Nucleotide diversity analysis highlights functionally important genomic regions

Tatiana V. Tatarinova1,2,* Evgeny Chekalin3,*, Yuri Nikolsky3,4,5, Sergey Bruskin1, Dmitry Chebotarov6, Kenneth L. McNally6 & Nickolai Alexandrov6

We analyzed functionality and relative distribution of genetic variants across the complete Oryza sativa genome, using the 40 million single nucleotide polymorphisms (SNPs) dataset from the 3,000 Rice Genomes Project (http://snp-seek.irri.org), the largest and highest density SNP collection for any higher plant. We have shown that the DNA-binding transcription factors (TFs) are the most conserved group of genes, whereas kinases and membrane-localized transporters are the most variable ones. TFs may be conserved because they belong to some of the most connected regulatory hubs that modulate transcription of vast downstream gene networks, whereas signaling kinases and transporters need to adapt rapidly to changing environmental conditions. In general, the observed profound patterns of nucleotide variability reveal functionally important genomic regions. As expected, nucleotide diversity is much higher in intergenic regions than within gene bodies (regions spanning gene models), and protein-coding sequences are more conserved than untranslated gene regions. We have observed a sharp decline in nucleotide diversity that begins at about 250 nucleotides upstream of the transcription start and reaches minimal diversity exactly at the transcription start. We found the transcription termination sites to have remarkably symmetrical patterns of SNP density, implying presence of functional sites near transcription termination. Also, nucleotide diversity was significantly lower near 3′ UTRs, the area rich with regulatory regions.

Understanding the relationship between genotype and phenotype is a key issue in life sciences with hugely important implications in biomedical R&D, healthcare and agriculture. Innate genetic variability is both the source and consequence of selection in populations of humans, crops and animals. There is a fine balance between the variability in DNA sequences and the evolutionary constraints for conservation of the original state or fixation of a new variant with a selective advantage. Most genetic variants occur as single nucleotide polymorphisms (SNPs) and small insertions and deletions (indels). In eukaryotic genomes, distribution of variants and the rate of spontaneous mutations are not uniform across individual sites1,2. Instead, they depend on functional impact of variants on either the protein structure or the RNA structure in the regulatory regions affecting transcription. Thus, a SNP causing a premature stop codon will truncate a protein sequence, which may be phenotypically significant. Not surprisingly, genetic variability in protein coding regions is two to three times lower than in intergenic regions2–5. Similarly, promoter regions (and especially the sequences within transcription factor binding sites, TFBSs) are less prone to SNPs and indels than intergenic regions, with the levels of sequence variation within and around TFBSs inferred from their position weight matrix6. Genetic variability can be also assessed from an evolutionary point of view, using a combination of phylogenetic and population genetic techniques7. Thus, mutational “hot” and “cold” spots were discovered by analysis of a population of 34 E. coli strains7. A lower mutation rate was observed in highly expressed genes and in those undergoing stronger purifying selection that reduces retention of deleterious mutations7. In another publication, mutational hotspots were linked to the regions of open chromatin in mammalian genomes8. Both in normal and cancerous tissues, mutation rates were found to be related to epigenetic features, such as DNA methylation and nucleosome occupancy9,10, with disproportionately higher numbers of...

1Center for Personalized Medicine and Spatial Sciences Institute, University of Southern California, Los Angeles, CA, USA. 2Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russian Federation. 3Vavilov Institute of General Genetics, Moscow, Russia. 4F1 Genomics, San Diego, CA, USA. 5School of Systems Biology, George Mason University, VA, USA. 6International Rice Research Institute, Los Baños, Laguna 4031, Philippines. *These authors contributed equally to this work. Correspondence and requests for materials should be addressed to T.V.T. (email: tatiana.tatarinova@usc.edu) or N.A. (email: n.alexandrov@irri.org)
transcription factor-encoding genes are highly conserved, whereas kinases and membrane-bound molecules. Arai-Kichise (sequencing and bioinformatics analyses of 104 rice varieties belonging to the major subspecies of Oryza sativa), McCouch et al. (1,568 diverse inbred rice varieties analyzed at 700,000 SNPs), Huang et al. (rice domestication analysis based on 1,083 cultivated indica and japonica varieties), Xu et al. (resequencing of 50 accessions at >15 × raw coverage), Duitama et al. (sequencing and bioinformatics analyses of 104 rice varieties belonging to the major subspecies of Oryza sativa), Arai-Kichise et al. (analysis of SNP in seven rice cultivars of temperate and tropical japonica), Jain et al. (whole-genome resequencing of three rice cultivars with contrasting responses to drought and salinity stress).

Most studies on genomic variability have been conducted on mammals and yeast, with much lesser attention to plants. Analysis of genomic variation in rice was recently presented by McCouch et al. (1,568 diverse inbred rice varieties analyzed at 700,000 SNPs), Huang et al. (rice domestication analysis based on 1,083 cultivated indica and japonica varieties), Xu et al. (resequencing of 50 accessions at >15 × raw coverage), Duitama et al. (sequencing and bioinformatics analyses of 104 rice varieties belonging to the major subspecies of Oryza sativa), Arai-Kichise et al. (analysis of SNP in seven rice cultivars of temperate and tropical japonica), Jain et al. (whole-genome resequencing of three rice cultivars with contrasting responses to drought and salinity stress).

Understanding DNA variability in plants is hugely important, as crops are the core of world's agriculture. Food consumption is predicted to double within the next 35 years, which translates in the demand for 2% annual yield growth rates for major crops. Yet, current breeding and selection methods, including transgenic technology, can only support on average a 1.1% annual growth rate. Rice is the key cereal for the majority of the global population, especially in Asia and Africa. Climate change and increasing lack of agricultural land pose further challenges to rice production. Achieving the required increase of global production by up to 50% by 2050 will require new technologies, such as genomic selection and genome editing. Genomic selection depends on selection of those SNPs that favorably contribute to phenotype. Generally, success of new computational methods relies on understanding the specifics of plant biology, not shared with animals, such as self-fertilization, polyploidy and higher genetic variability. Also, dicot and monocot plants have their own unique genetic features, and, therefore, one cannot liberally extrapolate knowledge from one plant to another, for example from the model dicot Arabidopsis thaliana to grasses.

Here, we provide a large-scale study on genetic variability in the rice genome, based on high quality data from the SNP-seek database, with over 40 million SNPs from 3,000 rice genomes, the largest source of plant SNPs to date. Accessing this unique database allowed us to precisely describe genome-wide patterns of variability and to detect pronounced patterns not seen before. We provide new evidence on close association between local genetic variability and functionality, and suggest that the patterns of SNP density can help to detect functionally important genomic regions that are essential for crop improvement. Thus, we discovered specific characteristic patterns near sites with well-defined biological functions, such as transcription and translation start and stop sites, promoters, intron/exon junctions and 5′ and 3′ UTRs. Some features are similar to those in animals, while the others are novel and unique. In addition, we calculated relative SNP density on different groups of genes and demonstrated that transcription factor-encoding genes are highly conserved, whereas kinases and membrane-bound transporters are the most variable gene families.

**Results**

**Genome-wide mutation rates.** We analyzed the distribution of SNPs in the genome of Oryza sativa. We found that the minor allele frequency (MAF) distribution varies between parts of genome. The original 29M dataset of bi-allelic SNPs mapped to the Nipponbare genome contains many rare SNPs: 22% of them have MAF < 0.00035 (occurring in one genome in 3,000 in a homozygous state), and 67% of them have MAF < 0.01 (present in less than 30 genomes). The genome-wide average MAF is 0.052 while the mean MAF for CDS is 0.028 and, contrastingly, 0.061 for introns. Exons contain more SNPs with lower MAF scores as compared to introns and UTRs. With an increase in the MAF cut-off (to control sequencing errors), a larger portion of SNPs in coding regions are excluded from the dataset. Rare exonic SNPs originated from five genomes of Oryza glaberrima, showing massive levels of genomic structural variation, with particularly high instability in defense-related genes. These accessions were excluded from our analysis since they are too evolutionary distant from O. sativa. Twelve other genomes were excluded since one had a very low coverage and others had excessive numbers of heterozygous singleton SNPs, which is highly unlikely for an inbred species. To reduce the number of possible false positive polymorphisms, we imposed two levels of constraints, as described in Materials and Methods, arriving at two datasets (“Base” 16M, and “Filtered” 5M) for the analysis (Table 1).
Transcription start and termination sites. Genomic regions in the vicinity of transcription start (TSS) and termination (TTS) sites are enriched in important regulatory elements. The complex formed by transcription factors and RNA polymerase binds to specific conserved sequences within the promoter (Fig. 1). Between the translation and transcription start and, subsequently, termination sites, there are untranslated regions (UTRs) containing conserved elements that regulate translation. Figure 2 shows the distribution of SNPs near the transcription start and end sites. In order to confirm the expected lower sequence variability in the vicinity of TSS

Figure 1. Regions near translation and transcription start and termination sites are more conserved than the surrounding genic and intergenic regions.

Figure 2. Density of SNP near the transcription start site (TSS), panels (A,C); and termination site (TTS), panels (B,D). Panels (A,B) correspond to the “Filtered” dataset, panels (C,D) to the “Base” dataset. TSS regions contain the following amount of SNPs: 553,800 “Filtered” and 1,443,188 “Base” datasets. TTS regions contain the following amount of SNPs: 482,685 “Filtered” and 1,231,570 “Base” datasets. Blue curve shows the distribution of 5′ (A,C) and 3′ (B,D) UTR lengths. Whole-genome fraction of variable positions is 1.3% for the “Filtered” and 4.4% for the “Base” datasets, respectively.
and TTS, we extracted 2 kb long fragments centered at TSS and TTS for 20 K high-confidence rice genes (see Methods) and computed the position-specific density of SNPs. At approximately 250 nucleotides upstream of the TSS (see Fig. 2A,C), the SNP density starts to decline, reaching the minimum at the TSS at approximately half of the intergenic value. The steep decline in the SNP rate in the promoter region is in line with the restrictions imposed by the presence of the conserved TFBS in the core promoter [−250, 0] upstream of the TSS (Fig. 3). Within the transcribed region, the density steadily increases from 5′ to 3′ end of the gene and plateaus approximately 300 nucleotides downstream from the TSS as shown in the Fig. 2(A,C). This region of linear increase roughly corresponds to the 5′ UTR. Lower sequence variability in the 5′ UTR regions may be explained by the observation that mutations in this region can change local mRNA structure near the 5′ cap and, therefore, affect translation process. The protein-coding region features fewer variants because of the demand to preserve the amino acid sequence of the product, which is supported by changes mostly in the third codon positions (Fig. 4).

The final stage of transcription is its termination, when the complete transcript dissociates and the RNA polymerase is released from the DNA template. The mechanism of termination is the least understood of the three transcription stages; two competing, yet not fully satisfactory43 models known as “allosteric” and “torpedo”44 are proposed as possible mechanisms. The distribution of SNPs around the transcription termination site almost mirrors the trend at the TSS, Fig. 2(B,D). SNP density begins to show a decline at about 500 bp upstream from the TTS, reaching the minimum just before the TTS. It then steadily increases for 300 nucleotides downstream from the TTS and then levels off reaching the intergenic level of SNP density. However, the plateau is achieved at over 1000 nucleotides downstream. We suggest that this difference is due to the intrinsic variability of the termination process, with the position of transcription termination being less precisely defined than the transcription start site. This profile of SNP density variation suggests the existence of evolutionary constraints protecting the TTS area, such as requirements to terminate transcription at the appropriate positions45, to interact with RNA-binding proteins to regulate mRNA translation46, and to accommodate miRNA target sites47.

SNPs in the protein coding regions. Overall, coding regions are more conserved than intergenic regions, promoters and UTRs (Table 1). The frequency of SNPs in introns is 30% higher than in exons, as calculated for the “Filtered” dataset. These results agree with previous observations2. There are several interesting trends in SNP distributions near the translation starts and terminations (Fig. 4). First, there is a conserved region immediately upstream of the first codon. The context around the start codon is critical for the translation initiation, as the ribosome may bind at the 5′-most ATG of the mRNA or proceed to the next start codon48. Across eukaryotes, the consensus around the start codon is (gcc)gccRccAUGG49 (Kozak consensus sequence). In our analysis of 3,000 rice genomes, the consensus is C(G/C)GC(G/C)AUGGC, adjusting (but not contradicting) the previously reported consensus for rice (g/c)(A/G)(A/C)(G/C)AUGGC50. The initiation of translation (and, therefore, conservation requirements for 5′ UTR) is affected by multiple regulatory mechanisms involving binding of regulatory proteins to specific regions within 5′ UTR, open reading frames and ribosome entry sites57.

Translation termination depends on codon-specific release factors that recognize the stop codon sequence in mRNA, as well as on GTP-binding release factors51. Efficiency of translation termination depends on the consensus sequence immediately downstream from the stop codon52,53. Kochetov et al.54 suggested that nonsense mutations occur significantly more often at the very beginning of 3′ UTR and in the same reading frame as the CDS. They found that A. thaliana and O. sativa genes with UGA stop codons had more nonsense codons in the first triplet position of 3′ UTR (that may result from weak natural selection for a “reserve” stop signal). Approximately 20–30 nucleotides (nt) upstream of the cleavage-polyadenylation site is the AAUAAA hexamer, which is the consensus signal for the transcript cleavage57. Therefore, there are constraints affecting the 3′ UTR sequences manifested as reduced SNP density55 (Fig. 4(B,D)). SNPs density gradually increases downstream from the TTS.
although the incline of the growth function is smoother than the decline upstream of the first codon. This may be explained by differences in distributions of the 5′ and 3′ UTR lengths (Fig. 2) and weaker constraints on the positional precision of the termination process as compared to transcription initiation.

The third position of the codon, as expected, is more variable than the first two, and that the first one is slightly more variable than the second, which follows from the redundancy of the genetic code and agrees with prior observations in human and mouse19.

We also analyzed sequence variability in introns (Fig. 5). In both “Filtered” and “Base” datasets, there are more SNPs in introns as compared to the adjacent exons. The exon/intron boundaries are particularly conserved, with the shape of SNP density function around the splicing sites similar to that around the translation and transcription initiation sites and the stop codon.

**Nucleotide composition and SNPs.** There are two classes of rice genes that differ in function, nucleotide composition, promoter organization, gene expression, and other features36,56–58. Here, we investigated distribution of SNPs as a function of the nucleotide composition asymmetry expressed as CG-skew and AT-skew59–62; both dependencies have remarkable shapes (Fig. 6). CG-skew was computed for each 500 nt long genomic region as a difference between the numbers of cytosines and guanines in this region divided by the sum of cytosines and guanines

\[
CG_{skew} = \frac{C - G}{C + G}
\]

AT-skew was similarly calculated as

\[
AT_{skew} = \frac{A - T}{A + T}
\]

Nucleotide asymmetry is an important measure, since CG-skew is correlated with the duration a particular DNA region stays in an unpaired, unprotected state during transcription and replication where the longer the duration, the higher the probability of that cytosine transitions to thymine57,59. Position-specific imbalance between cytosines and guanines was speculated to be related to translational efficiency63, avoidance of “kissing interactions,”64 DNA methylation67, distinct RNA polymerase pause sites in CpG island promoters65, and adaptation to extreme environments37,64,66,67. In addition, mRNAs tend to be purine-rich (where there is an excess of Gs over Cs)64,66,69. We evaluated the CG skew distribution in rice by dividing the genome into 500 nt long fragments and calculating the nucleotide composition and the number of SNPs in each window (Fig. 6B,D). Since the number of genes in both DNA strands is approximately equal, the plot is symmetric around zero CG-skew. Regions with high absolute values of the CG-skew
(0.3) tend to have high SNP density (Fig. 6B). Genes with high values of the CG-skew are less likely than the genes with low CG-skew to correspond to proteins of unknown function (Cramer’s V = 0.34, see Supplement for details). This observation can be explained by an earlier findings that the peak of CG-skew near the TSS is associated with high transcriptional efficiency36,38,59 and that highly expressed genes have a higher chance to be studied as compared to the genes expressed at a low level. Interestingly, twelve out of top twenty genes by SNP density with |CG-skew| > 0.3 are located on chromosomes 11 and 12, which are enriched in disease resistance genes and recent gene duplications 70. Chromosomes 11 and 12 constitute the most recently evolved part of the rice genome70, containing genes that may work differently in different ecotypes of rice subjected to distinct stress conditions present in their respective sources of origin.

Distribution of the AT-skew is shown in Fig. 6(A,C). Variability of the AT-skew was previously linked to tDNA insertion sites71, purine loading of mRNA64,68,69, and promoter activity72. We found that approximately two thirds of the genes occurring in regions with an absolute value of AT skew above 0.3 correspond to transposable elements (TE), in contrast to genes in genomic regions with AT skew between [−0.3, 0.3], where only 30% of genes correspond to TEs. This effect may also be explained by the observation that protein-encoding regions have distinct patterns of nucleotide composition (manifested as GC content, AT and CG skew patterns)37,59,73 and also have fewer SNPs than the intergenic regions.

Motifs in promoters associated with SNP density. In general, DNA sequences in promoters, enhancers and other regulatory elements in eukaryotes are less variable than the rest the genome74–77. However, the patterns of variability may be organism specific and are not well studied in plants. We applied the cisExpress algorithm78,79 to find regulatory motifs associated with altered SNP density in the region [−250, 50] around the TSS, corresponding to the core promoter and 5’ UTR. Presence of a TATA-box at −30 nt from the TSS and motif GCCC at [−250, −50] are negatively correlated with promoter sequence variability. On the other hand, an (AT)n repeat at 100 or more nucleotides away from the TSS is positively associated with promoter sequence variability (Fig. 7). To estimate statistical significance, we compared the mean number of SNPs in promoters with and without these motifs. The Z-score for the GCCC[ca] motif, which is possibly related to the GCC-box80–82 and involved in defense mechanisms, was −16.95, Z-score for the TATA box at −30 was −9.95, and the Z-score for TTAT[ca]...
was 20.6, suggesting that the GCCC[ca] and TATA-box containing promoters do not tolerate sequence variability, while TTAT[ca]-containing promoters are enriched in SNPs. We suggest that this effect is due to the possible alternative starts of transcription in various ecotypes of *O. sativa* that require a modified promoter to adapt gene expression level to changing environment. To test this hypothesis, one needs to analyze a large collection of 5′-full transcripts from various ecotypes.

**Gene sequence conservation is associated with functionality.** Sequence variability in protein-encoding genes may be related to their group-wise functionality, as defined by type-specific biological pathways and processes\(^{83,84}\). Using the complete 29M set of SNPs, we performed the one sample t-test to find possible associations between DNA conservation in rice genes with GO categories (Table 2). Functional bias was, indeed, pronounced, with the most conserved genes enriched for transcription regulation processes, whereas the least conserved genes encoded membrane located proteins that are involved in stress response and other processes associated with adaptation to the changing environment. GO analysis of the filtered set of SNPs (“Base”
dataset) shows similar trends and also indicates that the most conserved genes belonged to the functional category “sequence-specific DNA binding transcription factor” (Supp. Table 5).

Protein family analysis (gene enrichment with PFAM categories) showed that transcription factors feature fewer SNPs than other families. Genes with DUF1618, Myb DNA-binding, zf-C3HC4, and AP2 domains were among the least variable genes. Importantly, we saw the same variability patterns at transcription regulation level. SNP density was the lowest in the promoter regions of the genes belonging to “sequence-specific DNA binding transcription factors”.

Discussion

Variability of DNA and protein sequences reflects the balance between adaptation and conservation of essential functionality. Analysis of conserved patterns enables “reverse engineering” of function from the sequence data. Active sites of enzymes can be predicted using conserved motifs in protein families, transcription factor binding sites may be identified using conserved elements in homologous promoter sequences, and gene structures can be predicted from conserved fragments of orthologous genes. In general, an unusually high conservation in a DNA or protein sequence implies existence of a biologically important function.

Here, we analyzed the genetic variability within a unique collection of nearly 30M SNPs from the 3,000 Rice Genomes Project. Although the Human 1000 genome project discovered a larger number of SNPs (80M) from sequencing 2,504 persons, the rice SNP set has a much higher SNP density since the rice genome is almost 10 times smaller than the human genome. To the best of our knowledge, this study has the most precise resolution (one SNP per 20 nucleotides, on average using the “Base” set) of any such study of genome-wide patterns of genic SNP diversity to date.

| SNP density | Change from the baseline | log(P-value) | GO ID | Namespace | Name |
|-------------|--------------------------|-------------|-------|-----------|------|
| 5.40%       | −19.50%                  | −50         | GO:0006355 | Biological process | regulation of transcription, DNA-templated |
| 5.40%       | −19.50%                  | −49         | GO:0003677 | Molecular function | DNA binding |
| 5.10%       | −23.97%                  | −42         | GO:0003700 | Molecular function | transcription factor activity, sequence-specific DNA binding |
| 5.10%       | −23.97%                  | −27         | GO:0043565 | Molecular function | sequence-specific DNA binding |
| 5.20%       | −22.48%                  | −19         | GO:0046983 | Molecular function | protein dimerization activity |
| 5.60%       | −16.52%                  | −17         | GO:006351 | Biological process | transcription, DNA-templated |
| 5.30%       | −20.99%                  | −12         | GO:0003682 | Molecular function | chromatin binding |
| 5.70%       | −15.03%                  | −12         | GO:0003676 | Molecular function | nucleic acid binding |
| 6.20%       | −7.58%                   | −11         | GO:0005634 | Cellular component | nucleus |

Table 2. GO terms associated with SNP density. Analysis was performed on the 29M dataset; average mRNA mutation density is 6.7%. 
Here, we demonstrated patterns in genetic variability, established earlier for mammals and observed some novel trends. The rate of polymorphism in introns is known to be higher than in exons, which is consistent with evolutionary constraints in the mutability of protein-coding regions\(^1\). We confirmed this trend on a subset of common SNPs in rice (MAF > 0.01).

We identified promoter motifs associated with gene conservation using the cisExpress program\(^9\). We observed that TATA- promoters with TATA-boxes at –30 tend to have slightly fewer SNPs than those without the TATA-boxes. This might be due to the rigidly controlled expression mechanisms of these genes. TATA-box binding proteins recruit other transcription factors into the transcriptosome, with controlled expression profiles needed for genes of more conserved functions. This is consistent with our observation of the differences in distribution of SNPs between various GO classes: gene bodies and promoters of transcription-factor encoding genes have fewer SNPs than other functional classes. At the same time, a number of genes carry “alternative” TATA-boxes 100 or more nucleotides upstream from the usual location, and these regions feature higher SNP densities, perhaps indicating that they are evolving at faster rates. For example, TATA- genes annotated as having “kinase activity” have over 55 SNPs per 250 nt promoter region while TATA+ genes annotated as “sequence-specific DNA binding transcription factor” or “DNA-binding” have three or less. There is also a difference in cellular locations where variable TATA- genes predominate in vacuoles while conserved TATA+ genes occur preferably in cell walls.

In terms of gene families, we have shown that transcription factors feature unusually low SNP density agreeing with their importance of as hubs modulating various downstream cascades. The most genetically variable genes tend to encode membrane-bound proteins, ATPases and serine/threonine protein kinases, key regulators of plant responses to abiotic stresses. Similar trends were highlighted in comparative analysis of human and mouse genomes\(^5\). The relative conservation of transcription factors\(^6\), tier genetic variability in the genes responsible for interactions apparently correlates with the need to adjust rapidly changing environmental conditions.

Finally, CG-skew coupled with the associated transcriptional start sites, distributions of regulatory elements, expected lengths of UTR and other trends will be useful to refine gene prediction models in plants, especially the positions of TSSs and TSTs, the boundaries of gene models and the foundation for refinement of genotype to phenotype predictions.

**Materials and Methods**

We used a collection of SNPs obtained using the BW A-mem/GATK pipeline from the 3 K rice genome sequence data. SNP calls were made using the BW-MEM alignment of short reads against the Nipponbare IRGSP-1.0 RefSeq and GATK pipeline. SNPs are available from the SNP-Seek database (http://snp.seek.irri.org)\(^25\).

**SNP Filtering.** We downloaded 29M bi-allelic SNPs called on Nipponbare genome from the SNP-Seek portal. This data was used to generate datasets with varying quality cut-offs.

"Base" dataset (16M SNPs). We excluded SNPs detected in 5 genomes of Oryza glaberrima, restricting our analysis of O. sativa accessions. Twelve other genomes were excluded due to excessive amounts (i.e., >10,000) of heterozygous singleton SNPs. (see Supplementary Data (Supp. Fig. 1) for distribution of singleton heterozygotes per genome). Next, for each remaining SNP we measured the observed proportion of heterozygotes (H\(_{\text{obs}}\)) and computed the expected proportion of heterozygosity for its allele frequency (H\(_{\text{exp}}\)). The expected ratio H\(_{\text{obs}}\)/H\(_{\text{exp}}\) for rice is ∼0.05, because of high level of inbreeding; however, many SNPs had H\(_{\text{obs}}\)/H\(_{\text{exp}}\) > 1. SNPs that exhibit high degree of heterozygosity could represent alignment errors due to paralogs that do not occur in the reference genome. We estimated inbreeding coefficients for the 3 K dataset (F\(_{\text{3K}}\) = 0.9520), indica (F\(_{\text{indica}}\) = 0.9251), and japonica (F\(_{\text{japonica}}\) = 0.9689) as median value of 1 − H\(_{\text{obs}}\)/H\(_{\text{exp}}\) excluding all SNPs of low frequency (<0.05) and high heterozygosity (H\(_{\text{obs}}\)/H\(_{\text{exp}}\) < 10^(-10−1-F\(_{\text{3K}}\))). A number of genes annotated as having “kinase activity” have over 55 SNPs per 250 nt promoter region while TATA+ genes annotated as “sequence-specific DNA binding transcription factor” or “DNA-binding” have three or less. There is also a difference in cellular locations where variable TATA- genes predominate in vacuoles while conserved TATA+ genes occur preferably in cell walls.

"Filtered" dataset (5M). It was obtained by removing SNPs with MAF < 0.01 and maximum per-SNP missing rate 0.1 (PLINK options –maf 0.01 –geno 0.1). “Filtered” dataset contains 4,817,175 SNPs.

**Genomic data and gene selection.** The current MSUv7 annotation (http://rice.plantbiology.msu.edu/) of rice contains 55,986 predicted genes and 66,338 gene models\(^8\). Upon exclusion of pseudogenes, transposable elements, and genes with atypical lengths of 5’ UTR (below 20 nt or above 1000 nt long), we created a high-confidence subset of 20,367 expressed protein-coding genes. Rice genome coverage data were obtained from the Rice 3,000 Genomes project\(^9\). Genomic regions with zero coverage in all samples were excluded from the analysis. Next, the SNPs were stratified by quality and by genomic regions in terms of their coding, UTR, transcription start and termination, translation start and termination, promoter, etc. For each of the 20,367 genes the number of SNPs per nucleotide was by summing the number of SNPs from the 3000 genomes at this position. A subset of non-synonymous mutations was selected using the SnpEff\(^7\) tool.

**Gene ontology analysis.** Gene Ontology assignments were downloaded from the Gramene database (http://ftp.grame.com/CURRENT_RELEASE/data/ontology/go/go_ensembl_oryza_sativa_japonica.gaf.gz).
The map between MSU and RAP-DB locus names was downloaded from the RAP-DB site (http://rapdb.dna.affrc.go.jp/download/archive/RAP-MSU.txt.gz). Enrichment analysis was performed using one-sample t-test implemented in oracle as function stats_t_test_one. We excluded GO classes containing less than ten genes.

Identification of motifs in promoter. Motifs in promoter were found using the cisExpress tool, which is an improved and enhanced adaptation of an earlier algorithm, Motifer, specifically modified to process large datasets. cisExpress is based on an important assumption that function of promoter motifs is position specific. It works in two stages: (1) find ‘seed’ motifs associated with the phenotype of interest, and (2) optimize the ‘seed’ motifs using a genetic algorithm. cisExpress was originally developed to find motifs in promoter associated with gene expression patterns of interest. In this work, we used cisExpress in the ‘off-label’ fashion to find motifs in promoter associated with the SNP density. For each gene, two pieces of information were obtained: sequence of core promoter, defined as region \([-250, 50]\) around the TSS, and number of SNP in this region. Positions of TSS were obtained from the Rice Genome Annotation Project (http://rice.plantbiology.msu.edu/) database and validated using the NPEST algorithm.

Custom scripts for data analysis. Custom scripts for statistical data analysis and visualization were written in R and C++. In order to calculate the SNP density per genomic region, we wrote an R script that counted the number of SNPs per position for each of the 20,367 selected genes. The number of SNPs was divided by the total number of genes (20,367), producing the number of SNPs at each position per 1000 genes.

References

1. Schmidt, S. et al. Hypermutable non-synonymous sites are under stronger negative selection. PLoS genetics 4, e1000281, doi: 10.1371/journal.pgen.1000281 (2008).
2. Jain, M., Moharan, K. C., Shankar, R., Kumari, R. & Garg, R. Genomewide discovery of DNA polymorphisms in rice cultivars with contrasting drought and salinity stress response and their functional relevance. Plant Biotechnol J 12, 253–264, doi: 10.1111/pbi.12133 (2014).
3. Panunzi, L. G. & Aguero, F. A genome-wide analysis of genetic diversity in Trypanosoma cruzi intergenic regions. PLoS neglected tropical diseases 8, e2839, doi: 10.1371/journal.pntd.0002839 (2014).
4. Osada, N. & Wu, C. J. Inferring the mode of speciation from genomic data: a study of the great apes. Genetics 169, 259–264, doi: 10.1534/genetics.105.049231 (2005).
5. Palazzo, A. F. & Gregory, T. R. The case for junk DNA. Human mutation 33, 136–143, doi: 10.1002/humu.21616 (2010).
6. Hellman, A. & Chess, A. Extensive sequence-influenced DNA methylation polymorphism in the human genome. Epigenetics & chromatin 3, 11, doi: 10.1186/1756-8935-3-11 (2010).
7. Teykovitcha, S. et al. Human-macaque comparisons illuminate variation in neutral substitution rates. Genome biology 9, R76, doi: 10.1186/gb-2008-9-4-r76 (2008).
8. Hellmann, I. et al. Why do human diversity levels vary at a megabase scale? Genome research 15, 1222–1231, doi: 10.1101/gr.3461105 (2005).
9. Schabley, V. M. et al. The influence of genomic context on mutation patterns in the human genome inferred from rare variants. Genome research 23, 1974–1984, doi: 10.1101/gr.154971.113 (2013).
10. Linardopolou, E. V. et al. Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. Nature 437, 94–100, doi: 10.1038/nature04029 (2005).
11. Xia, J., Han, L. & Zhao, Z. Investigating the relationship of DNA methylation with mutation rate and allele frequency in the human genome. BMC genomics 13 Suppl 8, S7, doi: 10.1186/1471-2164-13-S8-S7 (2012).
12. Shabalina, S. A., Ogurtsov, A. Y., Rogozin, I. B., Koonin, E. V. & Lipman, D. J. Comparative analysis of orthologous eukaryotic mRNAs: potential hidden functional signals. Nucleic Acids Res 32, 1774–1782, doi: 10.1093/nar/gkh313 (2004).
13. Spicher, A. et al. Highly conserved RNA sequences that are sensors of environmental stress. Mol Cell Biol 18, 7371–7382 (1998).
14. Castle, J. C. SNPs occur in regions with less genomic sequence conservation. PLoS One 6, e20660, doi: 10.1371/journal.pone.0020660 (2011).
15. Atwell, S. et al. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature 465, 627–631, doi: 10.1038/nature08800 (2010).
16. Horton, M. W. et al. Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel. Nature genetics 44, 212–216, doi: 10.1038/ng.1042 (2012).
17. Weigel, D. & Nordborg, M. Population Genomics for Understanding Adaptation in Wild Plant Species. Annual review of genetics 49, 315–338, doi: 10.1146/annurev-genet-120213-092110 (2015).
18. Pasto, T. et al. Fine-scale population recombination rates, hotspots, and correlates of recombination in the Medicago truncatula genome. Genome biology and evolution 4, 726–737, doi: 10.1093/gbe/evs046 (2012).
19. Yoder, J. B. et al. Genomic signature of adaptation to climate in Medicago truncatula. Genetics 196, 1263–1275, doi: 10.1534/genetics.113.159319 (2014).
20. Alexandrov, N. et al. SNP-Seek database of SNPs derived from 3000 rice genomes. Nucleic Acids Res 43, D1023–D1027, doi: 10.1093/nar/gku1039 (2015).
21. McCouch, S. R. et al. Open access resources for genome-wide association mapping in rice. Nat Commun 7, 10532, doi: 10.1038/ncomms10532 (2016).
22. Huang, X. et al. A map of rice genome variation reveals the origin of cultivated rice. Nature 490, 497–501, doi: 10.1038/nature11532 (2012).
23. Xu, X. et al. Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nat Biotechnol 30, 105–111, doi: 10.1038/nbt.2050 (2012).
Scientific Reports | 6:35730 | DOI: 10.1038/srep35730

Nucleotide diversity analysis highlights functionally important genomic regions.

How to cite this article: Tatarinova, T. V. et al. Nucleotide diversity analysis highlights functionally important genomic regions. Sci. Rep. 6, 35730; doi: 10.1038/srep35730 (2016).

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/}

© The Author(s) 2016