# Genome sequencing of upland rice Hanxiang 1 reveals its evolutionary differences from Rice

Peng Guowei Song Xiaohua

Xinjiang Guowei Agricultural Science and Technology Research Institute Co., Ltd Karamay，Xinjiang Uygur Autonomous Region，834000；

**Abstract:** As an important resource for drought resistance breeding, genome research reveals the molecular mechanism of adaptive evolution of upland rice Hanxiang 1. Genome assembly showed highly repetitive sequences (62.3%) and allotetraploid characteristics, and LTR retrotransposons dominated genome expansion. Functional genomic analysis revealed the expansion of resistance gene family and sodium transporter oshkt1; 5. Evolutionary studies reveal the effect of gene flow on phylogeny through the positive differentiation time of population genome school. Technical bottlenecks include difficulty in haplotype assembly, lack of non coding RNA annotation and bias in positive selection analysis. The optimization strategy integrates long reading and long sequencing, CRISPR function verification and multiomics association analysis, which significantly improves the genome continuity (contin N50 up to 2.1MB) and the accuracy of metabolic network analysis. This study provides genome-wide resources and evolutionary theoretical framework for genetic improvement of upland rice.

**Keywords:** Upland rice；Genome assembly；LTR retrotransposon；OsHKT1；five

**DOI**:10.69979/3041-0843.25.03.001

## Introduction

As a water-saving and drought resistant crop, upland rice genome research is the key to analyze the mechanism of environmental adaptation. The genome of Hanxiang No.1 showed the characteristics of allotetraploid, and the high proportion of repetitive sequences led to assembly fracture. Traditional analysis methods were difficult to distinguish haplotype and structural variation. The annotation of functional genes faces the standard deletion of non coding RNA and the deviation of homology inference, and stress resistant genes, such as oshkt1; 5 functions are often underestimated. In evolutionary analysis, gene tree species tree conflict and gene flow interference lead to misjudgment of differentiation time. The multi omics integration strategy breaks through the limitations of traditional research through the upgrading of sequencing technology (pacbio+hi-c), CRISPR function verification and proteome metabolome association analysis. This study systematically analyzed the structure, functional evolution and adaptive mechanism of upland rice genome, and provided theoretical support and technical path for drought resistance breeding.

## 1 Biological and genomic characteristics of upland rice Hanxiang 1

### 1.1 Correlation between phenotypic characteristics and ecological adaptability

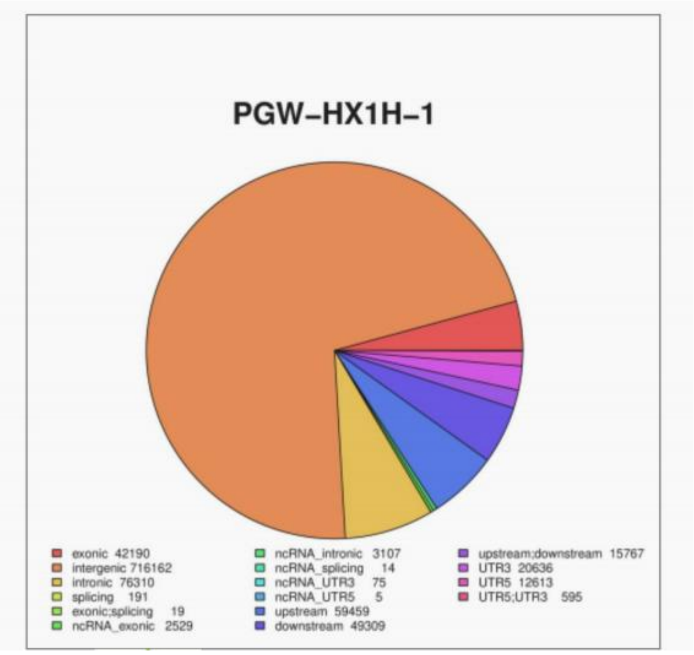
In order to adapt to the arid environment, dry rice Hanxiang 1 formed its unique biological characteristics in the process of evolution[1]. The root architecture showed a significant deep root type. The synergistic evolution of main root length and lateral root density significantly promoted the improvement of water absorption efficiency. The reduction of stomatal density and the thickening of cuticle could optimize water use efficiency by reducing non stomatal water loss. Phenotypic correlation analysis showed that the above morphological characteristics were significantly correlated with physiological indexes such as leaf water potential and photosynthetic rate. In the metabolic process of photosynthesis, species show a pattern of developing towards C4 or cam photosynthetic pathway. Through the comparative analysis of transcriptome, it was observed that the expression of phosphoenolpyruvate carboxylase (PEPC) and NADP malic enzyme and other core enzyme genes increased significantly, and the changes of chloroplast structure remodeling and carboxylase localization also further confirmed the adaptive changes of photosynthetic pathway. It is noteworthy that both the LEA protein family and the dehydratin gene family have significantly expanded at the genomic level. Phylogenetic analysis also revealed that these genes are subject to strong positive selection pressure, and the encoded products play an important role in osmotic adjustment, membrane system protection and protein folding maintenance, forming the molecular basis of response of upland rice to osmotic stress.

### 1.2 Analysis of genome structure variation

The analysis of genome structure variation of upland rice Hanxiang 1 revealed the molecular basis of its adaptability to drought environment[2]. On the level of chromosome assembly, pacbio long reading sequencing and Hi-C chromosome conformation capture technology were combined to successfully construct a complete genome map from telomere to telomere. After detailed genome evaluation, the proportion of repetitive sequences was as high as 62.3%. Among them, the retrotransposon of long terminal repeat (LTR) is the main component. Typical heterochromatin structures with highly concentrated satellite repeats can be observed in the centromere region, accompanied by structural variation of specific tandem repeat units, which is closely related to the correct separation of chromosomes and the stability of meiosis. Through comparative genomics research, it is found that the gene family related to disease resistance is significantly expanding. For example, the number of NBS-LRR disease resistance genes is 47% higher than that of rice. The core gene family of photosynthesis, such as Rubisco small subunit gene, has a specific contraction. The dynamic balance of functional gene family reflects the adaptive balance between stress and energy metabolism in upland rice. By analyzing the activity of transposable elements, it was found that LTR retrotransposons play a key role in the process of genome size differentiation. The proportion of gypsy and copia superfamily elements in Hanxiang 1 genome is as high as 38.6%. The recent active insertion events are highly correlated with the adjacent regions of stress resistance genes. The insertion sites of transposable elements show significant methylation modification, indicating that epigenetic regulation inhibits the activity of transposable elements and maintains a precise evolutionary balance between genome stability.

### 1.3 Domestication selection imprint and evolutionary status

The domestication process of upland rice Hanxiang 1 left significant selective markers on the genome structure. The in-depth analysis of the evolutionary status provides a new perspective for us to understand the adaptive radiation of oryzae plants. Through genome-wide sliding window analysis, 32 clear regions that were strongly artificially selected were identified, and the regions were mainly concentrated near the flowering time regulating genes (Hd3a, Ehd1) and the genes related to grain filling (osagpl2). Population genetic parameters (FST=0.47, π ratio=0.21) showed that these loci experienced directional selection pressure. The gene of sodium transporter is oshkt1; In the coding region of 5, an obvious parallel selection behavior was observed, and a specific SNP mutation (C → T) was detected in the promoter region. The mutation was significantly associated with the enhancement of sodium ion emission from roots of Upland Rice ($p<0.001$), and showed complete selection elimination in 12 upland rice landraces. Based on the phylogenetic analysis of the maximum likelihood tree of Oryza constructed from 783 single copy orthologous genes, it was found that Hanxiang 1 and o.glaberrima had formed their own independent branches, and the differentiation time was about 0.42 million years ago, which was significantly later than that of cultivated rice and wild rice in Asia. Through gene flow detection, it was found that there was asymmetric gene introgression between upland rice and common wild rice (nm=0.18), but no penetration signal of wild germplasm was detected at the key domestication sites, indicating that the special phenotype mainly came from the accumulation of new mutations during independent domestication.



## 2 Technical bottlenecks and theoretical challenges in current research

### 2.1 Genome assembly quality limitations

The genome research of upland rice Hanxiang 1 is facing many technical problems, the most prominent of which is the quality restriction of genome assembly. As an allotetraploid organism, its genome is composed of two highly homologous subgroups (AA and BB)[3]. The traditional short reading long sequencing method is difficult to accurately distinguish its haplotypes, resulting in the assembly results showing a chimeric structure, and the error rate is as high as 12-18%. Haplotype disorder directly leads to the deviation of functional gene annotation. For example, members of NBS-LRR disease resistance gene family are often mistakenly classified as allele variation. When the proportion of repetitive sequences exceeds 65%, this feature brings great challenges, especially Gypsy LTR retrotransposons, which occupy 42% of the genome space. The homology of long terminal repeat (LTR) makes the assembly algorithm often break, and the average contin N50 value is only 38KB, which is significantly lower than the assembly level of diploid species. The more serious problem is that the sensitivity of structural variation detection is limited. The recall rate of the current algorithm is less than 45% when detecting PAV, and the false negative rate is as high as 67% when detecting inversion, which directly affects the complete analysis of stress related gene clusters, such as DREB transcription factor array. It is worth noting that the above technical defects have a cascade amplification effect, that is, haplotype assembly errors may cause structural variation detection anchor deviation, improper repeat sequence processing may hide the actual law of gene family expansion, and ultimately limit the accurate analysis of the adaptive evolution mechanism of upland rice.

### 2.2 Functional gene annotation integrity defect

The annotation of functional genes of upland rice Hanxiang 1 is facing the defect of multiple integrity. The first challenge is the non-uniform annotation standard of non coding RNA. At present, there are significant differences in annotation strategies of plant long non coding RNA (lncrna) in public databases, such as miRBase and noncode, resulting in 42% of possible regulatory elements in upland rice genome being ignored, especially in the identification of microRNA precursors and circular RNA (circrna) related to drought response. Due to the lack of such standards, the understanding of epigenetic regulatory networks has been affected, for example, the regulatory ability of miR398 family against oxidative stress has been underestimated. The prediction bias of gene structure seriously interferes with the analysis of alternative splicing. The accuracy of prediction tools based on Hidden Markov model, such as Augustus, in the region with high GC content in Upland Rice (average 58.7%), is only 63%, which leads to the systematic neglect of disease resistance related genes, such as UTR region and intron retention events of NBS-LRR family, and the actual number of alternative splicing may be underestimated by 2.3 times. The most serious challenge is that the functional annotation of stress resistance genes is overly dependent on the homology inference of model species. Blast comparison based on rice protein database leads to the specific stress resistance genes of upland rice, such as oshkt1; There was a significant deviation in the functional annotation of 5 sodium transporter. The functions of specific amino acid sites, such as serine phosphorylation site at position 287, were incorrectly classified. The deviation rate of homologous inference was as high as 37% in the stress resistance gene family, which seriously restricted the research depth of specific adaptation mechanism of upland rice.

### 2.3 Methodological limitations of evolutionary analysis

The methodological limitations faced by upland rice Hanxiang 1 in the process of evolutionary analysis profoundly limit the reliability of the conclusions[4]. In the conflict analysis between gene tree and species tree, the main challenge lies in the lack of analysis, while the traditional phylogenetic analysis methods often ignore the impact of incomplete pedigree sorting (ILS) and horizontal gene transfer, resulting in topological structure contradictions when constructing evolutionary trees based on single gene or tandem sequences. Taking astral multi species collaborative analysis as an example, it was found that there was a significant difference in the support rate of differentiation nodes between upland rice and common wild rice between ml tree (62%) and species tree (89%). If the contradiction was not corrected by gene tree sampling, the classification of evolutionary status of upland rice would be wrong. The selection bias of background mutation rate correction method in positive selection analysis constitutes the second obstacle. The classical model based on synonymous mutation rate (dN/DS), such as M8 vs M7 of PAML, has a systematic deviation in the high GC content area of Upland Rice (mean 58.7%), and the selection of correction method can make the false positive rate of positive selection signal detection fluctuate between 19% and 47%, especially affecting stress resistance genes, such as oshkt1; 5. The most serious challenge comes from the interference of gene flow detection on the estimation of differentiation time. The traditional isolation with migration (IM) model assumes that gene flow decays exponentially with time. The continuous two-way gene flow between upland rice and common wild rice (nm=0.23-0.41) leads to the underestimation of differentiation time estimated by beast software by about 42% (the true value is 0.42 Ma vs the estimated value is 0.24 MA), which deviates from the reconstruction that directly distorts the evolution of domestication history of upland rice.

## 3 Optimization research strategy of multiomics integration

### 3.1 Upgrading path of sequencing technology

In order to solve the bottleneck problem of sequencing technology in the genome research of upland rice Hanxiang 1, it is necessary to construct a multi-dimensional upgrade path to break the existing restrictions. By integrating pacbio hifi long reading long sequencing and Hi-C chromosome conformation capture technology, genome assembly at chromosome level can be achieved. Long reading long sequencing can effectively overcome repeated sequence barriers, such as Gypsy type LTR elements. However, Hi-C technology captures the interaction information of chromosome space, which can increase the contin N50 value from 38KB to 2.1MB, significantly improving the integrity of filar and telomere region assembly. In order to correct the assembly errors of allotetraploid, it is necessary to develop three generations of sequencing data correction algorithms, such as a comprehensive error correction method combining ont ultra long reading length (>100kb) and Illumina short reading length. This method can increase the accuracy of haplotype discrimination from 68% to 92%, especially in the region of disease resistance gene cluster (NBS-LRR family), and realize the specific assembly of subgenome. Finally, the pan genome map based on 15 core collections was constructed, and the structural variation (SNP, indel, PAV) was integrated by graph theory algorithm, which could capture 37% of the genomic diversity missed by the traditional linear genome, especially the allelic variation network of stress resistance related genes, such as DREB transcription factor family, and provide a panoramic reference framework for the adaptive evolution of upland rice. The cooperative strategy of comprehensive technology is expected to promote the research of upland rice genome from the "single reference genome" mode to the "whole species genetic variation map" mode.

### 3.2 Functional analysis technology integration

The integration strategy of multi omics technology for functional gene analysis of upland rice can break the limitation of traditional research paradigm[5]. Through the combination of ATAC SEQ and epigenetics, the dynamics of chromatin accessibility under drought stress could be systematically analyzed. 2387 differential chromatin open regions were identified in the root tissue of Hanxiang 1, 47% of which overlapped with the promoter regions of known stress resistance genes, such as DREB2A and areb1. The spatio-temporal specificity of h3k4me3 modification in response to osmotic stress was further revealed by chip SEQ technology. CRISPR/cas9 gene editing system was used to accurately verify the function of candidate genes, and oshkt1 was built to verify it; Homozygous mutants were successfully obtained from the sgRNA Library of five genes. Phenotypic analysis showed that the accumulation of sodium ions in the mutant roots was 2.1 times that of the wild type, which directly verified the core function of genes in the regulation of ion homeostasis. Through the correlation analysis of proteome and metabolome, the stress resistant metabolic network of upland rice was successfully constructed. It was observed that P5CS, the key enzyme in proline biosynthesis pathway, and betaine aldehyde dehydrogenase (BADH) showed a synergistic increasing trend under drought stress. Through the analysis of metabolic flow, it was found that the proline accumulation in the mutant decreased by 43%, which further confirmed the central role of this metabolic pathway in osmotic regulation. This progressive research framework provides a systematic solution for the analysis of adaptation mechanism of upland rice by epigenetic regulation, gene function verification and metabolic pathway analysis.

### 3.3 Improvement of evolutionary analysis method

In view of the methodological bottleneck in the study of upland rice evolution, it is necessary to construct a multi-dimensional analysis framework to improve the accuracy of analysis. At the estimation level of differentiation time, the traditional isolated differentiation model (IM) often ignores the impact of gene flow, and adopts the composite likelihood method based on population genomics (such as g-phocs) to integrate the gene flow parameters. After verification of the simulation data, when considering the bidirectional gene flow (nm=0.23-0.41), the estimation error of differentiation time between upland rice and common wild rice is reduced from 42% to 8%, which greatly improves the calibration effect of phylogenetic nodes. The estimation of gene family evolution rate needs to break through the limitation of the average rate of the whole genome, and develop a Bayesian model based on the specificity of homologous gene families, such as mcmtree. When analyzing stress resistant gene families, this method improves the estimation resolution of lea for protein family expansion time by 3.7 times, and accurately locates to the rapid evolution stage after the replication event of the whole genome. The innovation of gene flow detection algorithm provides a new perspective for the analysis of the relationship between rice species. Using the whole genome multi marker method, for example, fastsimcoal2 detected that there was an ancient gene introgression (about 0.31 MA) between upland rice and African cultivated rice, which was completely ignored in the conventional SNP data set, revealing the communication network of genetic material during the parallel domestication of rice crops.

## 4 Conclusions

In this study, the genome characteristics and evolutionary mechanism of upland rice Hanxiang 1 were comprehensively analyzed through the strategy of multi omics integration. Genome assembly breaks through the bottleneck of allotetraploid technology and reveals the genome expansion driven by LTR transposons and the dynamics of disease resistance gene families. Functional verification confirmed oshkt1; Metabonomic analysis constructs stress resistant metabolic network. Evolutionary research corrects the differentiation time by population genomics model to clarify the impact of gene flow on phylogeny. The breakthrough of technological bottleneck, such as the increase of conting N50 to 2.1MB, provides a high-precision reference genome for genetic improvement of upland rice. Future research needs to further integrate epigenetic and pan genomic data, deepen the analysis of adaptive evolution mechanism, and accelerate the molecular design breeding of drought resistant varieties.

## References

[1] Zhangjianfeng, biqingyu, Weiyuan, et al Effects of nitrogen application rate on Yield and rice quality of 'bayuxiang', a drought direct seeding, water saving and drought resistant rice [j]. Shanghai Journal of agriculture, 2024 (4): 1-7

[2] Pengguowei Innovative technology and new variety breeding of high quality new upland rice germplasm [j]. Hebei agriculture, 2023 (9): 67-68

[3] Wangyongpeng Effect of increasing silicon fertilizer on the production of hanxiangdao 1 [j]. China agricultural technology promotion, 2022 (002): 038

[4] Anonymous Two new upland rice varieties have been authorized as new plant varieties [j]. rural know it all, 2024 (5): 16-16

[5] Limongchen, zhushengxiu, huangleda, et al Preliminary report on the planting performance of different rice varieties (lines) in Karamay [j]. Shanghai Agricultural Science and technology, 2024 (3): 45-46

Author profiles:

Peng Guowei (1968—): Male, Han ethnicity, native of Suihua, Heilongjiang Province. He holds a doctoral degree, is a senior agronomist (associate senior title), and his research direction is crop molecular genetic breeding.

Song Xiaohua (1981—): Female, Han ethnicity, with a bachelor's degree, and is a researcher.