Joint press release of the Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), the Martin Luther University Halle-Wittenberg (MLU) and the Leibniz Institute of Plant Biochemistry (IPB)

Summary:

- The correct assembly of plant genomes can be hampered by a large amount of repetitive sequences.

- Researchers have developed a bioinformatics tool for the automatic detection of repetitive genome regions, based on the identification of k-mers (nucleotide sequences of a pre-determined length).

- The tool was published under the name "Kmasker plants" in The Plant Journal.

- "Kmasker plants" is available as a web service or can be installed via https://github.com/tschmutzer/kmasker.

Advancing the application of genomic sequences through "Kmasker plants"

Gatersleben, 20.01.2020 **The development of next-generation-sequencing (NGS) has enabled researchers to investigate genomes that might previously have been considered too complex or too expensive. Nevertheless, the analysis of complex plant genomes, which often have an enormous amount of repetitive sequences, is still a challenge. Therefore, bioinformatics researchers from Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Martin Luther University Halle-Wittenberg (MLU) and Leibniz Institute of Plant Biochemistry (IPB) have now published "Kmasker plants", a program that allows the identification of repetitive sequences and thus facilitates the analysis of plant genomes.**

In bioinformatics, the term *k*-mer is used to describe a nucleotide sequence of a certain length "k". By defining and counting such sequences, researchers can quantify repetitive sequences in the genome they are studying and assign them to corresponding positions. As early as 2014, researchers at IPK in Gatersleben used this approach to develop the *in-silico (computer-based)* tool "Kmasker". It was used to detect repetitions in the characterisation of the barley genome (Schmutzer *et al.*, 2014).

The use of NGS is becoming more and more important, but the error-free composition of complex genomes from NGS results is still a challenge. For this reason, the researchers recently decided to revive and expand this initial proof-of-concept project. Under the leadership of Dr. Thomas Schmutzer, formerly from the research group "Bioinformatics and Information Technology" at IPK and now affiliated with the MLU, scientists from the MLU, the IPK, Wageningen University & Research and the IPB Halle worked in close cooperation on the redesign and development of "Kmasker plants". This collaboration was largely supported by the two service centres "GCBN" and "CiBi" from the German Network for Bioinformatics Infrastructure "de.NBI".

"Kmasker plants" allows for the rapid and reference-free screening of nucleotide sequences using genome-wide derived k-mers. In extension to the previous version, the bioinformatics tool now also enables comparative studies between different cultivars or closely related species, and supports the identification of sequences suitable as fluorescence *in situ* hybridisation (FISH) probes or CRISPR/Cas9-specific guide RNAs. Furthermore, "Kmasker plants" has been published with a web service that contains the pre-computed indices for selected economically important crop plants, such as barley or wheat. Dr. Schmutzer emphasises that "this tool will enable plant researchers all over the world to test plant genomes and thus, for example, identify repeat free parts of their sequence of interest." Rather, he believes that the enhanced features will make it possible to detect sequence candidate regions that have multiplied in the genome of one species but are missing in other species or occur in smaller copy numbers. This is a common effect that contributes to phenotypic variation of agronomic importance in various crops. A significant example is the Vrn-H2 gene, which is present in a single copy in winter barley, while it is missing in barley spring lines.

The "Kmasker plants" web-service is now available as part of the IPK Crop Analysis Tool Suite (CATS) and therefore as a service of the de.NBI Service Platform. Alternatively, the "Kmasker plants" source code can directly be accessed and installed via GitHub.

Figure for free use:
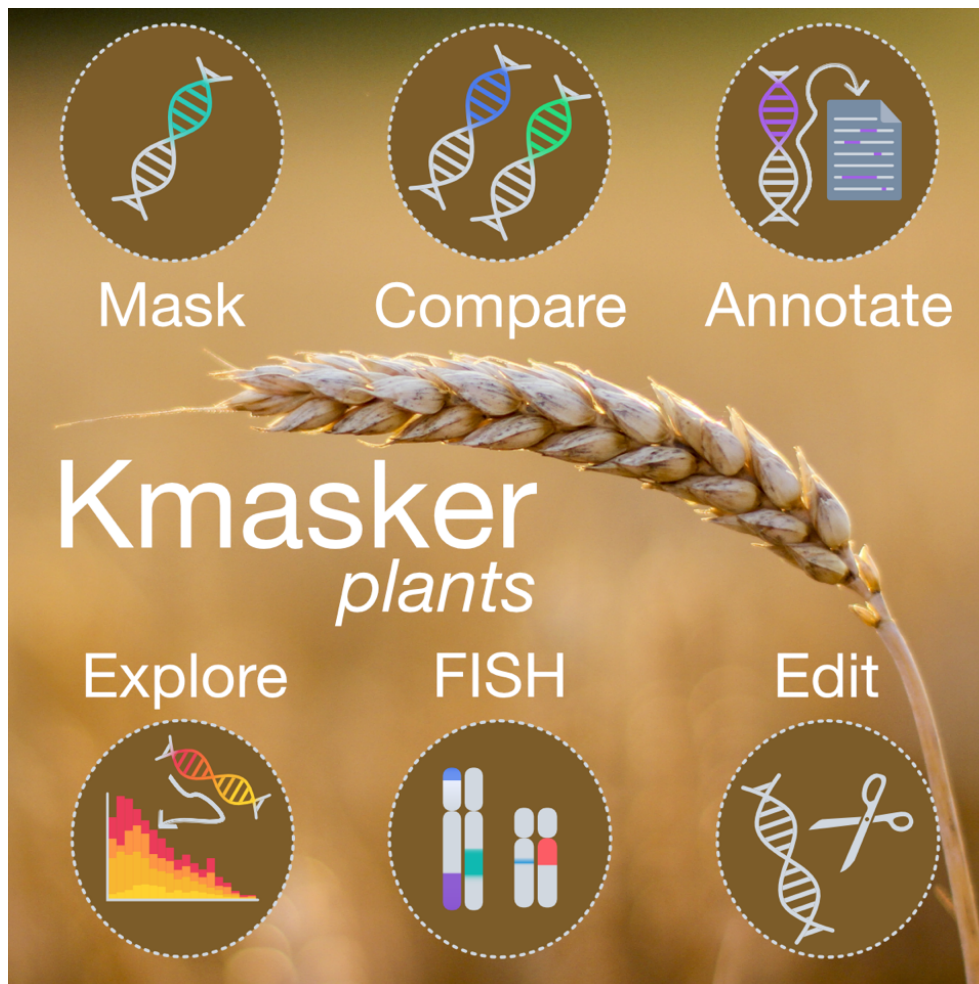
What short sequence pieces (k-mer) reveal. Applications and methods of the bioinformatics tool "Kmasker plants" for the analysis of sequence data. (Illustration: Chris Ulpinnis / IPB Halle & Pixabay).

Scientific Contacts:
Dr. Sebastian Beier
Leibniz Institute of Plant Genetics and Crop Plant Research (IPK)
Tel.: +49 39482 5659
E-mail: beiers@ipk-gatersleben.de

Dr. Thomas Schmutzer
Martin Luther University Halle-Wittenberg
Institute of agricultural and nutritional sciences
Tel.: +49 (0) 345 55 22 692
E-mail: thomas.schmutzer@landw.uni-halle.de

General Contact
Managing Office I Public Relations
Leibniz Institute of Plant Genetics and Crop Plant Research (IPK)
Tel. +49 39482 5424
E-Mail: geschaeftsstelle@ipk-gatersleben.de