymorphisms in rice displays a minor allele frequency (MAF) skew towards rare variants in high-ρ classes, such as those enriched in the more conserved promoters (for example, class 145; $\rho = 0.681$) and CDSs (for example, class 782; $\rho = 0.294$) (Fig. 4g and Supplementary Table 9) relative to low-ρ classes, such as the transposable element-enriched class 11 ($\rho = 0.016$). This skew was evident across classes in general ($r = 0.83$; two-tailed t-test; $P = 8 \times 10^{-64}$; $n = 246$) (Fig. 4h and Supplementary Table 9).

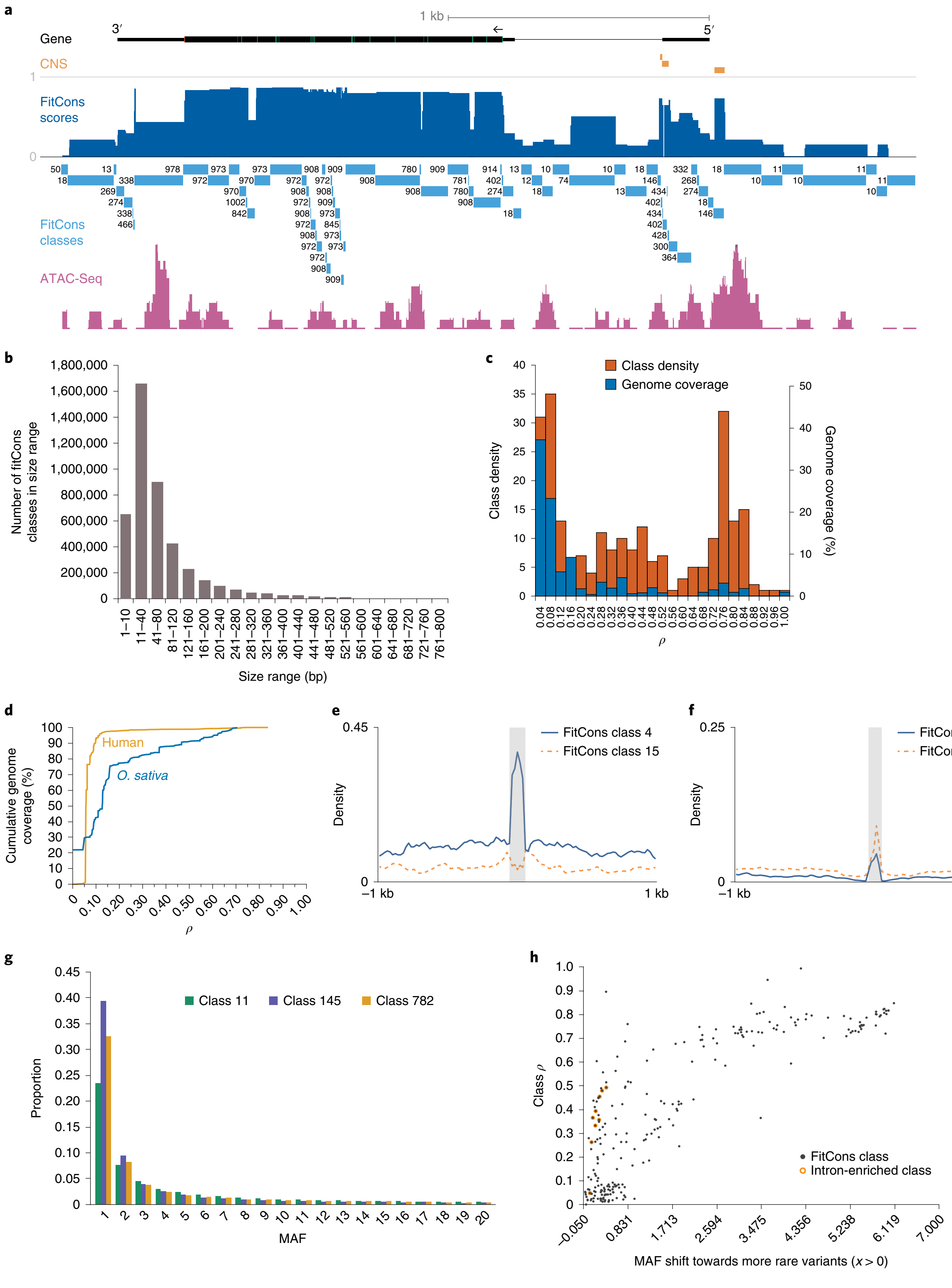
Overall, and as expected, strong negative selection was the main driver of high values of ρ at sites such as CNS and CDSs. However, about 2% of the rice genome appeared to be under weak negative selection; classes with lower ρ sometimes had notable levels (up to 8%) of sites under weak negative selection that potentially mark recently selected genomic sites (Supplementary Table 5).

Delineating putative noncoding regulatory regions. Among the 246 fitCons classes, we further defined three categories of potentially functional noncoding regions by considering their functional and epigenomic characteristics, as well as their ρ scores. The first category, termed conserved classes, includes 46 noncoding element classes with evidence of sequence conservation among *Oryza* species (phastCons > 0.82) (Supplementary Table 5). Conserved classes have high ρ and are prevalent in the promoter regions of a subset of protein-coding genes (within 0–350 bp upstream of transcription start sites) where they probably act as *cis*-regulatory elements[14,50] (Fig. 6a and Supplementary Fig. 7a). Gene Ontology enrichment analysis suggests that these are predominantly associated with transcription factor and developmental genes (false discovery rate = 1.98×10^{-40}) (Supplementary Table 10). The density of upstream conserved classes was strongly linked to genes with the highest fold change in expression between tissues (Fig. 6b), again suggesting tissue-specific or developmental roles. De novo motif analysis using Homer (N: 6–8 bp) of these 46 classes revealed generally complex motifs (Fig. 6c and Supplementary Fig. 7b,c), including well-characterized transcription factor-binding sites (for example, G-box, RY-repeat motifs, TATA box, and so on; see Supplementary Fig. 8)[22,23,51].

The second category comprises 17 classes, termed open chromatin classes, which have a broader range of ρ but have ATAC-Seq signals that are at greater than or equal to tenfold above background (Supplementary Table 5 and Supplementary Fig. 7a). While their ρ tends to be similar to ρ for UTRs and promoters (median: $\rho = 0.256$), this ranges from $\rho = 0.013$ (class 305) to $\rho = 0.946$ (class 201). These open chromatin classes are associated with stable gene expression (lowest fold change) profiles across multiple tissues (Fig. 6d) and are often enriched for simple tandem repeat motifs (Fig. 6c and Supplementary Figs. 7b,c and 9).

Fig. 4 | Distribution of ρ across the rice genome. a, Partitioning of a locus by genomic class and associated ρ, as illustrated in the rice locus *Os07g0153400*. A high-ρ class immediately upstream of the transcription start site coincides with both open chromatin and sequence conservation, while the short 3′ UTR peak lies immediately downstream of two canonical polyA signals, potentially coinciding with paused polymerases. **b**, Size distribution of the blocks that make up the 246 genomic classes used to partition the genome. **c**, Distribution of class ρ values across the rice genome. Most of the genome (right axis; percentage genome coverage) is contained in large, low-ρ classes (left axis; class density), contrasting with the much smaller high-ρ classes. **d**, Cumulative distribution of ρ in the 3.2-Gb human genome relative to the 0.4-Gb rice genome. **e**,**f**, Density of specific genome classes around type I (gypsy retrotransposons (**e**), shaded grey) and type II (mutator-like transposable elements (**f**)), shaded grey) transposable elements, in 50-bp windows. **g**, Standardized MAFs of rare single-nucleotide polymorphisms (subset of MAF distribution *f*:1–20) for three fitCons classes. Those with higher ρ (class 145 (0.69) > class 782 (0.27) > class 11 (0.001)) show an expected bias towards rare single-nucleotide polymorphisms (Supplementary Table 9). There is a MAF shift (x) towards rarer variants in each class relative to the control neutral class (class 11). **h**, As expected, this shift increases with class ρ, with the exception of classes that experience high levels of indirect rather than direct selection (a subset of such intron-enriched classes are highlighted).

The third category includes 11 classes enriched for intergenic bidirectional divergent PRO-Seq signals (Fig. 6e). These signals are often characteristic of mammalian enhancer RNA[52,53], but were also recently suggested in plants[36]. Using dREG[53] to identify enhancer RNA signals from PRO-Seq data, we found 1,000 high-scoring (>1.0) dREG sites in regions >1 kb from genes, suggesting that these



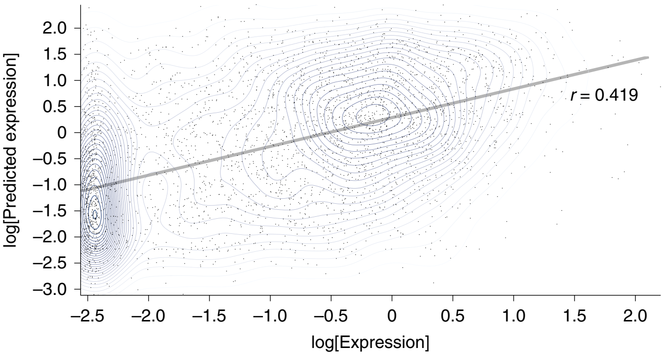


Fig. 5 | Proximal upstream chromatin class distribution correlates with downstream gene expression. Actual (*x* axis) and predicted (*y* axis) leaf tissue gene expression per class coverage on chromosome 1. The grey line indicates the best-fit line ($n=3,502$; Pearson's coefficient of determination, $r^2=0.18$). The predicted expression is derived from a multiple linear regression model of upstream chromatin classes used as covariates against gene expression for protein-coding genes on rice chromosomes 2–12.

sites form a set of putative rice enhancer elements (Supplementary Table 11; see also Methods). The dREG locations shared other enhancer-type characteristics, including moderate enrichment for open chromatin (~7.17-fold) (Fig. 6f), asymmetrically co-located H3K27ac marks (Fig. 6g) and enrichment for motifs similar to those found in open chromatin classes (Fig. 6c and Supplementary Fig. 10). As expected, these 11 classes (for example, classes 44 and 48; Fig. 3) had a greater than tenfold enrichment for these high-scoring dREG sites (Supplementary Fig. 7a and Supplementary Table 5), and individual dREG sites were generally found to overlap several of these 11 classes. However, it was rare for a dREG site to be homogenous across its length with respect to a single fitCons class (Supplementary Table 11).

These 11 classes, termed enhancer candidates, have weak correlations with the expression of nearby genes (for example, class 43 ($r=0.03$; two-tailed t-test; $P=1\times10^{-5}$; $n=24,296$)) (Supplementary Table 8) consistent with the majority of candidates being in a poised but inactive state. They are also associated with low ρ (<0.2) (Fig. 6h) and low phastCons conservation (Fig. 6i), but have a greater than twofold excess of sites under weak negative selection (Fig. 6j). The association with weak negative selection was also observed for dREG sites detected in human populations[54]. Taken together, this may suggest that emergent negative selection, consistent with rapid enhancer turnover, has recently acted on these classes within *O. sativa*.

Discussion

The INSIGHT and fitCons approaches provide a set of potentially powerful methods for identifying selective constraint on a genome-wide scale. INSIGHT has been used to identify different types of enhancers, such as exonic splicing enhancers in humans[55], shadow enhancers in *Drosophila*[56] and novel motifs such as the Coordinator motif found within human cranial neural crest cell-specific enhancers[57]. In the human genome, fitCons maps were revealed to have higher sensitivity than other methods for locating multiple types of functional noncoding elements with putative roles in transcriptional regulation[8].

Our fitCons map for rice provides a catalogue of putative functional sites that can allow patterns of selection between different genes, genomic regions and genetic pathways to be explored. The map we have developed has limitations; for example, since every base within a fitCons class will receive the same fitCons score (same ρ), we cannot determine which specific bases within a class are under selection. The validation of fitCons scores relative to the actual fitness impacts of mutations remains to be tested. Subsequent experimental analyses will help shed light on the function of the candidate noncoding regulatory elements that we have described in this study.

Nevertheless, some of the broad features of the distribution of ρ in the rice genome confirm what we know about the biology of specific genome elements, suggesting that the inferred ρ values are related to underlying biological features. Integrating evolutionary information with functional genomic and epigenomic data permits identification of important regulatory components of crop genomes[10], improving genome annotation and helping to guide molecular genetic studies. fitCons maps can also help in genetic mapping and crop breeding efforts, including the identification of candidate deleterious mutations[58] that can be targeted for removal in next-generation breeding efforts[59–61]. As more such maps are produced, it will be possible to undertake comparative analyses of genomic selection across species and help develop more precise genomic breeding programs. To this end, a web interface for greenINSIGHT is available at http://purugganan-genomebrowser.bio.nyu.edu/greenInsight/ and all functional genomic tracks, ρ scores and fitCons classes for rice can be viewed or downloaded from a custom genome browser (http://purugganan-genomebrowser.bio.nyu.edu/cgi-bin/hgTracks?db=Osaj&position=Osaj.1%3A166356-178595).

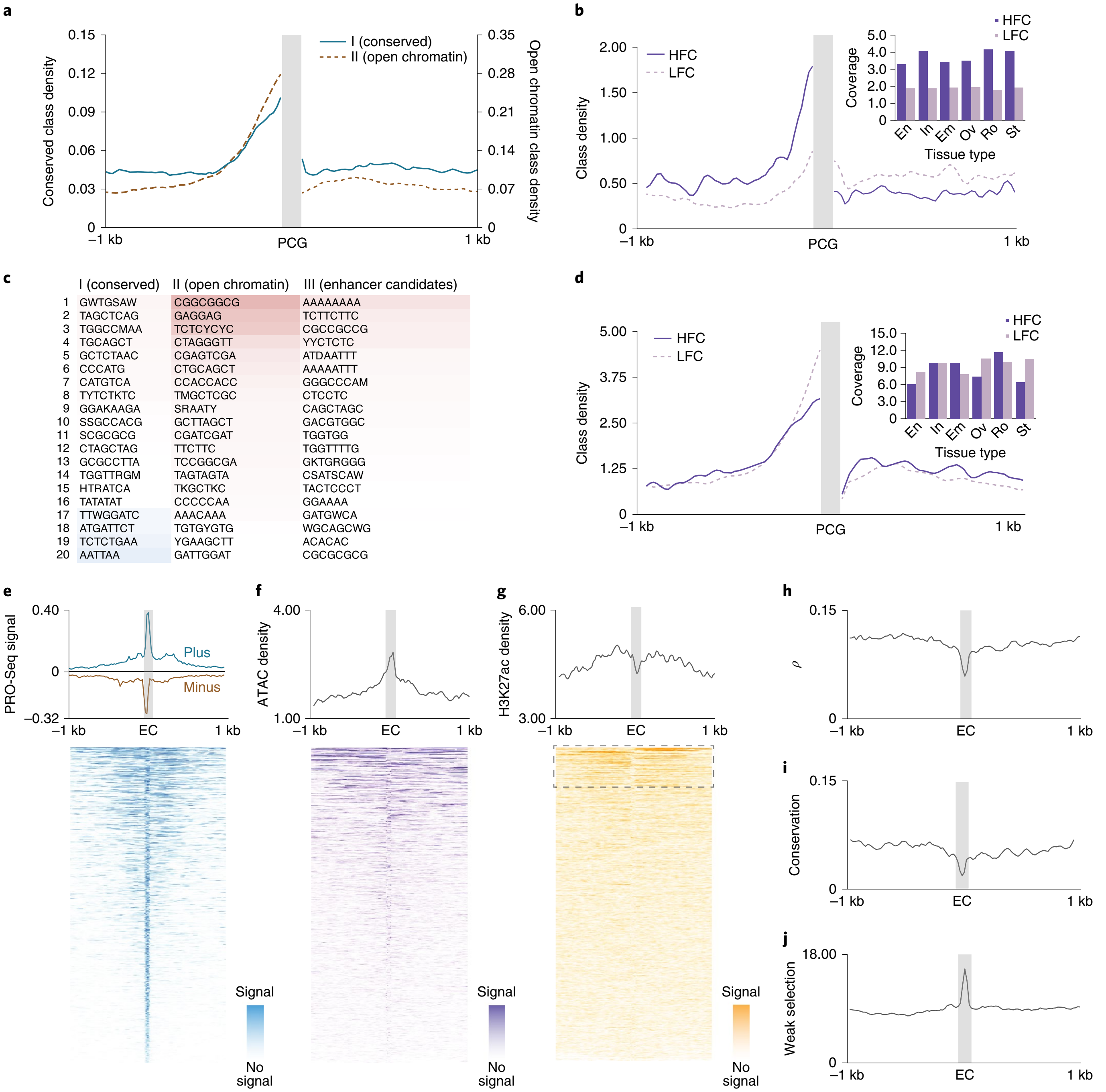
Methods

Plant material. Seeds of *O. sativa* landrace Azucena (IRGC 328; tropical Japonica), provided by the International Rice Research Institute (Los Baños, Philippines), were used for the functional genomic analyses. Seeds were incubated for 5 d at 50° C and germinated in water in the dark for 48 h at 30° C. These were subsequently sown on hydroponic pots suspended in 1× Peters solution and 1.8 mM $FeSO_4$ (pH = 5.1–5.8) (JR Peters). Plants were grown for 15 d in growth chambers (12-h days; 30 °C/20 °C day/night; 300–500 μmol quanta $m^{-2} s^{-1}$; relative humidity: 50–70%). Leaf tissue for library construction was collected from 17-d-old, young plants.

RNA-Seq. Total RNA was extracted using RNeasy Plant Mini kits (Qiagen), according to the manufacturer's instructions. RNA quality was determined by BioAnalyzer (Agilent). Contaminating DNA was removed from total RNA samples with Baseline-ZERO DNase (Epicentre), whereas ribosomal RNA was removed using a Ribo-Zero rRNA Depletion Kit (Epicentre). Strand-specific RNA-Seq libraries were synthesized using a Plant Leaf ScriptSeq Complete Kit (Epicentre). Three biological replicates were generated. Libraries were sequenced using Illumina protocols for 2 × 100-bp reads on an Illumina HiSeq 2500. Total RNA-Seq data were used to identify expressed regions rather than to quantify expression (for expression quantification, see expression correlation analysis below). Reads were 3′ trimmed for quality ($q<20$) and adapter sequences (Cutadapt 1.11; ref. [62]), and read pairs for which either end was shorter than 25 bp after trimming were rejected. Trimmed reads were aligned to the soft-masked IRGSP1.0/MSU7 build of the rice genome (downloaded from Ensembl) using Bowtie 2 (v.2.2.9)[63] with the option -sensitive-local. Alignments were converted to bam format (SAMtools 1.3), sorted and converted to bedGraph alignment format with bedtools v.2.25 (ref. [64]). Reads were subsequently converted to bigWig format (Kent tools; University of California, Santa Cruz (UCSC)[65]) for visualization in a custom UCSC Rice genome browser. A single sample was selected as input in the fitCons approach; the choice of replicate was based on signal strength and sequencing coverage. Note that for all subsequent analyses (sRNA-Seq, methylation, ATAC-Seq, CHIP-Seq and PRO-Seq), the alignment and visualization protocols were the same, with some modifications.

sRNA-Seq. For extraction of sRNA, total RNA was first isolated using Ambion Plant RNA Isolation Aid (Thermo Fisher Scientific) and sRNA subsequently extracted using a mirVana miRNA Isolation Kit (Thermo Fisher Scientific). A total of 35–70 ng sRNA was used to generate libraries using a TruSeq Small RNA Library Prep Kit (Illumina). Three biological replicates were generated. Libraries were sequenced using Illumina protocols for 2 × 50-bp reads on an Illumina HiSeq 2500. Reads were 3′ trimmed for quality ($q<20$) and adapter sequences (Cutadapt 1.11), and only reads longer than 15 bp were retained (-m 16). Trimmed reads were aligned to the IRGSP1.0/MSU7 build using Bowtie 2 (v.2.2.9) with the options -end-to-end, -sensitive and -k 100. To infer initial low-resolution chromatin states, the signal was averaged over 40-nucleotide blocks for input into ChromHMM[45].

DNA methylation. DNA was extracted using DNeasy Plant Mini kits (Qiagen) following the manufacturer's protocol. Extracted DNA was sheared into 350-bp fragments using an S220 Focused-ultrasonicator (Covaris). An Illumina TruSeq



	I (conserved)	II (open chromatin)	III (enhancer candidates)
1	GWTGSAW	CGGCGGCG	AAAAAAAA
2	TAGCTCAG	GAGGAG	TCTTCTTC
3	TGGCCMAA	TCTCYCYC	CGCCGCCG
4	TGCAGCT	CTAGGGTT	YYCTCTC
5	GCTCTAAC	CGAGTCGA	ATDAATTT
6	CCCATG	CTGCAGCT	AAAAATTT
7	CATGTCA	CCACCACC	GGGCCCAM
8	TYTCTKTC	TMGCTCGC	CTCCTC
9	GGAKAAGA	SRAATY	CAGCTAGC
10	SSGCCACG	GCTTAGCT	GACGTGGC
11	SCGCGCG	CGATCGAT	TGGTGG
12	CTAGCTAG	TTCTTC	TGGTTTTG
13	GCGCCTTA	TCCGGCGA	GKTGRGGG
14	TGGTTRGM	TAGTAGTA	CSATSCAW
15	HTRATCA	TKGCTKC	TACTCCCT
16	TATATAT	CCCCCAA	GGAAAA
17	TTWGGATC	AAACAAA	GATGWCA
18	ATGATTCT	TGTGYGTG	WGCAGCWG
19	TCTCTGAA	YGAAGCTT	ACACAC
20	AATTAA	GATTGGAT	CGCGCGCG

Fig. 6 | Characterization of three categories of intergenic fitCons classes. **a**, The densities of higher ρ/conserved classes (category I; solid blue line) and lower ρ/open chromatin classes (category II; red dashed line) around protein-coding genes (PCGs) (grey box) suggest a rice promoter with a mean size of ~320 bp. **b**, Regions upstream of genes (grey box) with a high fold change (HFC; purple line) across tissues are enriched for blocks of conserved classes relative to regions upstream of genes with stable expression across tissues (lowest fold change (LFC; dotted line)). A breakdown by tissue (inset) suggests that regions upstream of genes with differential expression relative to leaf (En, endosperm; In, inflorescence; Em, embryo; Ov, ovary; Ro, root; St, stamen tissues) consistently show greater promoter coverage with blocks of more conserved classes. **c**, Motif (6–8 bp) enrichment (blue, low; red, high) differs between the three categories of noncoding classes. Open chromatin classes are relatively enriched for simple repetitive motifs similar to those found in enhancer candidate regions. **d**, Genes whose upstream regions are enriched for open chromatin classes rather than conserved classes show broader activation across tissues, but no similar enrichment for differential expression across tissues. **e**–**g**, Density plots ±1 kb around the 1,000 enhancer candidate (EC) sites show a defining bidirectional diverged PRO-Seq signal (blue, plus strand; brown, minus strand; arbitrary strands) identified by dReg[53] (**e**), a marked enrichment for open chromatin (**f**) and a generally asymmetric H3K27ac location beyond the nucleosome-depleted core (as indicated by the dashed rectangle) (**g**). **h**–**j**, EC sites are also associated with low ρ (**h**), low conservation (**i**) and a twofold excess of weak negative selection (P_w) (**j**).

DNA Kit (catalogue number FC-121-3001) was used to construct the library and a Zymo Lightning Kit (catalogue number D5030) was used to perform the bisulfite treatment. KAPA Uracil Polymerase (catalogue number KK2623) was used to amplify the library with 12 cycles. Two biological replicates were generated. Libraries were sequenced using Illumina protocols for 2 × 100-bp reads on an Illumina HiSeq 2500. Reads were treated as two independent sets of forward and reverse reads, and aligned using a basic pipeline suitable for plant CG/CHG/CHH methylation. Reads were pre-processed in silico by trimming bisulfite sequencing adapters, after which unconverted (methylated) cytosine bases were converted to thymine bases. Reads were then aligned against both an A–G–T-transformed IRGSP1.0/MSU7 genome (in which all cytosine bases had been replaced by thymine bases) and an A–C–T-transformed genome (in which all guanine bases had been replaced by adenine bases) using Bowtie 2 (v.2.2.9) with the options -end-to-end, -sensitive and -k 1. Finally, alignments were processed to combine their location with offsets within the alignment of unconverted (methylated) cytosines. Refined locations were accumulated in a bedGraph file and subsequently converted to bigWig format (Kent tools; UCSC). For the purpose of inferring initial low-resolution chromatin states, the methylation signal was averaged over 40-nucleotide blocks for input into ChromHMM.

Chromatin accessibility. A dataset previously generated from leaf tissue of the same developmental stage in the Azucena background using the ATAC protocol[38] was used (SRR2981235, SRR2981233, SRR2981221, SRR2981227, SRR2981231 and SRR2981234). Reads were 3′ trimmed for quality ($q < 20$) and adapters, and were aligned with NovoAlign (v.3.04.04; novocraft.com) with the options -t 80 -a -i PE 500,400 -o SAM. Alignments were converted to bam format (SAMtools 1.3), sorted and converted to bedGraph alignment with bedtools v.2.25, followed by conversion to bigWig format (Kent tools; UCSC) for visualization in a custom UCSC Rice genome browser. ATAC peaks were called with MACS2 (v.2.1.1)[66], with genome size, q threshold and ATAC recommended parameters[67] '-g 4.0e8 -q 0.025–nomodel–shift −100–extsize 200 -B'. Signal was averaged over 40-bp blocks for input into ChromHMM.

ChIP-Seq. Leaf tissue (2 g) was fixed in 1% formaldehyde (*v/v*) for 15 min, after which glycine was added to a final concentration of 125 mM (5 min incubation). Tissues were rinsed three times with cold, de-ionized water before being flash frozen in liquid nitrogen. Chromatin extraction and chromatin shearing were performed using a Universal Plant ChIP-seq kit (Diagenode) following the manufacturer's instructions. Protease inhibitor cocktail (MilliporeSigma) was added to extraction buffer. Samples were sonicated for 4 min on a 30 s on/30 s off cycle using a Bioruptor Pico (Diagenode). Subsequent steps were performed as in the Universal Plant ChIP-seq kit protocol. Immunoprecipitation was done using anti-acetyl-histone H3 (Lys27) (H3K27ac; Cell Signaling Technology; catalogue number 4353S; lot 1), anti-trimethyl-histone H3 (Lys27) (H3K27me3; MilliporeSigma; catalogue number 07-449; lot 2919706), anti-trimethyl-histone H3 (Lys4) (H3K4me3; EMD Millipore; catalogue number 07-473; lot 2746331) and anti-acetyl-histone H3 (Lys18) (H3K18ac; Cell Signaling Technology; catalogue number 9675S; lot 1). The quality and fragment size of immunoprecipitated DNA and input samples were measured using agarose gel electrophoresis and TapeStation 2200 (Agilent). Three biological replicates were generated. Libraries were synthesized using a MicroPlex Library Preparation Kit (v.2; Diagenode). Libraries were sequenced as 2 × 50-bp reads on an Illumina HiSeq 2500 instrument. Reads were 3′ trimmed for quality ($q < 20$) and adapter sequences (Cutadapt 1.11), and read pairs for which either end was shorter than 16 bp after trimming were rejected. Trimmed reads were aligned to the soft-masked IRGSP1.0/MSU7 build of the rice genome (downloaded from Ensembl) using Bowtie 2 (v.2.2.9) with the options -end-to-end and -sensitive. ChIP-Seq peak calling was performed using MACS2, with input DNA used as a control and the additional parameters -g 4.0e8–bw 200 -B -m 3 50. For the purpose of inferring initial low-resolution chromatin states, signal was averaged over 40-bp blocks for input into ChromHMM.

PRO-Seq. *Nuclei isolation.* Nuclei isolation was as described by Hetzel et al.[68], with modifications. Briefly, ~20 g of leaf tissue from 17-d-old plants was collected in 4 °C, placed in ice-cold grinding buffer and homogenized using a Qiagen TissueRuptor. Samples were filtered and pellets were washed twice, followed by homogenization, resuspension in storage buffer (10 mM Tris (pH 8.0), 5 mM $MgCl_2$, 0.1 mM EDTA, 25% (*v/v*) glycerol and 5 mM DTT) and freezing in liquid N_2.

Nuclei sorting. Nuclei were stained with DAPI and loaded into a flow cytometer (Becton Dickinson FACSAria II). Around 15 million nuclei were sorted based on the size and strength of the DAPI signal, and subsequently collected in storage buffer. Nuclei were pelleted by centrifugation at 5,000*g* and 4 °C for 10 min, and resuspended in 100 µl storage buffer.

PRO-Seq library preparation. PRO-Seq was performed as described by Mahat et al.[30], generating strand-specific libraries with reads starting from the 3′ end of the RNA. Two biological replicates were generated. Amplified libraries were assessed for quality on a TapeStation before sequencing with 1 × 50-bp reads on a HiSeq 2500. Reads were trimmed and carefully aligned against the soft-masked IRGSP1.0/MSU7 build and supplemented with chloroplast and mitochondrial plastid sequences (alignments to plastids were not used) using NovoAlign (v.3.04.04; novocraft.com) with the options -o SAM -t 40 -r None -a TGGAATTCTCGGGTGCCAAGG -s 30 -l 25 (similar to the -local setting of Bowtie 2, but with a more quality-informed alignment location for short reads). Read data were analysed using the dREG algorithm[53].

Read preparation for genome assembly of wild rice *Oryza australiensis*. *Plant material and methods for genome sequencing and assembly.* A voucher specimen of *O. australiensis* (IRGC number 86534) was obtained from the International Rice Research Institute. Leaf tissue for library construction was collected from 17-d-old, young plants in sterilized, premixed soil (50% perlite:vermiculite). Seeds were incubated for 12 d at 50 °C in the dark and subsequently germinated. Plants were grown in growth chambers (11-h days; 29.6 °C/24.0 °C day/night; 300–500 µmol quanta m^{-2} s^{-1}; relative humidity: 60%).

Illumina fragment library and sequencing. Young leaf tissue was collected, and DNA was extracted using DNeasy Plant Mini kits (Qiagen) following the manufacturer's protocol. About 1 µg of DNA was sheared, and fragmented DNA was used to construct Illumina sequencing libraries. Fragment libraries, with a target insert size of 450 bp, were constructed using an Illumina TruSeq DNA PCR-Free Library Prep Kit and bead purified (Agencourt AMPure XP beads; Beckman Coulter). Fragment libraries were 2 × 250-bp-sequenced using a HiSeq 2500.

Assembly of wild rice genomes. *O. australiensis* reads were 3′ quality ($q < 20$) and adapter trimmed using cutAdapt (v.1.11), and were assembled using DiscovarDeNovo (https://software.broadinstitute.org/software/discovar/blog/) with default options. Assembled scaffolds < 3 kb in size were discarded along with scaffolds with GC% > 0.55 (usually corresponding to microbial contaminants). The remaining scaffolds were trimmed for doubly assembled regions. This generated an assembly of 785 megabases (99.98% called bases; 0.02% gaps) in 39,193 scaffolds, with scaffold and contig N_{50} values (that is, the minimum length needed to cover 50% of the genome) of 37 and 34 kb, respectively. Scaffold sizes ranged from 1.5–948 kb, similar to the size distributions for other plant genome assemblies used for comparative genomics[14].

Oryza officinalis was assembled through a hybrid assembly strategy. Paired-end reads were downloaded from the Sequence Read Archive (SRA) (DRR000711, DRR003647 and DRR003646), and were 3′ trimmed for adapters and low-quality sequences ($q < 20$) using cutAdapt (1.9.1). Reads were initially assembled with Ray[69] (v.2.3.1) to generate relatively short contigs that were fragmented in silico into overlapping reads (2 × 100 bp in a 180-bp insert). Overlapping and original reads were combined with longer-insert 6- and 8-kb mate-pair reads (DRR003207 and DRR003206) that were processed for 3′ adapter trimming. Reads were converted to unaligned BAMs (Picard tools; https://broadinstitute.github.io/picard/) and assembled using AllPaths-LG[70] (v.52488), resulting in an assembly of size 399 Mb (85% called bases; 15% gaps) in 13,189 scaffolds, a scaffold N_{50} value of 64 kb and scaffold lengths ranging from 0.8–450 kb.

greenINSIGHT. A detailed description of the method is given in Gronau et al.[3]. We adapted the pipeline for plant genomes (*A. thaliana* and *O. sativa*) by modifying human-specific elements, including criteria for selection of neutral sites, and introducing several minor adjustments to account for the typically lower depth of sequencing across plant genomes, which causes an increase in noise in the population genetic analysis. The greenINSIGHT pipeline is outlined in Supplementary Fig. 12, its input being a set of genomic alignments used to infer an ancestral probability, a set of base calls across a population of the focal species and a set of sites inferred to be neutral from annotation and functional evidence.

There are three differences in greenINSIGHT relative to the human INSIGHT pipeline. First, the INSIGHT approach benefits in several ways (for example, improved alignments and more powerful inference of recent and ancient selection) from a balance between evidence for constraint derived from recent population divergence and that derived from mutations that have accumulated since the ancestral state Z. Consequently, Z was targeted to be the most recent common ancestor of the rice AA genomes, even though this necessarily introduces a risk for noise due to interspecies introgression. We considered an adjusted INSIGHT model in which the topology of flanking neutral models was dynamically adjusted, but this required concatenation of neutral blocks in order for sufficient evidence for tree building to be generated. Therefore, we decided to interleave flanking neutral regions as closely as possible with their matched test sites, since the most significant problems with INSIGHT in species with a high level of inter-specific genomic introgression arise when flanking sequences have a different ancestry relative to matched test sequences.

Second, due to the low fidelity of alignments to transposable elements, flanking neutral sites had a different distribution in transposable elements relative to target regions. This created an asymmetry in alignment fidelity between neutral and non-neutral sites in which neutral sites probably saw different proportions of read noise. This was evident in the plot of Tajima's *D* (Supplementary Fig. 12), where site classes that also had an enrichment for transposable element content showed significant excess of rare variants. We consequently adjusted θ for neutral sites by blending per-block θ with mean θ across all neutral sites in the analysis using a sigmoid function that had a low impact for θ with a central tendency, but increasingly limited large departures in θ.

Third, for a base to be introduced into the human INSIGHT model, it required a valid ancestral state probability and a distribution of {A,C,T,G} base calls over virtually the entire population. In the greenINSIGHT scenario with relatively lower sequencing depth and potentially poorer alignments in some repetitive regions, we allowed slightly more missing population data (0–10%), and used the call-adjusted proportion of base counts to determine whether polymorphic bases were at high or low population frequency.

A notable difference in the use of this pipeline in rice is the combination of extensive linkage disequilibrium and short introns (mean intron size: ~400 bp in rice; ~15 kb in humans) in rice. This combination may introduce more indirect background selection into ρ estimates in the rice model.

Chromatin state modelling. Chromatin states were inferred based on a set of chromatin marks (ATAC-Seq, DNA methylation, H3K27ac, sRNA associated, H3K4me3, H3K18ac and H3K27me3) that were found through earlier work in *Arabidopsis* (unpublished) to be informative of chromatin states in both coding and noncoding regions of plant genomes. Because the ChromHMM pipeline uses a single sample per covariate, the replicate with the highest signal-to-background ratio was selected. This sample is available for viewing and downloading in the UCSC Browser. The hidden Markov modeller ChromHMM was used essentially as described by Gulko et al.[8] to binarize chromatin signals from 40-nucleotide genomic windows (the mean size of regulatory elements previously suggested in plants[14]) and produce a low-resolution map of chromatin states in the rice genome. We initially sought to use a parameter-count penalized log-likelihood ratio for each model to determine the number of distinct chromatin states in rice, cognizant of earlier principal component analyses that suggested that plant genomes have very few states[47] (a number of states similar to *Drosophila*[48]) and more recent ChromHMM analyses that described a relatively complex patterning of states[32]. While there was no distinct inflection in the penalized log-likelihood ratio curve that could inform a high-confidence cutoff, it appeared that after ~15 states, the addition of more states was less informative. We therefore sought to support selection of the number of states with the ChromHMM tool CompareModels[46], which compares the correlation between emissions of models with different numbers of states. Selecting a 50-state model as an overparameterized reference model, we found that by the time 20 states were incorporated in the model, the mean state correlation with the closest state in the 50-state model exceeded 0.9. We therefore selected a 20-state ChromHMM model, but recognize that the optimum value here may not be independent of the number and type of chromatin marks used as input.

Determination of fitCons classes from ChromHMM states. The 20 chromatin states determined by ChromHMM were intersected with additional annotation and functional genomic datasets by setting bits in a 16-bit-wide bitmap of the genome depending on the class value and the binary combination of additional binary annotation. Functional genomic tracks and annotations were binarized for the bitmap at signal (such as alignment depth) thresholds chosen to differentiate between the presence or absence of signal against a track-specific noise background for alignment classes, and split annotation classes marked as 0/1 (messenger RNA = 40 reads; PRO-Seq = 7 reads; phastCons = 0.7 score; Exon = 0.5; and CDS = 0.5). This had the potential to generate 640 classes (note that class labels do not run continuously due to padding left in the bitmap for additional chromatin states). However, many states were either not found or found only very rarely, such that they could not be informatively used to infer a state-specific value of ρ. Consequently, states with a coverage of <20 kb were excluded and 246 final states were characterized by INSIGHT.

Genomic annotations. *Neutral sites.* The collection of sites predicted to be neutral was obtained by eliminating from all genomic sites those likely to be under direct or indirect selection, including: (1) annotated protein-coding genes and 1,000-bp flanking regions on either side of IRGSP1.0.37 annotated genes; (2) CNSs (see below); (3) locations associated with an open chromatin (ATAC) signal; (4) sites with coding potential predicted by other approaches, including fgenesh, Phytozome XI, RAP-DB and blastp of all rice-lineage proteins; and (5) regions with a total expression depth of >50 reads. Due to limited intergenic annotation available in rice compared with humans, this leaves open the possibility that sites in neutral regions could more often intersect functional domains in rice than in humans.

Genome annotations. IRGSP1.0.37 gene and ncRNA (pre-miRNA, long ncRNA, small nucleolar RNA, ribosomal RNA and transfer RNA) annotations were used (Ensembl; ftp://ftp.ensemblgenomes.org/pub/plants/release-37/gff3/oryza_sativa) for annotation of the MSU7 assembly. Mature miRNA annotations were generated from the annotation in miRBase 22 (http://www.mirbase.org/). Where gene IDs required conversion between MSU7 and RAP-DB naming schemes (for example, in the expression correlation analysis), the 2018 translation table from RAP-DB (https://rapdb.dna.affrc.go.jp/download/irgsp1.html) was used. Consensus annotations for genomic features were obtained from featureBits (http://www.soe.ucsc.edu/~kent/src/unzipped/hg/featureBits) and the intersection of IRGSP, MSU7, RAP-DB and IOMAP annotations.

PhastCons and CNSs. PhastCons[1] (http://compgen.cshl.edu/phast/) scores of tribe-level interspecies conservation were generated as one input into the genome classification process. The process of multiple species alignment was carried out using Kent alignment pipeline with minor adjustments, essentially as described by Haudry et al.[14]. Following the MULTIZ[71] multiple alignment step, a global neutral evolutionary model was generated from neutral sequence locations (as described above) using phyloFit[72] and input into phastCons (-target-coverage 0.11 -expected-length 50) along with the multiple sequence alignment to infer a per-base score for non-neutral evolution across the genome. Because both direct and indirect introgression into the reference genome from the set of non-reference genomes being compared can be readily mistaken as a signal of constraint, we avoided using non-reference rice genomes with evidence in the literature for extensive introgression with *O. sativa*. The inference of conservation at medium resolution (>10 bp) can most clearly be made when neutral divergence has introduced substitutions at a higher rate in some parts of the genome than in others. However, for this inference, these neutrally diverged regions need to be similar enough to remain alignable to the reference genome. This creates an optimal neutral divergence level in the phylogeny where the total neutral branch length between reference and non-reference species is 0.1–0.3. Since diploid, assembled rice genomes were not readily available in this range, we chose to assemble two additional rice genomes (*O. australiensis* and *O. officinalis*) to a locus-level (N_{50} = 10–50 kb). This brought the total number of genomes used for the comparative analysis to eight (see Supplementary Fig. 1). Once phastCons scores had been generated, a set of CNSs were inferred by selecting regions with a phastCons score > 0.82 and a length longer than 11 bp (a combination intended to reduce the chance of coincidental undiverged alignments generating a conserved region) that did not overlap CDSs or CDS boundaries, and that did not overlap regions that had a possibility of having been protein coding in the immediate evolutionary past (inferred through blastp of all rice tribe proteins against the *O. sativa* genome).

Enrichment of classes by annotation features. Genome-wide enrichment between genomic classes and annotated features used featureBits (Kent tools; UCSC) with bed file inputs of class and feature locations, and optional parameter enrichment for estimation of enrichment relative to an independent and identically distributed assumption. Essentially the same process was used to determine enrichment of classes relative to ATAC-open and PRO-Seq-expressed regions. Raw ATAC and PRO-Seq data were first binarized using the ChromHMM binarizeSignal option (with ChromHMM selecting the most informative threshold relative to a fitted Poisson distribution) and the binarized signal was in turn used to create a bed-formatted annotation map for use with featureBits.

ρ density by annotation. A genome-wide bedGraph of ρ values was created from a union of all 246 class-specific bedGraphs of per-class ρ. This was intersected with the locations of seven annotations (neutral regions, distal (≥1 kb) CNSs, 500-bp upstream promoters, 5′ UTRs, introns, CDSs and 3′ UTRs) using bedtools 2 in intersect mode with 'awk' used to generate a single-line ρ score and annotation type for each base of the bedGraph ranges. The seven per-base ρ files were combined and displayed using a violin plot (geom_violin) function of the ggplot2 package, with the bandwidth set to 0.03 and the class median plotted as a single white point for each annotation.

Motif enrichment. Motif enrichment was determined using Homer (v.4.10; http://homer.ucsd.edu/homer/) findMotifs with the options -mset plants -len 6,7,8 enabled, and permuted sets of input sequences were used as controls.

Expression correlation analysis. Correlation between genomic class distributions and the expression of proximate genes used aggregated transcriptome profiles across all leaf tissues (excluding flag leaf) from the NCBI Gene Expression Omnibus (GSE21494). Expression was used as per-array median-normalized values. For immediate upstream, downstream and slightly more distal enhancer locations around each gene, featureBits (UCSC; Kent tools) was used to generate a count of total base coverage for each genomic class that was in turn correlated class wise with expression.

Multiple regressions of gene expression and genomic class. A multiple linear regression model for all 246 classes was generated in SPSS (IBM statistics; v.20) by combining individual models of class density around expressed genes (see the expression correlation analysis above) for rice chromosomes 2–12. To reduce the likelihood of overfitting, the model was tested for correlation between predicted levels of log[gene expression] from class density and actual expression of genes on chromosome 1 only (Fig. 2). To explore the possibility that different class distributions around silent and expressed genes are a consequence of gene expression (arising from the spread of activation epigenomic marks into upstream regions) rather than an indication of regulatory changes causative of expression, for the upstream model we masked the immediate 50-bp upstream promoter region where ATAC- and PRO-Seq signals from highly expressed genes can leak into the promoter region. The contour aerial density plot was generated with RAWGraphs (http://app.rawgraphs.io/).

dREG. The PRO-Seq signal was analysed on the Cornell community dREG GPU server for sites with an enhancer potential (https://dreg.dnasequence.org/). Signal was entered as alignment depth bigWig files (see PRO-Seq section in Methods) separately for nascent RNA with a Crick-and-Watson strand origin, and with other options set to defaults. This generated 58,920 candidate sites with a signal range from 0.32–1.46. Since many sites had only a weak PRO-Seq signal, we selected the top ~5% of sites (3,600 sites) with a signal ≥ 1.0 for further characterization. These were in turn refined for exactly 1,000 locations at least 1 kb distal of genes (featureBits); while enhancers can be located within gene and promoter regions, without high-resolution looping predictions, we were unable to distinguish these sites from regular promoters and splice elements.

Chromatin mark correlations. For each annotation class (transposable element, upstream500, upstream25, CDS and intronic), the chromatin signal for each chromatin mark used in the ChromHMM model was summed for each location and standardized for location length. To avoid spurious correlations in introns related to transposable element content, transposable elements were masked from intronic regions. Similarly, to avoid correlations characteristic of introns in transposable elements, only transposable elements 1 kb distal of genes were considered. Correlations between marks in each annotation class were then visualized using the R corrplot package (https://github.com/taiyun/corrplot).

Profiles of feature density. Profiles of bedGraph signal 1 kb either side of features of interest were generated by dividing the upstream and downstream regions into 50 × 20-bp blocks and the interior of the feature into five equally sized blocks. Signal within and around the features was then totalled from the bedGraph files at each region to create: (1) a composite profile of total signal integrated over all features; and (2) a representation of the distribution of signal at each location as a set of single lines where more intense colour represents more signal and total signal is used to order the lines.

MAF shift. For each of 246 fitCons classes, the frequency distribution of the minor allele was calculated from a summary of alleles derived from the 3K Rice Genome Project[11] variant call formats. To contrast these values with ρ generated for each class by INSIGHT, a metric was generated that contrasted the standardized minor allele distribution of each class relative to the standardized MAF distribution of class 11 (the largest neutral class covering 22% of the rice genome with a nominal ρ close to zero). The metric was generated by adding the excess of rare alleles in the class of interest in frequency classes 1 and 2 (that is, sites where there were only one or two members of the populations with segregating variants) relative to class 11 to the excess of more common alleles in frequency classes 4, 5 and 6 in class 11 relative to the class of interest.

Gene Ontology. ArgriGO v.1.2 (http://bioinfo.cau.edu.cn/agriGO/) was used for the Gene Ontology analysis.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The read data used to generate the ChromHMM model and genomic classes have been deposited at the NCBI SRA (https://www.ncbi.nlm.nih.gov/sra) and can be accessed through BioProject ID PRJNA586887. Genome assemblies of *O. officinalis* and *O. australiensis* are available from the CoGe CyVerse website (https://genomevolution.org/coge/) with genome IDs id56031 and id56030, respectively. Access to genomic class annotation and INSIGHT scoring of the rice genome is available via a genome browser linked from the project's website (http://purugganan-genomebrowser.bio.nyu.edu/insightJuly2018/greenInsight.html). All epigenomic data tracks, genome annotations, multiple alignments, conservation scores, fitCons scores and site classes are available for visualization and download on a local installation on the USCSC Genome Browser at http://purugganan-genomebrowser.bio.nyu.edu/cgi-bin/hgTracks?db=Osaj&position=Osaj.1%3A166356–178595, and are also available for download from the NCBI SRA (PRJNA586887). The greenINSIGHT-specific data used to generate the greenINSIGHT online tool are available in the "Additional information, scripts & data" section at http://purugganan-genomebrowser.bio.nyu.edu/insightJuly2018/greenInsight.html. The greenINSIGHT-specific code used to generate the greenINSIGHT online tool, as well as the code described in the Methods, are available in the "Additional information, scripts & data" section at http://purugganan-genomebrowser.bio.nyu.edu/insightJuly2018/greenInsight.html.

Code availability

The greenINSIGHT-specific data used to generate the greenINSIGHT online tool are available in the "Additional information, scripts & data" section at http://purugganan-genomebrowser.bio.nyu.edu/insightJuly2018/greenInsight.html. The greenINSIGHT-specific code used to generate the greenINSIGHT online tool, as well as the code described in the Methods, are available in the "Additional information, scripts & data" section at http://purugganan-genomebrowser.bio.nyu.edu/insightJuly2018/greenInsight.html.

Received: 18 July 2019; Accepted: 20 December 2019;
Published online: 10 February 2020

References

1. Siepel, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
2. Schrider, D. R. & Kern, A. D. Inferring selective constraint from population genomic data suggests recent regulatory turnover in the human brain. *Genome Biol. Evol.* **7**, 3511–3528 (2015).
3. Gronau, I., Arbiza, L., Mohammed, J. & Siepel, A. Inference of natural selection from interspersed genomic elements based on polymorphism and divergence. *Mol. Biol. Evol.* **30**, 1159–1171 (2013).
4. McDonald, J. H. & Kreitman, M. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351**, 652–654 (1991).
5. Sawyer, S. A. & Hartl, D. L. Population genetics of polymorphism and divergence. *Genetics* **132**, 1161–1176 (1992).
6. Bustamante, C. D. et al. Natural selection on protein-coding genes in the human genome. *Nature* **437**, 1153–1157 (2005).
7. Smith, N. G. C. & Eyre-Walker, A. Adaptive protein evolution in *Drosophila*. *Nature* **415**, 1022–1024 (2002).
8. Gulko, B., Hubisz, M. J., Gronau, I. & Siepel, A. A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nat. Genet.* **47**, 276–283 (2015).
9. Gulko, B. & Siepel, A. An evolutionary framework for measuring epigenomic information and estimating cell-type-specific fitness consequences. *Nat. Genet.* **51**, 335–342 (2019).
10. Wing, R. A., Purugganan, M. D. & Zhang, Q. The rice genome revolution: from an ancient grain to Green Super Rice. *Nat. Rev. Genet.* **19**, 505–517 (2018).
11. Wang, W. et al. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature* **557**, 43–49 (2018).
12. Stein, J. C. et al. Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus *Oryza*. *Nat. Genet.* **50**, 285–296 (2018).
13. Cao, J. et al. Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat. Genet.* **43**, 956–963 (2011).
14. Haudry, A. et al. An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. *Nat. Genet.* **45**, 891–898 (2013).
15. Huang, X. et al. A map of rice genome variation reveals the origin of cultivated rice. *Nature* **490**, 497–501 (2012).
16. Gutaker, R. M. et al. Genomic history and ecology of the geographic spread of rice. Preprint at https://www.biorxiv.org/content/10.1101/748178v1 (2019).
17. Josephs, E. B., Lee, Y. W., Stinchcombe, J. R. & Wright, S. I. Association mapping reveals the role of purifying selection in the maintenance of genomic variation in gene expression. *Proc. Natl Acad. Sci. USA* **112**, 15390–15395 (2015).
18. Flowers, J. M. et al. Natural selection in gene-dense regions shapes the genomic pattern of polymorphism in wild and domesticated rice. *Mol. Biol. Evol.* **29**, 675–687 (2012).
19. Caicedo, A. L. et al. Genome-wide patterns of nucleotide polymorphism in domesticated rice. *PLoS Genet.* **3**, 1745–1756 (2007).
20. Bradnam, K. R. & Korf, I. Longer first introns are a general property of eukaryotic gene structure. *PLoS ONE* **3**, e3093 (2008).
21. Rigau, M., Juan, D., Valencia, A. & Rico, D. Intronic CNVs and gene expression variation in human populations. *PLoS Genet.* **15**, e1007902 (2019).
22. Berendzen, K. W. et al. Bioinformatic *cis*-element analyses performed in *Arabidopsis* and rice disclose bZIP- and MYB-related binding sites as potential AuxRE-coupling elements in auxin-mediated transcription. *BMC Plant Biol.* **12**, 125 (2012).
23. Freeling, M., Rapaka, L., Lyons, E., Pedersen, B. & Thomas, B. C. G-boxes, bigfoot genes, and environmental response: characterization of intragenomic conserved noncoding sequences in *Arabidopsis*. *Plant Cell* **19**, 1441–1457 (2007).
24. Choi, H. I., Hong, J. H., Ha, J. O., Kang, J. Y. & Kim, S. Y. ABFs, a family of ABA-responsive element binding factors. *J. Biol. Chem.* **275**, 1723–1730 (2000).
25. Lu, T. et al. Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-Seq. *Genome Res.* **20**, 1238–1249 (2010).
26. Peng, T. et al. Differentially expressed microRNA cohorts in seed development may contribute to poor grain filling of inferior spikelets in rice. *BMC Plant Biol.* **14**, 196 (2014).
27. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
28. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-Seq: a method for assaying chromatin accessibility genome-wide. *Curr. Protoc. Mol. Biol.* **109**, 21.29.1–21.29.9 (2015).
29. Feng, S. et al. Conservation and divergence of methylation patterning in plants and animals. *Proc. Natl Acad. Sci. USA* **107**, 8689–8694 (2010).

30. Mahat, D. B. et al. Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-Seq). *Nat. Protoc.* **11**, 1455–1476 (2016).
31. Kwak, H., Fuda, N. J., Core, L. J. & Lis, J. T. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science* **339**, 950–953 (2013).
32. Liu, Y. et al. PCSD: a plant chromatin state database. *Nucleic Acids Res.* **46**, D1157–D1167 (2018).
33. Yan, W. et al. Dynamic control of enhancer activity drives stage-specific gene expression during flower morphogenesis. *Nat. Commun.* **10**, 1705 (2019).
34. Wen, M. et al. Expression variations of miRNAs and mRNAs in rice (*Oryza sativa*). *Genome Biol. Evol.* **8**, 3529–3544 (2016).
35. Zong, W., Zhong, X., You, J. & Xiong, L. Genome-wide profiling of histone H3K4-tri-methylation and gene expression in rice under drought stress. *Plant Mol. Biol.* **81**, 175–188 (2013).
36. Lozano, R. et al. RNA polymerase mapping in plants identifies enhancers enriched in causal variants. Preprint at https://www.biorxiv.org/content/10.1101/376640v1 (2018).
37. Xia, J. et al. Detecting and characterizing microRNAs of diverse genomic origins via miRvial. *Nucleic Acids Res.* **45**, e176 (2017).
38. Wilkins, O. et al. EGRINs (environmental gene regulatory influence networks) in rice that function in the response to water deficit, high temperature, and agricultural environments. *Plant Cell* **28**, 2365–2384 (2016).
39. Tan, F. et al. Analysis of chromatin regulators reveals specific features of rice DNA methylation pathways. *Plant Physiol.* **171**, 2041–2054 (2016).
40. Liu, C., Lu, F., Cui, X. & Cao, X. Histone methylation in higher plants. *Annu. Rev. Plant Biol.* **61**, 395–420 (2010).
41. Liu, N., Fromm, M. & Avramova, Z. H3K27me3 and H3K4me3 chromatin environment at super-induced dehydration stress memory genes of *Arabidopsis thaliana*. *Mol. Plant* **7**, 502–513 (2014).
42. Fang, H., Liu, X., Thorn, G., Duan, J. & Tian, L. Expression analysis of histone acetyltransferases in rice under drought stress. *Biochem. Biophys. Res. Commun.* **443**, 400–405 (2014).
43. Du, Z. et al. Genome-wide analysis of histone modifications: H3K4me2, H3K4me3, H3K9ac, and H3K27ac in *Oryza sativa* L. Japonica. *Mol. Plant* **6**, 1463–1472 (2013).
44. Lee, T., Zhai, J. & Meyers, B. C. Conservation and divergence in eukaryotic DNA methylation. *Proc. Natl Acad. Sci. USA* **107**, 9027–9028 (2010).
45. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).
46. Ernst, J. & Kellis, M. Chromatin-state discovery and genome annotation with ChromHMM. *Nat. Protoc.* **12**, 2478–1492 (2017).
47. Roudier, F. et al. Integrative epigenomic mapping defines four main chromatin states in *Arabidopsis*. *EMBO J.* **30**, 1928–1938 (2011).
48. Sequeira-Mendes, J. et al. The functional topography of the *Arabidopsis* genome is organized in a reduced number of linear motifs of chromatin states. *Plant Cell* **26**, 2351–2366 (2014).
49. Liu, C. et al. Genome-wide analysis of chromatin packing in *Arabidopsis thaliana* at single-gene resolution. *Genome Res.* **26**, 1057–1068 (2016).
50. Guo, H. & Moose, S. P. Conserved noncoding sequences among cultivated cereal genomes identify candidate regulatory sequence elements and patterns of promoter evolution. *Plant Cell* **15**, 1143–1158 (2003).
51. Liu, L., Xu, W., Hu, X., Liu, H. & Lin, Y. W-box and G-box elements play important roles in early senescence of rice flag leaf. *Sci. Rep.* **6**, 20881 (2016).
52. Ding, M. et al. Enhancer RNAs (eRNAs): new insights into gene transcription and disease treatment. *J. Cancer* **9**, 2334–2340 (2018).
53. Wang, Z., Chu, T., Choate, L. A. & Danko, C. G. Identification of regulatory elements from nascent transcription using dREG. *Genome Res.* **29**, 293–303 (2019).
54. Danko, C. G. et al. Dynamic evolution of regulatory element ensembles in primate CD4$^+$ T cells. *Nat. Ecol. Evol.* **2**, 537–548 (2018).
55. Savisaar, R. & Hurst, L. D. Exonic splice regulation imposes strong selection at synonymous sites. *Genome Res.* **28**, 1442–1454 (2018).
56. Cannavò, E. et al. Shadow enhancers are pervasive features of developmental regulatory networks. *Curr. Biol.* **26**, 38–51 (2016).
57. Prescott, S. L. et al. Enhancer divergence and *cis*-regulatory evolution in the human and chimp neural crest. *Cell* **163**, 68–83 (2015).
58. Mezmouk, S. & Ross-Ibarra, J. The pattern and distribution of deleterious mutations in maize. *G3 (Bethesda)* **4**, 163–171 (2014).
59. Wallace, J. G., Rodgers-Melnick, E. & Buckler, E. S. On the road to breeding 4.0: unraveling the good, the bad, and the boring of crop quantitative genomics. *Annu. Rev. Genet.* **52**, 421–444 (2018).
60. Moyers, B. T., Morrell, P. L. & McKay, J. K. Genetic costs of domestication and improvement. *J. Hered.* **109**, 103–116 (2018).
61. Morrell, P. L., Buckler, E. S. & Ross-Ibarra, J. Crop genomics: advances and applications. *Nat. Rev. Genet.* **13**, 85–96 (2012).
62. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10–12 (2011).
63. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
64. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
65. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
66. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
67. Raurell-Vila, H., Ramos-Rodríguez, M. & Pasquali, L. in *CpG Islands. Methods in Molecular Biology* Vol. 1766 (eds Vavouri, T. & Peinado, M. A.) 197–208 (Humana Press, 2018).
68. Hetzel, J., Duttke, S. H., Benner, C. & Chory, J. Nascent RNA sequencing reveals distinct features in plant transcription., *Proc. Natl Acad. Sci. USA* **113**, 12316–12321 (2016).
69. Boisvert, S., Raymond, F., Godzaridis, É., Laviolette, F. & Corbeil, J. Ray Meta: scalable de novo metagenome assembly and profiling. *Genome Biol.* **13**, R122 (2012).
70. Butler, J. et al. ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome Res.* **18**, 810–820 (2008).
71. Green, E. D. et al. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14**, 708–715 (2004).
72. Siepel, A. & Haussler, D. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol. Biol. Evol.* **21**, 468–488 (2004).

Acknowledgements

We thank the New York University Center for Genomics and Systems Biology GenCore Facility and the Next Generation Sequencing core at Cold Spring Harbor Laboratory for sequencing support. We thank O. Wilkins and C. Danko for valuable suggestions relating to the ATAC and PRO-Seq protocols, respectively. This work was supported primarily by a grant from the Zegar Family Foundation (no. A16-0051-004), as well as some support from the National Science Foundation Plant Genome Research Program (no. IOS-1546218) and NYU Abu Dhabi Research Institute to M.D.P., the National Science Foundation CAREER award (no. MCB-1552455), the US National Institutes of Health (no. R35GM124806) and US Department of Agriculture Hatch Program (no. 1012915) to X.Z., the US National Institutes of Health (no. R35GM127070) to A.S., and fellowships from the Gordon and Betty Moore Foundation and Life Sciences Research Foundation (no. GBMF2550.06) to S.C.G. and from the Natural Sciences and Engineering Research Council of Canada (no. PDF-502464-2017) to Z.J.-L.

Author contributions

M.D.P. conceived of the study idea. M.D.P., Z.J.-L., A.E.P. and A.S. designed the study. M.D.P. directed the study. Z.J.-L. and X.Z. collected the data, A.E.P., Z.J.-L., J.Y.C., B.G., S.C.G. and M.D.P. analysed the data. Z.J.-L., A.E.P., A.S. and M.D.P. wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41477-019-0589-3.

Correspondence and requests for materials should be addressed to M.D.P.

Peer review information *Nature Plants* thanks Robin Allaby, Peter Civan and Peter Morrell for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

nature research

Corresponding author(s): Michael D. Purugganan NPLANTS-19077157A

Last updated by author(s): Dec 7, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
□	☒	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
□	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

Illumina sequencing data was collected and demultiplexed using the standard illumina pipeline for the HiSeq2000.

Data analysis

The core INSIGHT code used in this study has previously been described in Human, we include links to this source code, source code for additional analysis and datasets used for the online greenInsight tool in the additional materials section at http://purugganan-genomebrowser.bio.nyu.edu/insightJuly2018/greenInsight.html

Other applications used in data analyses (further command line details are in methods):
- Picard tools, version 1.138, https://broadinstitute.github.io/picard/
- Cutadapt, version 1.11; DOI: https://doi.org/10.14806/ej.17.1.200
- Bowtie2 (version2.2.9); DOI: https://doi.org/10.1038/nmeth.1923
- bedtools (version 2.25) ; DOI: 10.1093/bioinformatics/btq033
- samtools (version 1.3) ; https://sourceforge.net/projects/samtools/files/samtools/1.3/
- UCSC Kent tools; DOI: 10.1101/gr.229202
- ChromHMM; DOI: 10.1038/nmeth.1906
- CompareModels; DOI: 10.1038/nprot.2017.124
- Novoalign (version 3.04.04); URL: http://www.novocraft.com/
- MACS2 (version 2.1.1); DOI: https://doi.org/10.1186/gb-2008-9-9-r137
- dREG algorithm; DOI: 10.1101/gr.238279.118; URL: https://dreg.dnasequence.org/
- DiscovarDeNovo; URL: https://software.broadinstitute.org/software/discovar/blog/
- Ray (version 2.3.1); DOI: 10.1186/gb-2012-13-12-r122
- AllPathsLG (version 52488); DOI: 10.1101/gr.7337908
- featureBits program; URL: http://www.soe.ucsc.edu/~kent/src/unzipped/hg/featureBits
- PhastCons; DOI: 10.1101/gr.3715005 ; URL: http://compgen.cshl.edu/phast/
- Multiz: DOI: 10.1101/gr.1933104
- phyloFit; DOI: 10.1093/molbev/msh039

- R corrplot package; URL: https://github.com/taiyun/corrplot
- ArgriGO (version 1.2); URL: http://bioinfo.cau.edu.cn/agriGO/
- Homer (version 4.10); URL: http://homer.ucsd.edu/homer/

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Multiple alignments, conservation scores, fitCons scores and site classes are available for visualization and download on a local installation on the USCSC Genome Browser at http://purugganan-genomebrowser.bio.nyu.edu/cgi-bin/hgTracks?db=Osaj . Assembled genomes of O. australiensis and O. officinalis are available from the COGE Cyverse website (https://genomevolution.org/coge/GenomeInfo.pl?gid=56030 and https://genomevolution.org/coge/GenomeInfo.pl?gid=56031). Sequencing data is available from the SRA through BioProject PRJNA586887 .

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Three biological replicates were generated to ensure robustness to sample preparation failure for total RNA-seq, small RNA-seq, and ChIP-seq assays. Two biological replicates were generated for the methylation and PRO-seq assays. For each of the functional datasets a single sample was selected as an input in the FitCons approach. The choice of replicate was based on signal strength and sequencing coverage. For ATAC-seq, we used previously published ATAC-seq data (SRR2981235, SRR2981233, SRR2981221, SRR2981227, SRR2981231, SRR2981234). The approach discussed (fitCons) utilizes (as in the human studies eg https://www.nature.com/articles/ng.3196), replication of equivalent sites across the genome rather than replication across samples. We undertook sample replication for the purpose of selecting samples in which the ratio of specific to non-specific signal was highest.
Data exclusions	Because the chromHMM/INSIGHT pipeline uses a single sample per covariate, the replicate with the highest signal-to-background was selected as described in the above section. This sample is available for viewing and downloading in the UCSC Browser. All raw sequencing data will be available at the SRA.
Replication	Samples were generated in triplicate for the purpose of selecting a single best-performing sample for introduction into the downstream pipeline. The INSIGHT/fitCons approach uses coherence across very many genomic sites to estimate error terms on the output metrics (eg rho).
Randomization	Seeds of the cultivated rice Oryza sativa landrace Azucena (IRGC#328; tropical japonica) were germinated and grown on hydroponic pots. Every two days, pots were randomly moved in the trays and between shelves to avoid edge bias. Leaves from multiple (N>=3) pots were selected at random for sequencing.
Blinding	No blinding was used.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
□	☒ Antibodies
☒	□ Eukaryotic cell lines
☒	□ Palaeontology
☒	□ Animals and other organisms
☒	□ Human research participants
☒	□ Clinical data

Methods

n/a	Involved in the study
□	☒ ChIP-seq
☒	□ Flow cytometry
☒	□ MRI-based neuroimaging

Antibodies

Antibodies used

- H3K27ac: Cell Signaling Technology, cat. #4353S, Lot # 1
- H3K4me3: Millipore, cat. #07-473, Lot # 2746331
- H3K27me3: Millipore Sigma, cat. #07-449, Lot # 2919706
- H3K18ac: Cell Signaling Technology, cat. #9675S, Lot # 1

Dilutions:
H3K4me3: 2ug of antibody per chip
H3K27ac: 4ug of antibody per chip
H3K27me3: 4ug of antibody per chip
H3K18ac: 1:25 dilutions as per manufacturer instructions

RRIDs:
H3k4me3:AB_1977252
H3k27me3: AB_310624
H3k27ac:AB_10545273
H3k18ac: AB_331550

Validation

- H3K4me3 antibody has been previously used in ChIP-seq studies in Oryza sativa using this antibody: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1010149, DOI: https://doi.org/10.1105/tpc.112.102269
- Antibodies for H3K27acc have been used in studies in Oryza sativa: DOI: 10.1016/j.bbrc.2013.11.102
- H3K27me3 antibody has been previously used in ChIP-seq studies in Oryza sativa using this antibody: DOI: https://doi.org/10.1105/tpc.112.102269
- Antibodies for H3K18ac have been used in studies in Oryza sativa: DOI: 10.1016/j.bbrc.2013.11.102

ChIP-seq

Data deposition

☒ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☒ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links
May remain private before publication.

Prior to publication all processed epigenomic data tracks, genome annotations, multiple alignments, conservation scores, fitCons scores and site classes will be available for download from the project's genome browser (http://purugganan-genomebrowser.bio.nyu.edu/cgi-bin/hgTracks?db=Osaj&lastVirtModeType=default&lastVirtModeExtraState=&virtModeType=default&virtMode=0&nonVirtPosition=&position=Osaj.1%3A166356-178595). All raw data will be available via the NCBI's SRA

Files in database submission

Provide a list of all files available in the database submission.

Genome browser session
(e.g. UCSC)

Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.

Methodology

Replicates

Three biological replicates were generated to ensure robustness to sample preparation failure for total RNA-seq, small RNA-seq, ATAC-seq, and ChIP-seq assays. Two biological replicates were generated for the methylation and PRO-seq assays. For each of the functional datasets a single sample was selected as an input in the FitCons approach. The choice of replicate was based on signal strength and sequencing coverage. Three biological replicates were generated to ensure robustness to sample preparation failure for total RNA-seq, small RNA-seq, and ChIP-seq assays. Two biological replicates were generated for the methylation and PRO-seq assays. For each of the functional datasets a single sample was selected as an input in the FitCons approach. The choice of replicate was based on signal strength and sequencing coverage. For ATAC-seq, we used previously published ATAC-seq data (SRR2981235, SRR2981233, SRR2981221, SRR2981227, SRR2981231, SRR2981234).

Sequencing depth

(see also methods)
H3K27ac - 16,44,0676 reads, 14,897,894 aligned

H3K4me3 - 17,866,235 reads, 14,947,598 aligned
H3K27me3 - 15,595,820 reads, 8,681,535 aligned
H3K18ac - 9,553,296 reads, 8,766,304 aligned

Antibodies

- H3K27ac: Cell Signaling Technology, cat. #4353S, Lot # 1
- H3K4me3: Millipore, cat. #07-473, Lot # 2746331
- H3K27me3: Millipore Sigma, cat. #07-449, Lot # 2919706
- H3K18ac: Cell Signaling Technology, cat. #9675S, Lot # 1

Peak calling parameters

ChiP seq data was used per the human and drosophila INSIGHT protocols (eg https://www.nature.com/articles/ng.2658).

Data quality

Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.

Software

While peaks were assessed with MACS2, they are not reported. ChIP-seq data was used in a Chromatin state pipeline that was not peak focused . See methods (ChromHMM)