

**Genomics, transcriptomic and micro RNA target side prediction of CRKs and their role under cotton leaf curl disease provide deep insight in cotton plant**

Athar Hussain<sup>1,2\*</sup>, Naila Asif<sup>1</sup>, Abdul Rafay Pirzada<sup>1</sup>, Azka Noreen<sup>2</sup>, Javeria Shaukat<sup>1</sup>, Akif Burhan<sup>1</sup>, Madiha Zaynab<sup>3</sup>, Ejaz Ali<sup>4</sup>, Koukab Imran<sup>1</sup>, Ayesha Amin<sup>5</sup>, Muhammed Arsalan<sup>2</sup>, Aquib Nazar<sup>1</sup>, M. Shahid Mukhtar<sup>6</sup>

<sup>1</sup>Genomics Lab, Department of Life Sciences, School of Science, University of Management and Technology (UMT), Lahore, 54000, Pakistan

<sup>2</sup>National Institute of Biotechnology and Genetic Engineering (NIBGE) constituent College of Pakistan Institute of Engineering and Applied Sciences (PIEAS), Faisalabad, 38000, Pakistan

<sup>3</sup>Shenzhen Key Laboratory of Marine Bio resource & Eco-Environmental Sciences, College of Life Sciences and Oceanography, Shenzhen University, Shenzhen 51807, China

<sup>4</sup>Center of Excellence in Molecular Biology, University of Punjab, 54000, Lahore, Pakistan

<sup>5</sup>Office of Research Innovation and Commercialization, University of Management and Technology (UMT), Lahore, 54000, Pakistan

<sup>6</sup>Department of Biology, University of Alabama at Birmingham, 1300 University Blvd., Birmingham, AL, 35294, USA

\*Corresponding author: [Atharmutahari@gmail.com](mailto:Atharmutahari@gmail.com)

**Abstract**

Cysteine-rich receptor-like-kinases (CRKs) are a transmembrane subfamily of receptor-like kinase, play important roles in plant adaptation. Cotton is the major source of fiber for textile industry. But environmental stresses are limiting its growth and production. Here, we have performed deep computational analysis of CRKs in five *Gossypium* species, including *G. arboreum* (60 genes), *G. raimondii* (74 genes), *G. herbaceum* (65 genes), *G. hirsutum* (118 genes) and *G. barbadense* (120 genes). All identified CRKs were classified into 11 major classes and 43 subclasses with finding of several novel CRK-associated domains including *ALMT*, *FUSC\_2*, *Cript*, *FYVE* and *Pkinase*. Of these, *DUF26\_DUF26\_Pkinase\_Tyr* was common and had high expression under different biotic and abiotic stresses. The 35 land plants comparison identified several new CRKs domain-architectures. Many SNPs and InDels observed in CLCuD resistant *G. hirsutum*. The miRNA target side prediction and their expression profiling in different tissues predicted miR172 as major CRK regulating miR. The expression profiling of CRKs identified multiple clusters with co-expression under certain stresses. The expression analysis under CLCuD highlighted role of *GhCRK057*, *GhCRK059*, *GhCRK058* and *GhCRK081* in resistant accession. The results provided primary data for functional analysis as well as a reference study for other important crops.

**Keywords:** Genome-wide, CRKs, Cysteine-rich receptor like-kinases, expression profiling, miRNA, PPI, RT-qPCR, Docking, CLCuD

**Introduction**

The number of people globally affected by hunger has been rising since 2014 (FAO et al., 2020). Almost 690 million people—8.9% of the world population—are estimated to have been

undernourished in 2019. Without change, the world is on track to reach 840 million undernourished people by 2030 (FAO et al., 2020). The reasons for an increasingly undernourished population are manifold including unequal resource distribution, food waste, and crop loss arising from climate change (Barrera & Hertel, 2021; Hasegawa et al., 2016; Janssens et al., 2020). While an integrated approach is vital to successfully curb this alarming trend, climate-change-resilient crops are needed to counter more frequent and extreme weather events that affect mostly populations with already high rates of undernourishment (FAO et al., 2020). During recent decades, significant technological progress has been made in plant breeding to develop new cultivars including marker-assisted selection (MAS), which has been in use since the early 1990s (Dudley, 1993; Ribaut & Hoeslin, 1998). In MAS, genomic markers are identified *in silico* and used to select individuals for use in specialized breeding programs (Hillel et al., 1990; Tanksley & Nelson, 1996). Originally, trait-associated genomic markers targeted by these approaches were identified as quantitative trait loci (QTL) (Geldermann, 1975), which in their earliest iteration could contain many genes within the same locus (Beckmann & Soller, 1983; Westman & Kresovich, 1997). The ability to identify candidate genes was accelerated by the construction of reference assemblies representing the DNA sequence of an individual's genome. Reference genomes were used to identify single-nucleotide polymorphisms (SNPs), copy-number variations (CNVs), and insertion-deletions (InDels). These markers became the foundation for genome-wide association studies (GWAS) and genomic selection (GS), where diversity sequencing datasets are compared with reference genomes and the identified variations statistically associated with phenotypes (Crosa et al., 2017; Hayes & Goddard, 2010; Ozaki et al., 2002; Varshney et al., 2005, 2009). The identification of trait-associated alleles and genes has also been used to drive a revolution in green biotechnology. The first genetically modified crop, the tomato (*Solanum lycopersicum* L.) 'FlavrSavr', was released to U.S. markets in the 1990s (Kramer & Redenbaugh, 1994). Since then, commercialization and distribution of genetically modified and, now, gene-edited crops have the potential to increase yield and improve traits such as biotic and abiotic stresses tolerance (Liu et al., 2021; Shi et al., 2017; Singh et al., 2018; Singh et al., 2020; Varshney et al., 2011; Wang et al., 2014; Zeng et al., 2019). The approaches described above require the identification of candidate genes or sequences for modification; however, they usually rely on a single reference genome, which does not contain the full extent of genetic variation present in the species, especially in polyploid crops (Bayer et al., 2020; Golicz et al., 2016a). This limitation has led to the rise of pangenomes, which combine the genomic data derived from multiple accessions and cultivars to detail the full extent of sequence variation within a species, finding genes and alleles to accelerate crop breeding. The pangenome concept was Core Ideas. Pangenomes allow integration of many types of DNA variants in a single reference. Pangenome studies highlight the importance of structural variations affecting agronomical traits. Many structural variations occur in regulatory regions and affect gene expression. Adopting pangenome graphs will help understand coding and noncoding variation. first proposed by Tettelin et al. (2005) to describe a *Streptococcus agalactiae*, or group B *Streptococcus*, reference combining datasets derived from eight different bacterial isolates (Tettelin et al., 2005). The genes present in all individuals were defined as 'core' genes, while the variable fraction was termed 'dispensable' (also referred to as 'accessory' or 'variable'). Functional characterization of group B *Streptococcus* core genes highlighted their involvement in essential processes, while a significant portion of dispensable genes was found to be the cause of newly acquired traits such as antibiotic resistance (Tettelin et al., 2005). Later pangenomics studies confirmed the critical roles of several variable genes in the organism's adaptation to a certain environment, making them effectively indispensable to the specific strain or cultivar fitness (Scheben et al., 2016). Tettelin et

al. (2005) also noted that increasing the sample size would expand the pangenome size indefinitely, making it an 'open' pangenome as opposed to a 'closed' one, whereby after inclusion of a sufficient number of samples, the addition of further datasets would not result in the identification of novel sequences (Tettelin et al., 2005). Since then, the adoption of high-throughput and cost-effective DNA sequencing technologies has resulted in the proliferation of pangenomes, including major crop species such as rice (*Oryza sativa* L.), maize (*Zea mays* L.), soy-bean [*Glycine max* (L.) Merr.], and rapeseed (*Brassica napus* L.) (Hirsch et al., 2014; Li et al., 2014; Song et al., 2020; Wang et al., 2018; Yao et al., 2015). Compared with single-reference genomes, pangenomes enable more accurate identification and representation of complex DNA polymorphisms within a species including large insertions, deletions, inversions, duplications, translocations, presence-or-absence variations (PAVs,) and copy-number variations (CNVs). Structural variations (or structural variants, SVs), ranging from few base pairs to several megabase pairs are the result of a variety of mechanisms including transposable elements (TEs) insertion, recombination, and double-strand breaks repair (Saxena et al., 2014). Structural variations are one of the major determinants of phenotypic variability within a species almost 50% of them absent from the 'Darmor-bzh' reference (Hurgobin et al., 2018). Dispensable resistance genes include BnaA03g43460.1D2, a potential orthologue of a clubroot resistance gene in *Brassica rapa* L., CRa (Ueno et al., 2012). A later study by Dolatabadian et al. (2020) expanded on this discovery, identifying 753 variable resistance gene analogs (RGA) in the rapeseed pangenome, with 106 resistance gene analog candidates predicted to contribute to blackleg resistance, one of the major diseases affecting *Brassica* species (Dolatabadian et al., 2020; Howlett et al., 2001).