

## RESEARCH HIGHLIGHT



## The rice pangenome branches out

Kenneth M. Olsen <sup>1</sup>✉

© CEMCS, CAS 2022

*Cell Research* (2022) 32:867–868; <https://doi.org/10.1038/s41422-022-00699-7>

**The power of genome-enabled research depends not only on the quality of individual genome assemblies, but also the extent to which they collectively represent the natural variation of a study system. In a recent *Cell Research* paper, Shang et al. make an important advance in our understanding of rice genomes with the release of the first rice ‘super-pangenome’, which encompasses the pangenome of domesticated Asian rice plus three closely related species.**

This year marks two decades since the first draft genome of any crop species was released. The 2002 publication of the *japonica*<sup>1</sup> and *indica*<sup>2</sup> rice draft genomes was a seminal event in plant biology, setting the stage for the first completed rice reference genome<sup>3</sup> and an explosion in studies in fields ranging from rice genetics and breeding to genome evolution and phylogenetics of the grass family. As a diploid species with a relatively small genome size, the rice reference genome also provided a critical foundation for tackling more complicated plant genomes, including the polyploid genomes that characterize many crop species.

The rice reference genome was one of the last to be generated by assembling Sanger-sequenced BAC and PAC clones — an expensive and laborious approach that was superseded by the arrival of high-throughput “next-gen” short-read sequencing technology. This shift to short-read sequencing led to a massive increase in genome sequence data. But it also came at a cost, as short-read data are poorly suited for accurately sequencing genomic regions of structural variations (SVs), including presence/absence variations (PAVs), copy number variations (CNVs), and chromosome-level variations such as inversion polymorphisms and segmental duplications. As genome-enabled research advanced through the 2000’s and 2010’s, it became increasingly apparent that SVs and other difficult-to-characterize variants are not only abundant in eukaryotic genomes<sup>4</sup> but also important contributors to phenotypic variation, ranging from crop domestication traits<sup>5</sup> to human disease risk.<sup>6</sup>

At the same time that the importance of SVs was becoming recognized, there was also increasing realization that a single reference genome, even of the highest quality, is insufficient for genomic characterization of a species. The problem here — which is compounded by reliance on short-read sequences — is that a single reference genome creates an ascertainment bias against detecting SVs and other variations in individuals that genetically diverge from the reference genotype.<sup>7</sup> The result is to systematically miss rare variants and those characterizing the subpopulations that are often the least well studied in the species.

An important remedy for both of these problems has come in the last several years with technical advances in long-read sequencing, which can more accurately capture previously-missed SVs and other genomic complexity, and by projects focused on the collective

grouping of these high-quality assemblies into pangenomes, which (ideally) represent the complete range of genetic variation, including SVs, within a gene pool. Importantly, in the case of crops, this gene pool often extends beyond individuals of the single domesticated species to its wild progenitor(s) as well as other reproductively-compatible wild relatives — all of which are potential contributors to the genetic and phenotypic diversity of the domesticate. Recognizing the need to consider this broader gene pool in the crop pangenome, Khan et al. in a 2019 commentary,<sup>7</sup> called for the development of phylogenetically-expanded pangenomes for crop species, coining the term “super-pangenome” to describe pangenomes of crop species that include their broader gene pool of wild relatives.

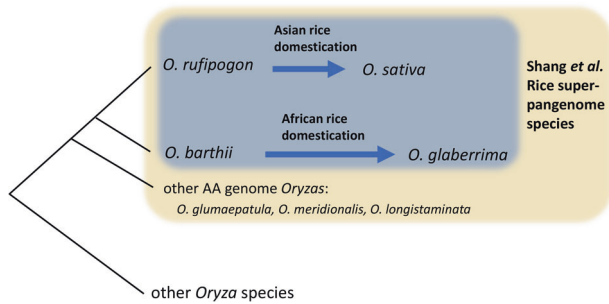
Development of a crop super-pangenome is a quantum advance in principle but not so easily put into practice. Now, as described in their *Cell Research* paper,<sup>8</sup> Shang et al. have made the super-pangenome a reality for the first time in rice — the cereal crop that feeds more people on the planet than any other. For this endeavor the authors focused on Asian rice (*O. sativa*) and a key subset of three other ‘AA genome’ *Oryza* species that are phylogenetically closest to the crop: its wild progenitor (*O. rufipogon*), the closely related but independently-domesticated African cultivated rice (*O. glaberrima*), and the African domesticate’s wild progenitor (*O. barthii*) (Fig. 1).

Having a super-pangenome that includes two domesticated species plus their respective wild progenitors creates an especially powerful dataset for examining the genetic underpinnings of domestication. For example, the authors are able to demonstrate that independent selection for reduced seed shattering (an important domestication trait in cereals) involved selection acting in parallel on the same genetic network in both rice species, and that this occurred at least partly through selection targeting the same *SHAT1* ortholog in the two domesticated species.

This rich new pangenome dataset for rice also greatly enhances our knowledge of the tremendous genetic diversity contributed by SVs in *Oryza* genomes. This knowledge can have immediate practical value. For example, incorporating SVs into association analyses can not only greatly improve the efficiency of trait mapping, but it can also be critical for identifying the underlying causal genetic variants. As an illustration of this approach, the authors documented that a 1.3 kb deletion polymorphism at a candidate locus for thousand-grain-weight in Asian rice appears to directly affect gene expression and the resulting grain phenotype.

Perhaps an equally important finding from Shang et al.’s analyses is that the diversity captured in their sample of 251 *Oryza* genomes, while impressive, is far from exhaustive. Their simulation analyses suggest that among their four sampled species, expanded sampling

<sup>1</sup>Department of Biology, Washington University in St. Louis, St. Louis, MO, USA. ✉email: kolsen@wustl.edu



**Fig. 1 The rice super-pangenome.** Phylogenetic breadth of the AA genome *Oryza* species represented in the rice super-pangenome of Shang et al.<sup>8</sup>

of the wild crop progenitors is especially warranted. Beyond its value for basic research, augmenting ex-situ collections of wild progenitor populations also holds immediate practical value, as the genetic resources offered by wild germplasm for crop improvement can only be as good as the diversity represented in

available germplasm.<sup>9</sup> The germplasm and data already encompassed by Shang et al.'s study can provide an excellent scaffold for further rice super-pangenome initiatives.

## REFERENCES

1. Goff, S. A. et al. *Science* **296**, 92–100 (2002).
2. Yu, J. et al. *Science* **296**, 79–92 (2002).
3. International Rice Genome Sequencing Project. *Nature* **436**, 793–800 (2005).
4. Feuk, L., Carson, A. R. & Scherer, S. W. *Nat. Rev. Genet.* **7**, 85–97 (2006).
5. Muñoz-Amatriain, M. et al. *Genome Biol.* **14**, R58 (2013).
6. McCarroll, S. A. & Altshuler, D. M. *Nat. Genet.* **39**, S37–42 (2007).
7. Khan, A. W. et al. *Trends Plant Sci.* **25**, 148–158 (2020).
8. Shang, L. et al. *Cell Res.* <https://doi.org/10.1038/s41422-022-00685-z> (2022).
9. Ramirez-Villegas, J. et al. *Nat. Plants* **8**, 491–499 (2022).

## ADDITIONAL INFORMATION

**Correspondence** and requests for materials should be addressed to Kenneth M. Olsen.

**Reprints and permission information** is available at <http://www.nature.com/reprints>