LETTER

Ecological Nitrogen Limitation Shapes the DNA Composition of Plant Genomes

Claudia Acquisti,† James J. Elser,† and Sudhir Kumar*†

*Center for Evolutionary Functional Genomics, Biodesign Institute, Arizona State University; and †School of Life Sciences, Arizona State University

Phenotypes and behaviors respond to resource constraints via adaptation, but the influence of ecological limitations on the composition of eukaryotic genomes is still unclear. We trace connections between plant ecology and genomes through their elemental composition. Inorganic sources of nitrogen (N) are severely limiting to plants in natural ecosystems. This constraint would favor the use of N-poor nucleotides in plant genomes. We show that the transcribed segments of undomesticated plant genomes are the most N poor, with genomes and proteomes bearing signatures of N limitation. Consistent with the predictions of natural selection for N conservation, the precursors of transcriptome show the greatest deviations from Chargaff's second parity rule. Furthermore, crops show higher N contents than undomesticated plants, likely due to the relaxation of natural selection owing to the use of N-rich fertilizers. These findings indicate a fundamental role of N limitation in the evolution of plant genomes, and they link the genomes with the ecosystem context within which biota evolve.

Ammonia, nitrates, and nitrites constitute the primary sources of the nitrogen (N) atoms incorporated in the nitrogenous bases that comprise the building blocks of plant genomes (Berg et al. 2006). These sources of inorganic N are severely limited in the natural ecosystems (Elser et al. 2007), which may promote the use of bases requiring fewer N atoms. Thymine (T) contains two N atoms, cytosine (C) three, and adenine (A) and guanine (G) five. Although the avoidance of N-rich nucleotides in DNA sequences can reduce the N cost, the savings differ for double- and single-stranded sequences. Due to the conformational requirements of double-stranded DNA, genomic regions afford minimal opportunities for N conservation (0.5 N atoms per base): A and T require 3.5 N atoms per base and G and C require 4 N atoms per base in double-stranded configurations. In contrast, transcribed genomic regions give rise to single-stranded RNA in which the difference in the N cost is as high as three N atoms per base. Thus, the effects of natural selection for N conservation, if any, should be more pronounced in these templates.

To test these predictions, we analyzed the *Arabidopsis thaliana* (thale cress) genome and found that the transcribed segments of the genome have a significantly lower N content than the whole genome (5.3% difference; $P \ll 0.01$, *Z*-test; fig. 1*A*). Only 5% of the transcribed segments show a higher N content than the whole genome. In contrast, animal genomes and transcriptomes are almost identical in N content (<0.02% difference; table 1 and fig. 1*B*). This is reasonable because the biosynthesis of nucleotides in animals is not likely N limited because animals feed on an N-rich biomass already containing N in preformed amino acids (Berg et al. 2006). In these analyses, we excluded exons when estimating the N content of transcribed regions because their nucleotide composition is dictated by the protein sequence encoded (considered separately in the amino acid sequence analysis below) and by the secondary and tertiary structural constraints on the mature transcripts.

Key words: nitrogen limitation, plant genome, crops, biological stoichiometry.

E-mail: s.kumar@asu.edu.

Mol. Biol. Evol. 26(5):953–956. 2009
doi:10.1093/molbev/msp038
Advance Access publication March 2, 2009

In comparisons of N-limited (plants) and N-sufficient (animals) organisms, the difference in N content is an order of magnitude higher in transcriptomes than in genomes. This disparity reflects differences in the overall contribution of RNA and DNA to cellular biomass: DNA generally contributes less than 2% to overall organismal biomass, whereas RNA can constitute up to 15% of the biomass in multicellular eukaryotes (Sterner and Elser 2002; Elser et al. 2003). These results are also consistent with the observation that, on average, the most highly expressed proteins in plant species show the lowest N content, whereas in animals use of N-rich amino acids is not a function of the amount of gene expression (Elser et al. 2006). Plant proteomes contain 7% fewer ($P \ll 0.01$, Student's *t*-test) N atoms than the two animal taxa (table 1). Taken together, our results add to accumulating evidence for an imprint of natural selection for N conservation on a genomic scale in plants.

Comparison of local deviations from Chargaff's second parity rule in the transcriptome is an independent tool for testing the natural selection hypothesis for N conservation in plants. Chargaff's second parity rule asserts (approximate) equality of the frequencies of A and T nucleotides as well as the equality of frequencies of C and G nucleotides in any single strand of DNA. Deviations from this rule are known to have functional correlations, for example, with the direction of transcription (Bell and Forsdyke 1999; Paz et al. 2007). Under an N-conservation regime, we expect to find a large deviation from Chargaff's rule in the sense strand, such that it conserves N in plants. This is indeed the case: the *A. thaliana* transcriptome shows a 5-fold greater deviation toward low-N nucleotides than that in animals (fig. 1*C*; $P \ll 0.01$, *t*-test). As expected, whole-genome analyses do not show appreciable deviations from Chargaff's second parity rule. Furthermore, animal transcriptomes do not show significant deviations from Chargaff's second parity rule (fig. 1*C*).

Laboratory evidence for the role of natural selection for N conservation in the genomes of plants is not yet available. However, we can take advantage of a natural experiment by examining the N content of crop plant genomes as massive nitrogen enrichment by fertilization of cultivated soils is tantamount to removing the selection pressure exerted by N limitation. In this case, purifying selection would

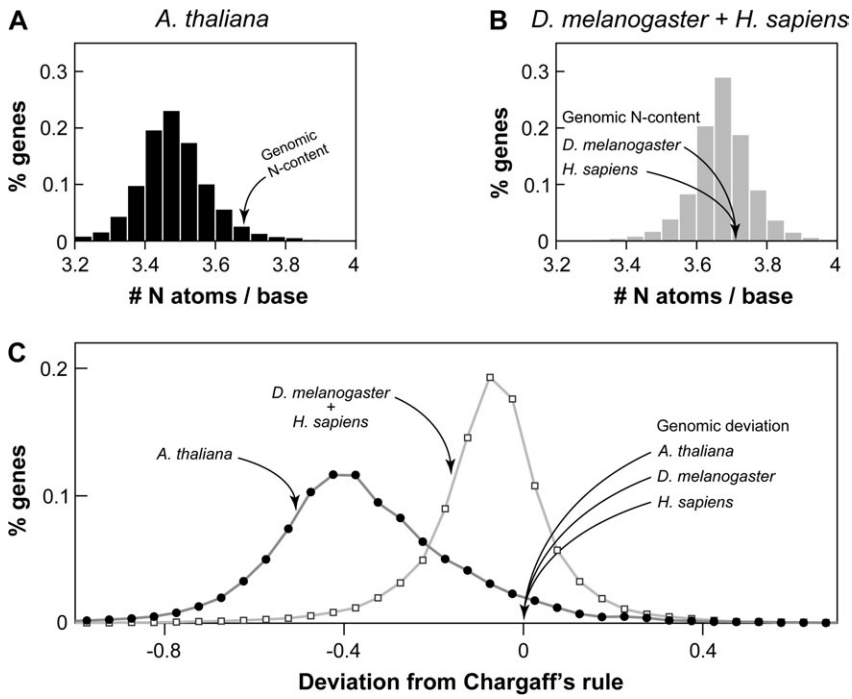

FIG. 1.—Patterns of N content in the DNA sequences of plant and animal genomes. The N contents for the transcriptome (histogram) and the whole genome are shown for a plant (*A*) and two animals (*B*). The mean and standard deviation (SD) of the transcriptomic N content derived from introns are given in table 1. Distributions of the deviations from Chargaff's second parity rule (difference between the N contents of the sense and antisense strands in introns) are shown in (*C*). The mean, SD, and standard error (respectively) for the deviations from Chargaff's second parity rule are *Arabidopsis thaliana* (−0.352, 0.226, and 0.001), *Drosophila melanogaster* (0.090, 0.191, and 0.001), and *Homo sapiens* (−0.058, 0.118, and 0.0007).

no longer act with high intensity against mutations leading to N-rich nucleotides and amino acids. Consequently, crop species should show a higher N content and lower deviation from Chargaff's rule than *A. thaliana*.

Analyses of the completely sequenced and annotated genome of domesticated rice (*Oryza sativa*) produces N content estimates that are significantly higher than those found in *A. thaliana* (fig. 2*A*; $P \ll 0.01$), which is consistent with our predictions. In addition, deviations from Chargaff's rule have been reduced by half in *O. sativa* relative to *A. thaliana* (fig. 2*B*). Furthermore, proteomic N content is higher in *O. sativa* than in *A. thaliana* ($P \ll 0.01$, *t*-test). In fact, a variety of domesticated species show higher proteomic N contents than undomesticated plants (fig. 2*C*). Plants harboring N-fixing bacteria also show a higher N content than the nondomesticated taxa (fig. 3). Interestingly, phylogenetically divergent crop species (e.g., the dicots *Medicago truncatula* and *Lotus japonicum*) show an N content that is more similar to a monocot crop (*Zea mays*) than to another undomesticated dicot (*A. thaliana*) (fig. 2*C*). Therefore, the patterns observed for rice should extend to other domesticated plants as well. It is remarkable that the removal of natural selection (via the use of fertilizers or the presence of N-fixing symbiotic bacteria) has quickly erased ancestral signatures of N limitation and altered the genomic architecture of many species at a fundamental level.

In summary, our findings directly implicate ecological limitations in altering the composition of molecular moieties associated with the flow of genetic information from genome to transcriptome. Thus, environmental growth limitations directly impact the biochemical structure of information storage and processing in both microbial and multicellular forms of life (McEwan et al. 1998; Baudouin-

Table 1
N Content of Genetic Informational Molecules

	Genome		Transcriptome				Proteome			
					N Content				N Content	
Species	No. of Bases	N Content	No. of Genes	No. of Bases	Mean	SD	No. of Proteins	No. of Residues	Mean	SD
Arabidopsis thaliana	119×10^6	3.680**	26,544	24×10^6	3.485**	0.113	31,921	13×10^6	0.364**	0.079
Drosophila melanogaster	128×10^6	3.711	18,042	113×10^6	3.643	0.101	20,736	11×10^6	0.386	0.085
Homo sapiens	2.8×10^9	3.704	25,022	1.5×10^9	3.701	0.076	25,305	14×10^6	0.377	0.078

NOTE.—N content is the number of N atoms per monomer. The number of genes is lower than the number of proteins because genes without introns cannot be used for transcriptome analyses. Only experimentally validated isoforms were included for each gene. Standard deviations (SD) are shown, wherever appropriate, because the standard errors are close to zero due to very large sample sizes.

** Cases where the observed differences between the plant and animal genomes were significant at $P \ll 0.01$ in Student's *t*-test or the Fisher's exact test.

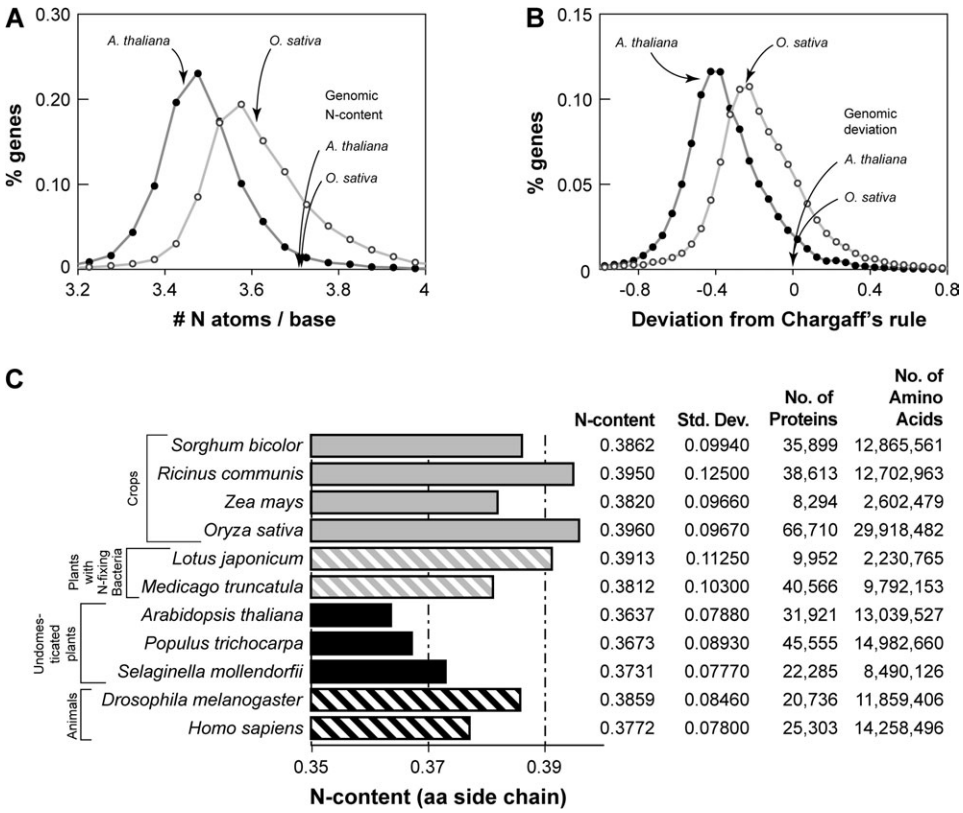

FIG. 2.—N contents in the genome, transcriptome and proteome of undomesticated and crop species. (*A*) Distribution of N content in the transcriptomes and (*B*) deviations from Chargaff's second parity rule in *Arabidopsis thaliana* (undomesticated) and *Oryza sativa* (crop). The mean, SD, standard error (SE), and sample size (number of genes) of N content distribution in (A) are *A. thaliana* (3.485, 0.114, 0.0007, and 26,544) and *O. sativa* (3.616, 0.139, 0.0006, and 54,712). The overall genomic N content was 3.680 (119×10^6 bp) for *A. thaliana* and 3.718 (372×10^6 bp) for *O. sativa*. In (B), the mean, SD, and SE of the difference between the N contents of the sense and antisense strands are −0.352, 0.226, and 0.001 for *A. thaliana*, and −0.171, 0.249, and 0.001 for *O. sativa*, with the whole-genome deviation close to zero in both species (*A. thaliana* = 0.0008 and *O. sativa* = 0.0002). (*C*) N content per amino acid side chain of protein sequences in crops plants known to be symbiotically related to N-fixing bacteria, undomesticated plants, and animals. The mean, SD, and the sample size (number of amino acids and proteins) for each species are shown on the right.

Cornu et al. 2001, 2004; Bragg and Hyder 2004; Bragg et al. 2006; Elser et al. 2006; Kolkman et al. 2006; Acquisti et al. 2007; Bragg and Wagner 2007). In the future, availability of completely sequenced and better annotated plant genomes (including those that are undomesticated, symbiotic with N-fixing bacteria, and crop species) will permit further quantification of the genomic impact of selection pressures due to N limitation across multiple plant phyla and ecosystems.

Methods

Completely sequenced and annotated genomes of *O. sativa* (rice, a crop species) and *A. thaliana* (thale cress) were used to represent plants. Although the sequencing of several other plant genomes is under way, the absence of well-assembled and well-annotated genomic data and gene models precludes their use at present. Genomic sequences and gene models were retrieved from the TIGR database (ftp://ftp.tigr.org/pub/data) for *O. sativa* (release 5.0) and from the TAIR database (ftp://ftp.arabidopsis.org/home/tair/) for *A. thaliana* (release TAIR 7). The available protein sequence data of many other domesticated and undomesticated plants were obtained for *Populus trichocarpa*, *Sorghum bicolor*, and *Selaginella moellendorffii* from the Eukaryotic Genomics database (http://genome.jgi-psf.org/) and for *Z. mays*, *M. truncatula*, *L. japonicum*, and *Ricinus communis* from the TIGR database (ftp://ftp.tigr.org/pub/data). The genomes of two highly divergent animal species were analyzed: *Homo sapiens* (a vertebrate) and *Drosophila melanogaster* (an invertebrate). Sequences and gene models were obtained from the UCSC database (ftp://hgdownload.cse.ucsc.edu/goldenPath/currentGenomes/) for *D. melanogaster* (Release 5, FlyBase Gene Models) and for *H. sapiens* (hg18, RefSeq Gene Models). In a previous analysis of many complete animal genomes (Elser et al. 2006), these genomes were found to be typical animal representatives, with the exception of species in the genus *Caenorhabditis*, which, for unknown reasons, stand out as outliers among animals in their atomic composition (see Elser et al. 2006; Acquisti et al. 2007). For the transcribed genomic segments (referred to as transcriptome), we used the sequences of introns in the protein-coding, tRNA, and rRNA genes for estimating transcriptome N content.

N content for DNA, RNA, and protein sequences was measured in atoms per base or amino acid residue using the formula $\sum(n_i \times p_i)$, where the n_i is the number of N atoms in the i-th base and p_i is the proportion of each base in the data ($\sum p_i = 1$). For genomes (double-stranded

DNA), $n_A = n_T = 3.5$ and $n_C = n_G = 4$. For transcriptomes (single-stranded RNA), $n_A = 5$, $n_T = 2$, $n_C = 3$, and $n_G = 5$. For amino acid side chains in proteomes, $n = 1$ for asparagine, glutamine, lysine, and tryptophan; $n = 2$ for histidine; $n = 3$ for arginine; and $n = 0$ for the rest. For the transcribed sequences, the strand bias (deviation from Chargaff's second parity rule) for each gene was calculated as the difference between the N contents of the transcript (sense strand) and its complement (antisense strand).

The lack of dependence of N content on the genomic G + C content has been discussed previously (Elser et al. 2006) for proteomes. We found that this extends to the transcriptome as well. The genomic G + C content of *O. sativa* was the highest of all the species we examined, but its transcriptome showed a significantly lower N content per nucleotide than animals, which contrasts with an expectation of high N content if G + C content dictated the N content. In addition, the analysis of the deviation from Chargaff's second parity rule (figs 1, 2*A*, and 2*B*) showed that the strand bias composition, and not the GC content alone, was the main factor determining the difference in the patterns observed in the transcriptome of plants and animals.

Acknowledgments

We thank William Fagan, James Gilbert, Thomas Wiehe, Alan Filipski, Peter Stadler, Antonio Marco, Fabia Battistuzzi, Bernhard Haubold, and Gregory Babbitt for scientific discussions; Bernard Van Emden and Revak Raj Tyagi for technical support; and Kristi Garboushian for editorial support. This work was funded by the National Science Foundation (J.J.E. and S.K.) and National Institutes of Health (S.K.).

Literature Cited

Acquisti C, Kleffe J, Collins S. 2007. Oxygen content of transmembrane proteins over macroevolutionary time scales. Nature. 445:47–52.

Baudouin-Cornu P, Schuerer K, Marliere P, Thomas D. 2004. Intimate evolution of proteins. Proteome atomic content correlates with genome base composition. J Biol Chem. 279:5421–5428.

Baudouin-Cornu P, Surdin-Kerjan Y, Marliere P, Thomas D. 2001. Molecular evolution of protein atomic composition. Science. 293:297–300.

Bell SJ, Forsdyke DR. 1999. Deviations from Chargaff's second parity rule correlate with direction of transcription. J Theor Biol. 197:63–76.

Berg J, Tymoczko J, Stryer L. 2006. Biochemistry. New York: W.H. Freeman & Co Ltd.

Bragg JG, Hyder HC. 2004. Nitrogen versus carbon use in prokaryotic genomes and proteomes. Proc R Soc Lond B. 271:S374–S377.

Bragg JG, Thomas D, Baudouin-Cornu P. 2006. Variation among species in proteomic sulphur content is related to environmental conditions. Proc R Soc Lond B. 273:1063–1070.

Bragg JG, Wagner A. 2007. Protein carbon content evolves in response to carbon availability and may influence the fate of duplicated genes. Proc R Soc Lond B. 274:1293–1300.

Elser JJ, Acharya K, Kyle M, et al. (12 co-authors). 2003. Growth rate-stoichiometry couplings in diverse biota. Ecol Lett. 6:936–943.

Elser JJ, Fagan WF, Subramanian S, Kumar S. 2006. Signature of ecological resource availability in the animal and plant proteomes. Mol Biol Evol. 23:1946–1951.

Elser JJ, Bracken MES, Cleland EE, et al. (10 co-authors). 2007. Global analysis of nitrogen and phosphorus limitation of primary producers in freshwater, marine and terrestrial ecosystems. Ecol Lett. 10:1135–1142.

Kolkman A, Daran-Lapujade P, Fullaondo A, Olsthoorn MMA, Pronk JT, Slijper M, Hecket AJR. 2006. Proteome analysis of yeast response to various nutrient limitations. Mol Syst Biol. 2:doi: 10.1038/4100069

McEwan CE, Gatherer D, McEwan NR. 1998. Nitrogen-fixing aerobic bacteria have higher genomic GC content than non-fixing species within the same genus. Hereditas. 128:173–178.

Paz A, Mester D, Nevo E, Korol A. 2007. Looking for organization patterns of highly expressed genes: purine-pyrimidine composition of precursor mRNAs. J Mol Evol. 64: 248–260.

Sterner RW, Elser JJ. 2002. Ecological stoichiometry. Princeton (NJ): Princeton University Press.

Carlos Bustamante, Associate Editor

Accepted February 24, 2009