

# The Dark Side of the Genome: Revealing the Native Transposable Element/Repeat Content of Eukaryotic Genomes

Dear Editor,

The majority of genome assemblies to date fail to represent the true structure of native genomes. This lack of completeness is largely due to the inability to assemble the variable (often significant) fraction of nuclear genomes that is composed primarily of repeated sequences (with either a structural function such as satellite DNA and simple sequence repeats or “selfish DNA” such as high-copy transposable elements [TEs]), herein defined as the “dark side of the genome.” To address this problem, we developed a method to detect and quantify the dark side of the genome and used it to infer the genomic composition and dynamic evolution of the majority of native repeats and TEs present within several test eukaryotic genomes.

Eukaryotic genomes range in size by about four orders of magnitude, with flowering plants having the widest variation (Fedoroff, 2012). Without taking into account whole-genome duplication and polyploidization events, it is well established that genome size is highly correlated with TEs and repeated sequences (e.g., centromeric, satellite, ribosomal DNA) content (Kidwell, 2002). TE activity and dynamics can have major effects on the host’s genetic material by acting as mutagenic agents, as substrates for inducing changes in gene content and regulation, and as a general source of genetic variability (Lisch, 2013).

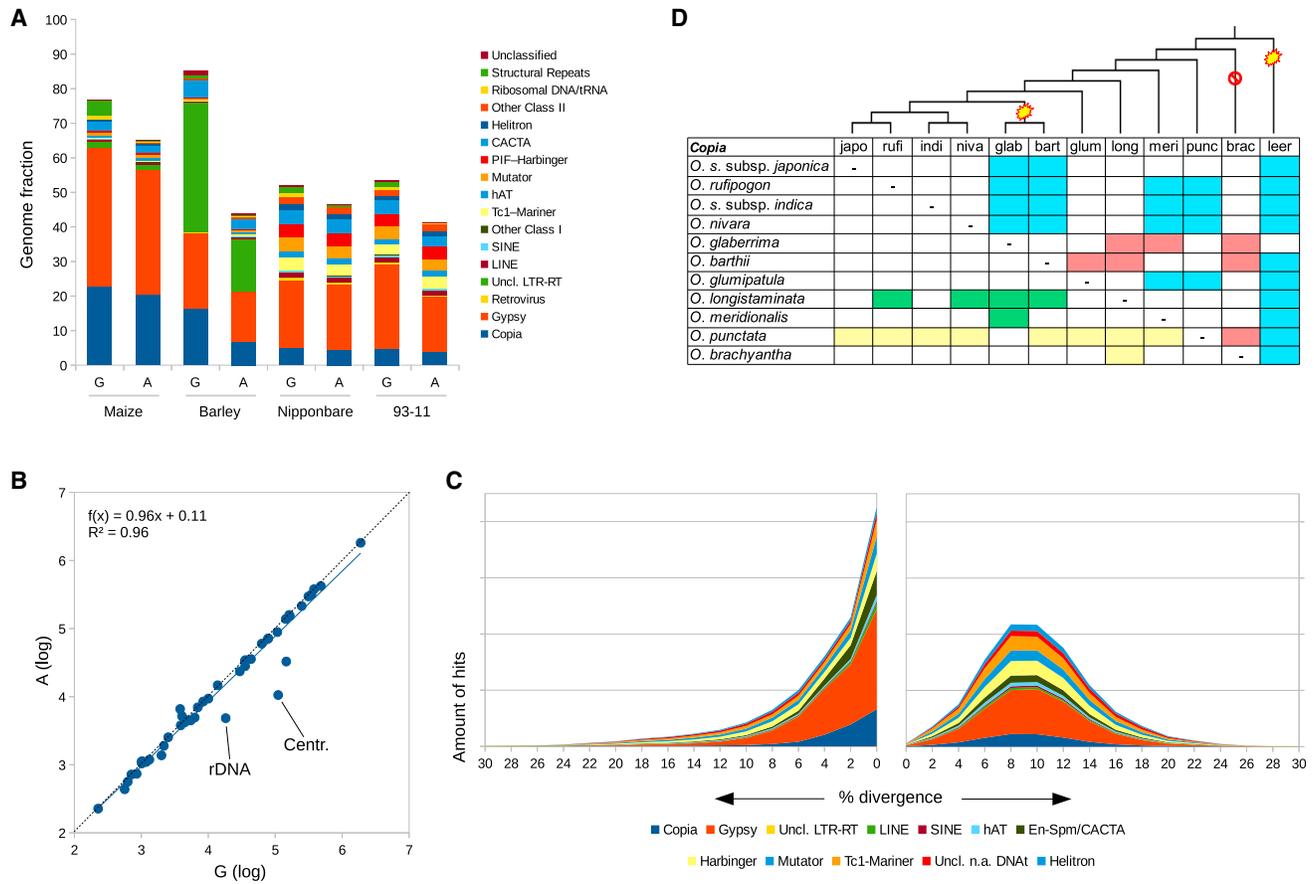
Despite the documented importance of repetitive sequences and TEs in genome biology, the majority of publicly available eukaryotic genome assemblies today lack high-quality and comprehensive representations of these sequences. In some extreme cases, significant portions of a genome’s repeat fraction are barely present in its draft assembly. This results in biased analyses of genome composition, based solely on the assembled portion of a given genome.

The goal of our study was to develop a method to discover (with the least bias conceivable) the total repetitive/TE sequence content of native genomes (including the unexplored dark side), and to compare these results with the TE/repeat content of a corresponding set of genome assemblies. To accomplish these goals we aligned sets of either unassembled or genome-derived single-sequence reads (i.e., a low-coverage short-read genome skim) to highly curated repeat libraries and tallied the hits to each repeat/TE category. Of note, since the comprehensiveness and accuracy of classification of a repeat library is crucial in identifying and quantifying all of the genomic elements, we developed repeat libraries using orthogonal approaches (both structure- and homology-based methods; see Copetti et al., 2015 and this

work) on all species surveyed. A more detailed description of our methods, datasets, and software adopted is outlined in Supplemental Figure 1 and Supplemental Information.

The serial application of our method to a set of 16 heterogeneous assemblies (Supplemental Table 1) led to three key observations: (1) although variable, the amount of repeats and TEs in a given assembly (A) was consistently lower than those found in the corresponding genome (G) (Figure 1A, Supplemental Figure 2, and Supplemental Table 1); (2) genome assemblies were identified where the repeat and TE differences between A and G were negligible, or were statistically significant (i.e., the repeat composition of A was significantly under-represented with respect to G, Supplemental Table 2); and (3) these differences were associated with the assembly strategy rather than with the size of the genome. The first finding was expected, as genome assemblers often fail to unambiguously place repeats and discard them from the assembly. The latter two observations should prompt investigators to consider not only the assembly, but also to include the dark side of the genome when describing a genome assembly as a true representation of a native genome. For example, our analyses of the maize (genome size [GS] ~2.7 Gb) and barley (GS ~5.4 Gb) genome assemblies demonstrated that the assumption that all large genome assemblies are depleted in repeats is not supported if a local assembly strategy is applied to assemble a genome (such as BAC-by-BAC rather than a whole-genome shotgun).

Plotting the log values of G and A abundances provided another view as to how to quantify the completeness of an assembly. When the majority of repeat/TE types are assembled accurately, the quantity of G and A will be the same and will graph along the bisector (Figure 1B and Supplemental Figure 3A). When repeat/TE sequences present in G are not assembled in A, the data will scatter (toward higher G values), thereby decreasing the coefficient of determination, which also affects the line equation (Supplemental Figure 3B and 3C). In addition to reaching the conclusion drew by Ross-Ibarra’s group (Figure 1A in Tenailon et al., 2011), our data suggest that  $R^2$  is not the only metric to consider when measuring differences among repeat abundances. Our results on 16 species systematically confirmed (see Figure 1B and Supplemental Figure 3A for some examples) how the content imbalance in G and A affects the line equation, and that all components of the



**Figure 1. Analysis of the Dark Side of the Genome Reveals Structural and Evolutionary Properties of Genomes.**

(A) Bar graph presenting the total and specific amounts of repeated sequences and transposable elements in native genome and assembled genomes (columns G and A, respectively) of four species. More examples are given in Supplemental Figure 2.

(B) Plot of log values of the hit counts in native genome reads (G, x axis) versus genome assembly reads (A, y axis) in *Oryza sativa* subsp. *japonica* Nipponbare. A linear regression (blue line) and the coefficient of determination describe the properties of each species, i.e., how well the native repeats are represented in the genome assembly. For comparison, a dashed line representing a perfect 1:1 relation is shown. Only repeat categories significantly over-represented in G are detailed. Plots for other species are given in Supplemental Figure 3.

(C) The distribution of repeat similarity *Oryza barthii* in the native genome and the assembly (left and right panels, respectively). Distributions are mirrored from the center of the image with increasing values of divergence. See also Supplemental Figure 4.

(D) Matrix of *Copia* retroelement abundance between *Oryza* species and *Leersia perrieri*. Red and blue cells represent over- and under-represented values in the native genome, respectively. Yellow and green species combinations depict over- and under-represented values in the assembly, respectively. Combining tree topology (obtained from Geering et al., 2014) and the abundance pattern, proliferation (yellow bursts) and loss (red circle) events can be inferred and placed in time. Supplemental Figure 5 contains comparisons for the other types of repeats.

regression must be taken into account when describing the features of a given genome.

By measuring similarity among reads, our method could also detect the presence of different quantities of repeats and TEs across species, and provided information on the amount of recently duplicated sequences in a genome and their representation in an assembly. For example, in all species examined, footprints of recent and possibly ongoing TE activity could be observed in the large amount of hits with low sequence divergence (left panels of Figure 1C and Supplemental Figure 4). The different profiles obtained from assembly-derived reads (right panels of Figure 1C and Supplemental Figure 4) revealed that most of these recently duplicated sequences are not present in their corresponding assemblies, a known flaw of assembly algorithms that our analyses were able to demonstrate in a practical example.

By comparing differential repeat/TE abundance in pairs of closely related *Oryza* genomes, our method was also able to detect instances of significant over- or under-representation. Looking at the resulting matrix together, the overall patterns were coherent with the species' phylogeny of the *Oryza* genus (Figure 1D and Supplemental Figure 5), enabling us to detect signatures of past repeat/TE burst and sequence removal. Moreover, each case showed an evolutionary pattern that was independent of both the host genome and other repeat/TE class evolution.

In conclusion, using low-coverage SGS reads as a proxy for genome composition, we developed an “assembly-independent” method to quantify repetitive sequences and TEs in native genomes. Previous studies (Tenailon et al., 2011; Sveinsson et al., 2013) attempted to describe genome features by adopting a similar principle, but in our opinion their results are affected by the incompleteness of the repeat libraries used in

terms of both the species represented and the lack of non-assembled repeats. With the additions and modifications implemented here, we demonstrated that our approach can detect and quantify repeats and TEs in a more comprehensive way. Our analyses demonstrated that by using a standard set of bioinformatics tools, coupled with highly curated and comprehensive repeat libraries, the native repeat and TE content of a given genome can be easily measured in a routine fashion. We also demonstrated how the genome assembly strategy, and not the genome size *per se*, has a major impact on the repeat content present in a given genome assembly. Lastly, given that repeats are a key component of eukaryotic genomes, we emphasize how important it is to specify how inclusive a repeat analysis is, and especially to distinguish whether surveyed repeats represent the content of a whole genome or an assembly.

### SUPPLEMENTAL INFORMATION

Supplemental Information is available at *Molecular Plant Online*.

### FUNDING

This work was supported by the US National Science Foundation under grants #0321678, #0638541, #0822284, and #1026200, the Bud Antle Endowed Chair of Excellence in Agriculture and Life Sciences, and the AXA Chair for Evolutionary Genomics and Genome Biology to R.A.W. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the US National Science Foundation.

### AUTHOR CONTRIBUTIONS

Conceptualization, D.C.; Methodology, D.C.; Validation, D.C.; Investigation, D.C. and R.A.W.; Writing – Original Draft, D.C.; Writing – Review & Editing, D.C.; Funding Acquisition, R.A.W.

### ACKNOWLEDGMENTS

We would like to thank Robert Edwards and Isaak Jenkins for their statistical consultation, and Olivier Panaud, Petr Novak, and Florian Maumus for their fruitful comments. No conflict of interest declared.

Received: June 10, 2016  
Revised: September 12, 2016  
Accepted: September 22, 2016  
Published: September 27, 2016

**Dario Copetti<sup>1,2</sup> and Rod A. Wing<sup>1,2,\*</sup>**

<sup>1</sup>Arizona Genomics Institute, BIO5 Institute and School of Plant Sciences, University of Arizona, Tucson, AZ 85721, USA

<sup>2</sup>International Rice Research Institute, T.T. Chang Genetic Resources Center, Los Baños, Laguna 63108, Philippines

\*Correspondence: Rod A. Wing ([rwing@mail.arizona.edu](mailto:rwing@mail.arizona.edu))  
<http://dx.doi.org/10.1016/j.molp.2016.09.006>

### REFERENCES

- Copetti, D., Zhang, J., El Baidouri, M., Gao, D., Wang, J., Barghini, E., Cossu, R.M., Angelova, A., Maldonado L, C.E., Roffler, S., et al. (2015). RiTE database: a resource database for genus-wide rice genomics and evolutionary biology. *BMC Genomics* **16**:538.
- Fedoroff, N.V. (2012). Transposable elements, epigenetics, and genome evolution. *Science* **338**:758–767.
- Geering, A.D.W., Maumus, F., Copetti, D., Choisne, N., Zwickl, D.J., Zytnicki, M., McTaggart, A.R., Scalabrin, S., Vezzulli, S., Wing, R.A., et al. (2014). Endogenous florendoviruses are major components of plant genomes and hallmarks of virus evolution. *Nat. Commun.* **5**:5269.
- Kidwell, M.G. (2002). Transposable elements and the evolution of genome size in eukaryotes. *Genetica* **115**:49–63.
- Lisch, D. (2013). How important are transposons for plant evolution? *Nat. Rev. Genet.* **14**:49–61.
- Sveinsson, S., Gill, N., Kane, N.C., and Cronk, Q. (2013). Transposon fingerprinting using low coverage whole genome shotgun sequencing in cacao (*Theobroma cacao* L.) and related species. *BMC Genomics* **14**:502.
- Tenaillon, M.I., Hufford, M.B., Gaut, B.S., and Ross-Ibarra, J. (2011). Genome size and transposable element content as determined by high-throughput sequencing in maize and *Zea luxurians*. *Genome Biol. Evol.* **3**:219–229.