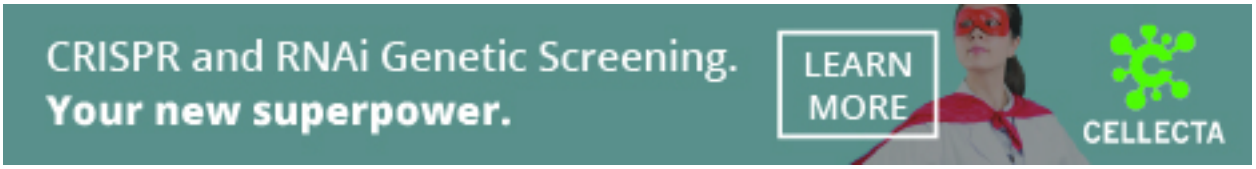
Genomics: More Than the Sum of the Parts

Pablo D. Rabinowicz and Ravi Scahidanandam

Genome Res. 2002 12: 1015-1016
Access the most recent version at doi:10.1101/gr.432502



Genomics: More Than the Sum of the Parts

Pablo D. Rabinowicz[1] and Ravi Scahidanandam[2]

Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA

It has been known for some time that DNA composition varies across a given genome as well as between genomes (Filipski et al. 1973; Wagner and Capesius 1981). Genomic sequencing projects allow this observation to be confirmed at the sequence level (The Arabidopsis Genome Initiative 2000; Ashikawa 2001). However, the cause and function of these compositional differences are still obscure. Among the theories that may explain these phenomena (Eyre-Walker and Hurst 2001), mutation bias from C to T due to deamination of methylated C has been commonly used to account for them (Coulondre et al. 1978). As methylation is probably involved in a mechanism to silence transposable elements (Martienssen 1998), it makes sense that inactive methylated transposons can easily undergo C to T transition because they are under no selective pressure. However, this theory cannot explain other related compositional biases such as the CpG suppression observed in animal mitochondria (Cardon et al. 1994), where there is no DNA methylation.

The recent completion of the draft sequence of the genome of the rice *indica* subspecies (Yu et al. 2002) allowed Wong and coworkers (2002) to uncover a new kind of fine-scale GC heterogeneity. By analyzing in detail GC frequencies in a collection of rice full-length cDNAs and aligning them to the genome, they discovered that genes are richer in GC at the 5′ end than at the 3′ end. Interestingly, this trend is not only observed in the coding sequence but also in introns. As a consequence of these GC gradients, codon and amino acid usage are also affected, showing 5′ to 3′ gradients. When testing this observation against other plant genes, they found GC gradients in all grasses tested but not in the phylogenetically distant dicots. To see this phenomenon, a careful sequence analysis must be performed using a window size smaller than the average gene to scan intragenic GC frequencies. Traditionally, GC content is measured in longer stretches of DNA, which would overlook such fine-scale gradients. In a recent report on GC content among different plants genomes, these 5′ to 3′ gradients in GC frequency were not detected even using a small window, because the study was focused on CpG islands (Ashikawa 2001).

In addition to the old questions on the genomic GC bias, this discovery certainly prompts speculation on the reasons why grass genes show these GC and codon usage gradients whereas dicots do not. For example, what is the biological significance of these gradients? What proportion of all rice genes show them? Are grass genes clustered according to the presence of compositional gradients? Is there a connection between the lower GC content in dicots and their lacking gradients? A single discovery can raise many new questions or, as a lawyer in a Coen brothers' movie put it, "the more we look, the less we really know".

A related observation was made by Yu et al. (2002) when performing sequence similarity searches between rice and *Arabidopsis*. Using `TBLASTN`, they saw that for about 80% of *Arabidopsis* genes, a homolog in rice could be found. However, only nearly 50% of rice genes showed a homolog in *Arabidopsis*. Yu and coworkers proposed that the gradients in amino acid usage may be part of the reason why so many rice genes do not find a match in *Arabidopsis*. Additional, not mutually exclusive possibilities may also explain this fact. One of these possibilities was observed in the sequence of the rice *japonica* subspecies, whose draft sequence was published at the same time (Goff et al. 2002). The analysis of this version of the rice genome showed a similar situation in terms of homology between rice and *Arabidopsis* genes. In this case, most of the rice genes with no match in *Arabidopsis* were low-evidence, predicted genes. So, some of them may not be genes at all. Another possibility that can explain part of the asymmetry between these two plant genomes is that a fraction of these rice genes without a homolog in *Arabidopsis* actually corresponds to previously unknown rice-specific transposable elements that are decayed and/or in low copy number. It is not unusual that hypothetical genes annotated at early stages of a genome sequencing project turn out to be repeats when annotation of the same or other genomes is improved. Some of such repetitive elements could be active and thus expressed. In this way they could be present in cDNA libraries used to help gene annotation.

Perhaps more immediate is the impact of the discovery of these GC gradients on gene annotation. Annotation is often controversial because of its importance as the link between sequence and biology (Stein 2001). In particular, the presence of the GC gradients may affect the accuracy of gene prediction software. Gene modeling programs typically rely on previously known information, called the training data set, about the genome under analysis. Such approaches are thus only as good as their training sets. To build gene models, the software uses the training set to extract statistics for features such as compositional bias and codon and dicodon usage (which are peculiar for each organism) in exonic, intronic, and nongenic regions (Milanesi and Rogosin 1998). After their discovery by Wong and coworkers (2002), GC gradients become a compositional feature to be incorporated into the training sets. However, because the compositional gradient is not observed in all genes, two training sets will probably be needed: one for genes with GC gradients (where the codon bias would change with distance from the 5′ end) and another for genes without them. The use of a single training set which includes both kinds of genes may create an average statistic for codon usage that does not reflect the reality for either type of genes. Prior knowledge of the peculiarities of grass genes, such as the presence of GC gradients, will allow for much more accurate gene predictions, not just for rice, but also for other grass genomes that may be sequenced in the future.

Light will be shed on many of the uncertainties posed by these findings when the highly accurate sequence of the rice genome is completed by the International Rice Genome Sequencing Project (IRGSP, Sasaki and Burr 2000). Certainly new questions will arise from the new data but more importantly, what we learn from the rice genome will pave the way for tackling other plant genomes, which are the future targets for partial or complete sequence. Maize, which will probably be the next grass genome to be ap-

[1] Corresponding author.
E-MAIL rabinowi@cshl.edu; FAX (516) 367-8369.
[2] sachidan@cshl.org.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.432502.

proached for sequencing (Bevan 2002), will undoubtedly demonstrate that the biological knowledge gathered from both rice and maize sequences is incredibly bigger than the sum of the two pieces of data. Putting it in a lawyer's terms, the more we look, the more we can predict.

ACKNOWLEDGMENTS

We thank Erik Vollbrecht for critical reading of the manuscript.

REFERENCES

Ashikawa, I. 2001. Gene-associated CpG islands in plants as revealed by analyses of genomic sequences. *Plant J.* **26:** 617–625.

Bevan, M. 2002. The first harvest of crop genes. *Nature* **416:** 590–591.

Cardon, L.R., Burge, C., Clayton, D.A., and Karlin, S. 1994. Pervasive CpG suppression in animal mitochondrial genomes. *Proc. Natl. Acad. Sci.* **91:** 3799–3803.

Coulondre, C., Miller, J.H., Farabaugh, P.J., and Gilbert, W. 1978. Molecular basis of base substitution hotspots in *Escherichia coli*. *Nature* **274:** 775–780.

Eyre-Walker, A. and Hurst, L.D. 2001. The evolution of isochores. *Nat. Rev. Genet.* **2:** 549–555.

Filipski, J., Thiery, J.P., and Bernardi, G. 1973. An analysis of the bovine genome by Cs2SO4-Ag density gradient centrifugation. *J. Mol. Biol.* **80:** 177–197.

Goff, S.A., Ricke, D., Lan, T.H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H., et al. 2002. A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). *Science* **296:** 92–100.

Martienssen, R. 1998. Transposons, DNA methylation and gene control. *Trends Genet.* **14:** 263–264.

Milanesi, L. and Rogozin, I.B. 1998. In *Guide to human genome computing*, 2nd ed. (ed. M.J. Bishop), pp. 215–259. Academic Press, San Diego, California.

Sasaki, T. and Burr, B. 2000. International Rice Genome Sequencing Project: The effort to completely sequence the rice genome. *Curr. Opin. Plant Biol.* **3:** 138–141.

Stein, L. 2001. Genome annotation: From sequence to biology. *Nat Rev Genet.* **2:** 493–503.

The Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408:** 796–815.

Wagner, I. and Capesius, I. 1981. Determination of 5-methylcytosine from plant DNA by high-performance liquid chromatography. *Biochim. Biophys. Acta* **654:** 52–56.

Wong, G.K., Wang, J., Tao, L., Tan J., Zhang J., Passey D.A., and Yu, J. 2002. Compositional gradients in *Gramineae* genes. *Genome Res.* **12:** 851–856.

Yu, J., Hu, S., Wang, J., Wong, G.K., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X., et al. 2002. A draft sequence of the rice genome (*Oryza sativa L. ssp. indica*). *Science* **296:** 79–92.