



Crowning the King of Fruits: Genomic Evidence of Population Structure and Relatedness in Malaysian Durians

Hung-Hui Chung¹ · Li Yuan Liew² · Adriana Batrisyia Mohd Faisal² · Leonard Whye Kit Lim¹ · Han Ming Gan²

Received: 11 November 2025 / Accepted: 28 June 2026

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2026

Abstract

Durian (*Durio zibethinus*), often dubbed the King of Fruits, is a high-value tropical crop native to Southeast Asia and widely cultivated in Malaysia. Revered for its unique flavor yet polarizing for its pungent aroma, durian holds significant cultural and economic importance across the region. Despite this, genomic research on Malaysian durian cultivars remains limited. We performed whole-genome sequencing of 14 durian samples consisting of several renowned and local durian cultivars, including Musang King (D197), Sultan (D24), Black Thorn (D200), Muar Gold (D101), and Red Prawn (D13) collected from distinct geographic regions across Malaysia. Illumina sequencing yielded approximately 20 to 30x coverage per sample, sufficient for single-nucleotide polymorphism (SNP) detection and population genomic analyses. Principal component analysis (PCA) based on genome-wide SNPs revealed visible population stratification and varying degrees of genetic relatedness among Malaysian durians. Subsequent KING kinship analysis indicated that most cultivars are relatively clonal, with the notable exception of the reference Musang King genome, which appeared genetically heterogeneous compared to other Musang King samples, a finding further supported by population structure inference from admixture analysis and ITS1-5.8 S-ITS2 haplotyping. Notably, the Black Thorn cultivar showed no relatedness to other Malaysian varieties but exhibited a distant genetic similarity to the Kan Yao variety of Central Thailand's Nonthaburi region. These results provide new genomic insights into the population structure, cultivar differentiation, and potential hybridization among Malaysian durians, forming a valuable foundation for future molecular marker development and breeding programs aimed at different variety authentication and genetic improvement.

Keywords Durian · *Durio* · Whole genome sequencing · Illumina · Population genomics

Introduction

Durian (*Durio zibethinus*) is a highly popular seasonal fruit in Malaysia, with strong domestic demand and growing export value, particularly to China (Fig. 1). As a key contributor to Malaysia's agricultural economy, durian holds both economic and cultural importance (Thorogood

et al. 2022). Decades of selective breeding have produced numerous cultivars that differ in taste, texture, aroma, and bitterness, reflecting regional adaptation and consumer preference (Khoo 2025). Despite Malaysia's central role in durian domestication and cultivation, the first reference durian genome was published by Teh et al. (2017), using a specimen originally sourced from Pahang, Malaysia. The study offered insights into the molecular basis of durian's distinctive aroma and established a well-annotated reference for future resequencing and comparative analyses.

Subsequent studies have expanded the genomic understanding of durian where Lin et al. (2022) performed ddRAD-seq analyses on durians cultivated in southern China, tracing their genetic origins to Southeast Asia. More recently, Zhong et al. (2025) performed whole-genome resequencing of over 100 durians planted in China and

✉ Han Ming Gan
gan.mypb@gmail.com

¹ Faculty of Resource Science and Technology, University Malaysia Sarawak, Kota Samarahan 94300, Sarawak, Malaysia

² Patriot Biotech Sdn Bhd, Bandar Sunway, Petaling Jaya 47500, Selangor, Malaysia



Fig. 1 Durian fruit (kampung variety) from Sarawak, Malaysia, displayed for sale. The fruit shows the characteristic thorny husk and is fully ripe, ready to be opened for consumption

identified three genetic clusters, suggesting considerable genomic diversity among cultivated materials. However, because the samples were all introduced and cultivated in China, with limited representation of native Southeast Asian germplasm, the study offers only a partial view of durian's natural diversity. The genetic clustering observed may largely reflect recent propagation and planting histories rather than deep biologically meaningful evolutionary differentiation. The weak correspondence between genetic clusters and geographic provenance, together with signs of reduced nucleotide diversity in one group, further suggests founder effects and a restricted genetic base among the Chinese-planted durians (Zhong et al. 2025). On the contrary, in Thailand, three Thai durian genomes were sequenced, representing the Kradumthong, Monthong, and Puangmanee cultivars and similarly revealed substantial intraspecific variation among durian cultivars (Nawae et al. 2023). Building on this, another effort reported a chromosome-scale, haplotype-resolved assembly of the Thai cultivar Kan

Yao, currently the most complete durian genome available (Ji et al. 2025).

Within Malaysia, however, population-level genomic research remains limited. The only relatively comprehensive durian genetic study to date used microsatellite markers to assess durian diversity across regions, revealing considerable genetic heterogeneity. Yet, the small number and modest informativeness of these loci restricted the resolution needed to distinguish among all cultivars (Siew et al. 2018). Beyond the Musang King reference genome, no other Malaysian durian genomes have been sequenced and/or reported, and none by a Malaysian-led research team. Given the deep cultural symbolism and economic significance of durian to Malaysia, this represents an opportunity for expanded Malaysian-led contributions in the genomic characterization of one of its most iconic crops.

In this study, we address this by generating whole-genome sequences for 14 durian samples representing major cultivars collected across Malaysia, including one from Sarawak, to capture the breadth of the country's genetic and geographical diversity. Using Illumina whole-genome sequencing, we present the first population-scale genomic analysis of Malaysian durians, revealing patterns of genetic diversity and population stratification among key cultivars.

Materials and Methods

Sample Collection and DNA Extraction for Whole Genome Sequencing

A total of 14 durian samples were collected across Malaysia for DNA extraction and whole-genome sequencing (Table 1). Fruit flesh was obtained from ripe, ready-to-eat fruits purchased directly from local sellers and sampled immediately upon opening on the same day of purchase. Flesh tissue was placed into sterile 1.5 mL tubes and stored at -20 °C until DNA extraction. Young leaf tissues were collected from actively growing shoots and preserved in silica gel prior to DNA extraction. Approximately 150 mg of flesh or leaf tissue was homogenized and washed in sorbitol buffer (Inglis et al. 2018) to minimize polysaccharide and polyphenol carryover. Plant cell lysis was performed in a buffer containing 1% SDS, 25 mM Tris-HCl (pH 8.0), 10 mM EDTA, 15 mM NaCl and 2% PVP followed by incubation at 65 °C for 30 min. Subsequently, 0.5x volume of saturated potassium chloride was added to precipitate residual mucopolysaccharides (Sokolov 2000). DNA was precipitated from the supernatant by adding 0.6x volume of isopropanol, centrifuged at maximum speed for 10 min, and the resulting DNA pellet was washed twice with 70% ethanol and resuspended in TE buffer.

Table 1 Summary of sequencing metrics, genome depth and sampling information sorted by percentage of chloroplast-aligned reads

Sample ID	Cultivar	Mapped reads	Nuclear Genome Depth (% genome covered $\geq 5\times$)	Chloroplast Genome Depth (x)	Chloroplast Reads (%)	Tissue Type	Preservation Method	Location	Reference
KGSwk	Kampung	132,369,811	26 (89.5%)	31,579	22	Flesh	Ethanol	Kuching, Sarawak	This study
MKSD	Musang King	176,380,995	35 (94.4%)	36,046	19	Flesh	Frozen (-20 C)	Sungai Dua, Pahang	This study
D101MT1	D101	145,065,423	29 (92.6%)	24,548	16	Flesh	Frozen (-20 C)	Mantin, Negeri Sembilan	This study
D24SD	D24	204,662,301	41 (94.0%)	34,177	15	Flesh	Frozen (-20 C)	Sungai Dua, Pahang	This study
D101SD	D101	162,594,600	33 (93.4%)	20,460	11	Flesh	Frozen (-20 C)	Sungai Dua, Pahang	This study
KGSD	Kampung	186,865,730	37 (94.7%)	16,082	8	Flesh	Frozen (-20 C)	Sungai Dua, Pahang	This study
MKJ7	Musang King	135,854,529	27 (94.8%)	10,374	7	Leaf	Silica Gel	Johor	This study
MKRef	Musang King	127,518,986	26 (95.5%)	6,790	5	Leaf	Unknown	Raub, Pahang	(Teh et al. 2017)
D101J2	D101	154,637,389	31 (94.1%)	5,605	3	Leaf	Silica Gel	Johor	This study
D13J3	D101	184,288,643	37 (94.4%)	5,876	3	Leaf	Silica Gel	Johor	This study
D13J4	D13	147,580,309	30 (94.3%)	4,701	3	Leaf	Silica Gel	Johor	This study
MKJ8	Musang King	174,925,861	35 (95.8%)	6,117	3	Leaf	Silica Gel	Johor	This study
BTJ6	Black Thorn	178,192,152	36 (94.7%)	4,296	2	Leaf	Silica Gel	Johor	This study
BTP	Black Thorn	163,339,770	33 (94.5%)	4,246	2	Leaf	Silica Gel	Nibong Tebal, Pulau Pinang	This study
D101J1	D101	162,248,528	32 (94.4%)	3,739	2	Leaf	Silica Gel	Johor	This study
KanYao	KanYao	182,566,142	37 (94.0%)	4,454	2	Leaf	Unknown	Thailand	(Ji et al. 2025)

Illumina Whole Genome Sequencing

Approximately 100 ng of high-quality genomic DNA was sheared to an average fragment size of 350 bp using a Bioruptor sonicator (Diagenode, Belgium). Library preparation was performed using the NEBNext Ultra II DNA Library Preparation Kit (New England Biolabs, USA) following the manufacturer's protocol. Briefly, fragmented DNA was end-repaired, followed by the addition of an adenine overhang and ligation of Illumina sequencing adapters. Indexed PCR amplification was then carried out to attach sample-specific barcodes and complete the Illumina adapter structure. The libraries were sequenced on an Illumina NovaSeq 6000 platform (Illumina, USA) using a paired-end 2×150 bp configuration, generating approximately 20 Gb of raw data per sample.

Reference Mapping and Variant Calling

The reference durian nuclear genome of *Durio zibethinus* and the chloroplast genome from the *Durio zibethinus* Mon Thong variety (GenBank accession MT321069.1) (Shearman et al. 2020; Teh et al. 2017) were combined into a single reference assembly for read alignment. Raw Illumina reads were adapter- and quality-trimmed using fastp version 0.20.1 (Chen et al. 2018), retaining high-quality reads longer than 50 bp. Filtered reads were then aligned to the combined reference using BWA-MEM2 version 2.2.1 (Li 2013). The resulting SAM files were sorted and converted

to BAM format using SAMtools version 1.13 (Danecek et al. 2021). Genome-wide coverage breadth was estimated for each sample using mosdepth v0.3.14 (Pedersen and Quinlan 2018) with a $5\times$ depth threshold to quantify the proportion of the reference genome covered by at least five reads. Variant calling was performed jointly on all aligned BAM files using bcftools v1.22 (Danecek et al. 2021). We chose bcftools for variant calling due to its efficiency and robustness for joint multi-sample SNP calling in population-scale resequencing datasets, with benchmarking studies reporting comparable performance to GATK and FreeBayes in non-human genomes under standard filtering conditions (Lefouili and Nam 2022; Liu et al. 2022). Briefly, variants were first generated with the mpileup command to produce uncompressed binary variant call output. The resulting variants were subsequently filtered to retain only high-confidence (Q30), biallelic single-nucleotide polymorphisms (SNPs) with a minor allele frequency (MAF) ≥ 0.05 and missing data fraction ≤ 0.2 .

Population Structure and Genetic Relatedness Analyses

Population structure among durian cultivars was first examined using principal component analysis (PCA) based on genome-wide single nucleotide polymorphisms (SNPs). Prior to PCA, linkage disequilibrium (LD) pruning was performed on the quality-filtered SNP set using the snp-gdsLDpruning function in SNPRelate v1.34.1, applying a

sliding-window approach (up to 50 SNPs per window and 1 Mb physical distance) with an LD threshold of 0.2 (composite method). The LD-Pruned SNP set was converted from VCF to GDS format, and PCA was performed using the `snpGdsPCA` function from the `SNPRelate` v1.34.1 R package (Zheng et al. 2012). An admixture analysis was also conducted on the LD-Pruned SNP set using `ADMIXTURE` v1.3 (Alexander and Lange 2011). The program was run across a range of clusters from $K=2$ to $K=8$ and the resulting ancestry proportions were visualized in a structure plot-style representation to reveal the genetic composition and possible admixture among cultivars. To further refine these relationships, kinship coefficients were subsequently computed using the KING-robust algorithm implemented in `PLINK` v2.0 (Chang et al. 2015; Manichaikul et al. 2010). Unlike the `ADMIXTURE` and PCA frameworks which require linkage-pruned markers, the KING algorithm operates optimally on unpruned genome-wide variants to accurately separate background population stratification from recent family relatedness. Therefore, following official software guidelines, the initial set of quality-filtered SNPs was used for the KING kinship inference.

ITS1-5.8S-ITS2 Haplotype Reconstruction

Conventional variant calling approaches are unsuitable for the ITS1-5.8S-ITS2 region because, although the genome itself is diploid, this locus exists in multiple paralogous copies that vary within the same genome. As a result, standard diploid-based variant callers fail to capture the full haplotype diversity, and *de novo* assembly approaches often collapse these repetitive sequences into consensus contigs. To better resolve this complex region, we adopted a haplotype reconstruction strategy inspired by viral quasispecies analysis, where the underlying computational challenge similarly involves disentangling closely related sequence variants from a mixed read population.

First, a subsample of filtered pair-end reads (1 million reads) from sample BTP was assembled using `MEGAHIT` v1.2.9 (Li et al. 2015). Contigs containing the ITS region were identified using `ITSx` v1.1b (Bengtsson-Palme et al. 2013), which also extracted the full ITS1-5.8S-ITS2 region. Subsequently, 1 million filtered paired-end reads from each sample were then aligned to this reference sequence using `BWA-MEM`. Haplotype reconstruction was performed using `CliqueSNV` v2.0.3 (Knyazev et al. 2021), which infers local haplotypes and estimates their frequency based on patterns of co-occurring variants within the aligned reads. Finally, all reconstructed haplotypes from all samples were concatenated and re-clustered at 100% sequence identity and identical sequence length using `CD-HIT` v4.8.1 (Fu et al. 2012), producing a non-redundant haplotype set for downstream analyses.

Fig. 2 2D PCA plots showing different pairwise combinations of the principal components PC1-PC4 derived from linkage disequilibrium (LD)-pruned genome-wide SNP dataset. Each point represents an individual sample and is colored according to its durian cultivar. Clustering in each plot reflects variation along the specific PCs shown and does not necessarily represent absolute genetic relationships among samples. Percentage values on the axes indicate the proportion of total genetic variance explained by each principal component

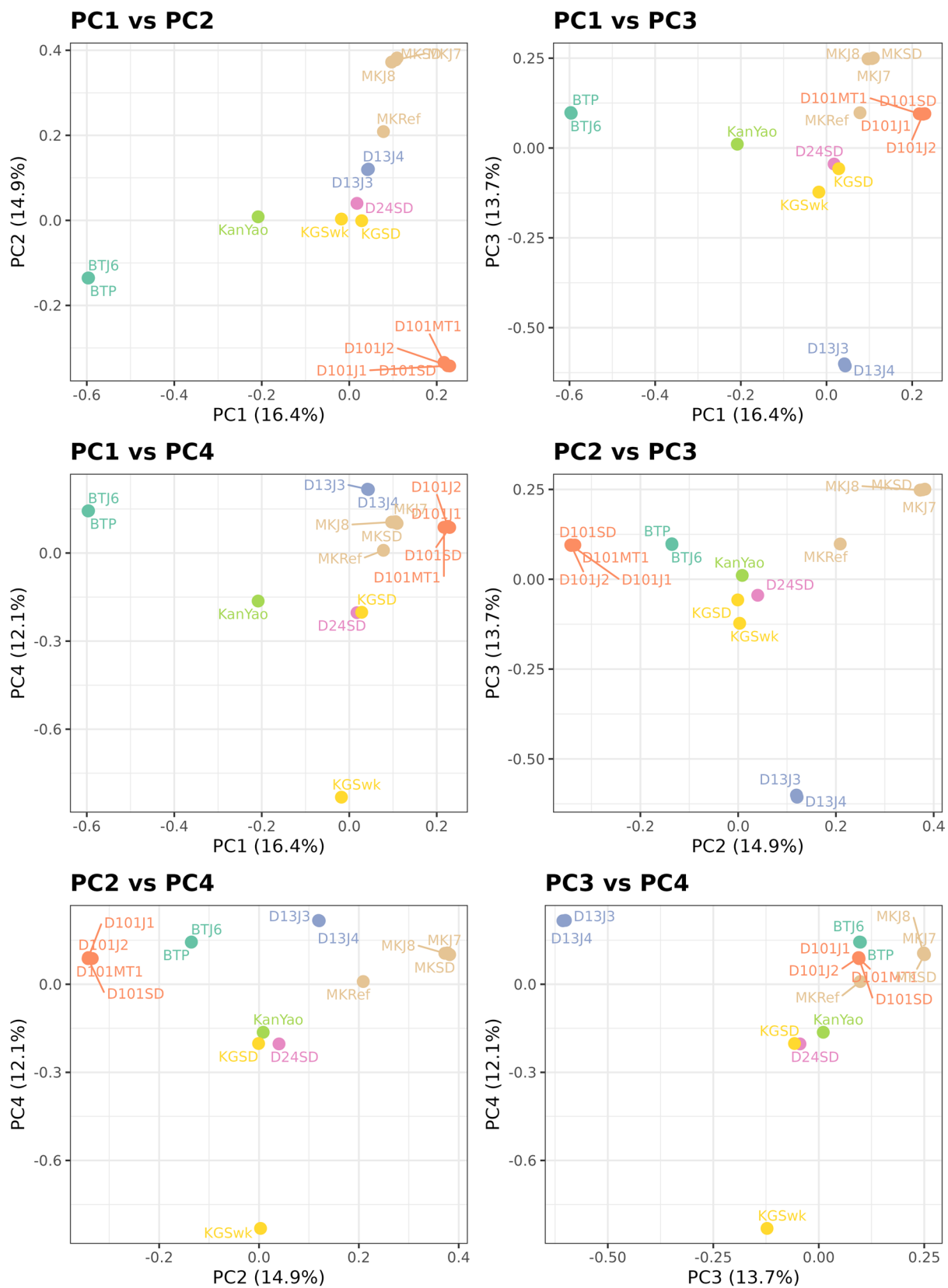
Results

Durian Flesh Yields Usable NGS Data and Higher Chloroplast Read Counts

On average, 191 million (range: 166–222 million) raw paired-end reads were generated per sample, including those derived from durian flesh, which is typically more challenging due to the presence of polysaccharides and other PCR inhibitors. Interestingly, a higher proportion of reads from flesh samples aligned to the chloroplast genome compared to leaf samples, despite leaves being among the most chloroplast-rich tissues. Mapping analysis showed that 8–22% of reads from flesh samples aligned to the durian chloroplast genome, compared to only 2–7% from leaf samples (Table 1). Across all samples, nuclear genome sequencing depth ranged from $26\times$ to $41\times$, with consistently high genome representation as reflected by $\geq 5\times$ breadth coverage of approximately 89.5–95.8% (Table 1), indicating robust and uniform recovery of the nuclear genome despite differences in tissue type and DNA source.

PCA Reveals Population Stratification among Malaysian Durian Cultivars

An initial set of 25.24 million high-quality SNPs was retained following quality filtering. After LD-pruning, 1.39 million representative SNPs (5.5%) were retained for downstream analyses, representing a reduction of 94.5% (23.85 million variants) due to linkage disequilibrium. Principal component analysis (PCA) of LD-pruned SNP data revealed that the genetic structure among cultivars can be resolved across pairwise projections of PC1 to PC4 (Fig. 2). PC1 effectively separates BlackThorn (BTJ6, BTP) from all other cultivars on the negative end of the axis, while separating D101 and Musang King from the rest on the positive end. In both PC2 and PC3, the reference MK sample (MKRef) was positioned slightly further back toward the center of the MK cluster. In addition, PC3 primarily resolved D13 (Udang Merah) on the extreme negative end of the axis, marking it as a distinct genetic lineage within the dataset. Interestingly, while both KGSD (Kampung from Selangor) and KGSwk (Kampung from Sarawak, Borneo) are grouped as “Kampung” types, their



wide geographic separation is reflected genomically. PC4 clearly distinguishes the Sarawak sample from the Peninsular Malaysia one. However, the close clustering of D24 with KGSD across all principal components suggests a highly similar genetic profile, which is unexpected given their distinct cultivar identities.

The Reference Musang King Genome Exhibits Unexpected Genetic Heterogeneity Compared to other Musang King Samples

Admixture analysis based on LD-pruned genome-wide SNPs was used to explore the genetic structure and potential ancestry components among the *Durio zibethinus* samples (Fig. 3). Cross-validation (CV) error analysis indicated that a three-cluster model ($K=3$, $CV=0.50730$) represents the mathematically optimal population structure for this dataset. At this resolution, the samples were partitioned into three broad ancestral lineages. The first cluster grouped the Black Thorn (BT), Kampung (KG), and D13 samples together, the second cluster cleanly isolated the D101 variants and the third cluster defined the Musang King (MK) field samples. Notable admixed profiles were observed for the single D24SD individual and the Kan Yao variety, both of which exhibited primary affinity to the Black Thorn/Kampung/D13 cluster while carrying distinct secondary ancestral components from the other lineages. Interestingly, at $K=5$, the majority of the analyzed field samples exhibited high assignment probabilities to their respective single ancestry clusters despite higher CV error. For example, the sequenced Musang King field samples (MKJ7, MKJ8, and MKSD) shared an almost identical genetic background despite their disparate geographic collection sites. Similarly, the sampled individuals of Black Thorn (BTJ6, BTP), D13 (D13J3, D13J4), and D101 displayed distinct, cultivar-associated profiles with minimal inferred admixture within the limits of this dataset. Conversely, two notable exceptions to this uniform clustering were observed. First, the single D24SD individual clustered entirely within the Kampung (KGSD, KGSwk) background. In contrast to the uniform profiles of our MK field samples, the reference Musang King genome (MKRef) displayed an admixed profile at $K=5$, comprising approximately 70 to 75% core Musang King ancestry alongside minor signals sharing components with the Black Thorn and Kampung backgrounds. Given our limited sample size, further extensive population sampling is required to determine whether this admixture profile is widespread among alternative Musang King lineages or specific to the reference clone assembly.

KING Kinship Reveals Clonal Structure and Lineage-level Differentiation among Durian Cultivars

Pairwise KING kinship coefficients (ϕ) were inferred from genome-wide SNP data (24.8–25.1 million SNPs) across 16 durian samples to robustly estimate recent co-ancestry while accounting for allele frequency variation. A kinship network was constructed with edges retained only for $\phi > 0.04$, a moderately conservative threshold designed to capture both clonal and close pedigree relationships while excluding background relatedness. Edge thickness and color intensity were scaled by IBS0, the proportion of loci with zero alleles identical by state, such that lower IBS0 values denote greater genomic similarity (Fig. 4). The resulting network revealed a central hub anchored by KGSD (a wild-type variety), connected to multiple samples. Musang King derivatives (MKJ7, MKJ8, MKSD, and MKRef) formed a tight clonal cluster characterized by low IBS0 values. Notably, the reference genome (MKRef) was connected to this cluster through slightly higher IBS0, suggesting close but non-clonal pedigree relationships. The D101 lineage (D101J1, D101J2, D101MT1, D101SD) formed a compact subclade radiating from KGSD, with uniformly low IBS0 (lightest edges), a pattern similarly observed for the D13 cluster. D24SD was linked only to KGSD node and with relatively high IBS0, indicating distant but detectable co-ancestry. A separate Kan Yao/Black Thorn triad (KanYao, BTP, BTJ6) formed an independent module, with BTP and BTJ6 displaying very low IBS0 (clonal), while KanYao exhibited higher IBS0, indicating moderate divergence within this lineage. In contrast, KGSwk, originating from Sarawak, Malaysia, remained completely isolated from all other lineages.

ITS1-5.8 S-ITS2 Haplotyping Uncovers Multiple Paralogous Variants

To further elucidate genetic relationships among cultivars and complement the high-resolution nuclear SNP analyses, we examined ITS1-5.8 S-ITS2 haplotype composition derived from genome-skimming assemblies of the durian samples. Initial variant calling and haplotype reconstruction identified 26 haplotypes. After excluding those occurring at $< 5\%$ within-sample frequency to minimize potential sequencing or assembly artifacts, 12 dominant haplotypes (Haplotype_0, _1, _2, _6, _7, _8, _9, _10, _11, _15, _21, and _22) were retained for downstream analyses. A presence/absence heatmap (Fig. 5) revealed partial lineage-specific patterns. Haplotype_0, _2, and _6 were relatively common and shared among multiple cultivars, indicating limited diagnostic value.

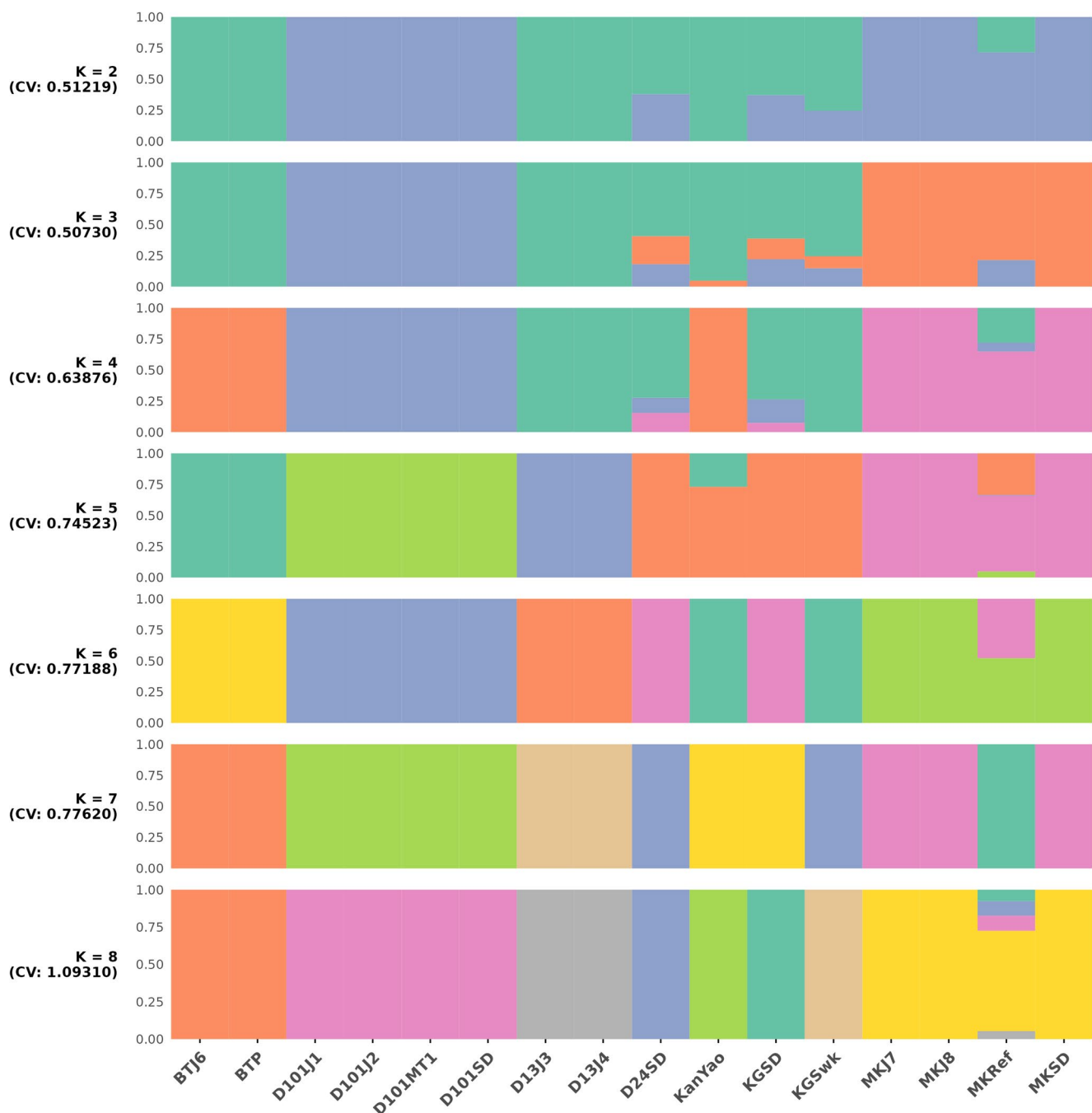


Fig. 3 Genetic structure and ancestry estimation of durian (*Durio zibethinus*) cultivars. Admixture analysis structure plots showing the proportional ancestral contributions to individual samples across a range of modeled clusters ($K = 2$ to $K = 8$). The height of each distinct colored segment within a vertical bar represents the estimated membership coefficient (ancestry percentage) of that specific genetic cluster

In contrast, Haplotype_1 was unique to the Black Thorn cluster (BTP and BTJ6), potentially serving as a strong diagnostic marker for this premium variety pending additional genomic sampling. Notably, the MKRef sample was distinguished by a private Haplotype_22 not detected in any other samples.

for the corresponding sample. The cross-validation (CV) error for each run is provided in parentheses on the left axis. Cultivar abbreviations on the horizontal axis correspond to sampled groups, including Black Thorn (BT), D101, D13, D24, Kan Yao, Kampung (KG), and Musang King (MK)

Discussions

The successful generation of high-quality Illumina whole genome sequencing data from durian flesh, historically regarded as a challenging issue due to its high polysaccharide

Fig. 4 KING kinship network plot illustrating the pairwise kinship relationships among durian samples. Connections (edges) are drawn only between sample pairs with a KING kinship value ≥ 0.04 . Edge colors represent the IBS0 values calculated from the KING analysis, where lower IBS0 indicates a higher proportion of shared alleles and higher IBS0 indicates fewer shared alleles. Nodes are colored according to cultivar

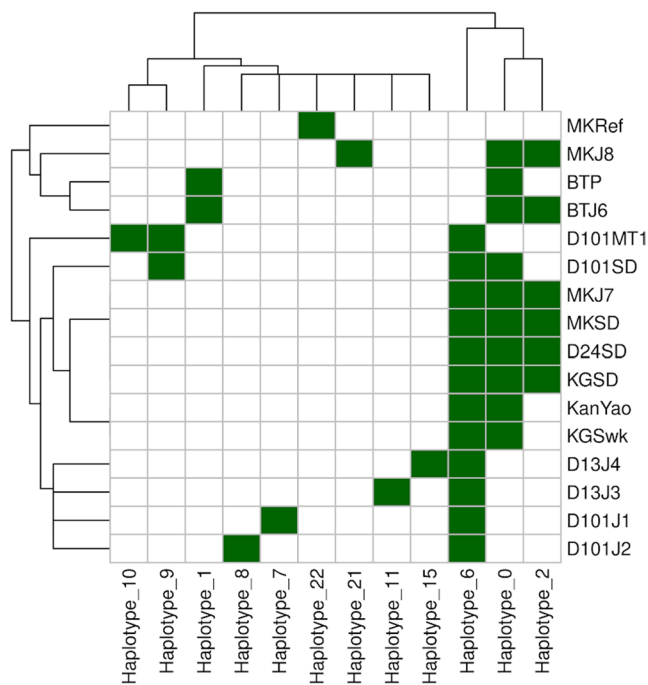
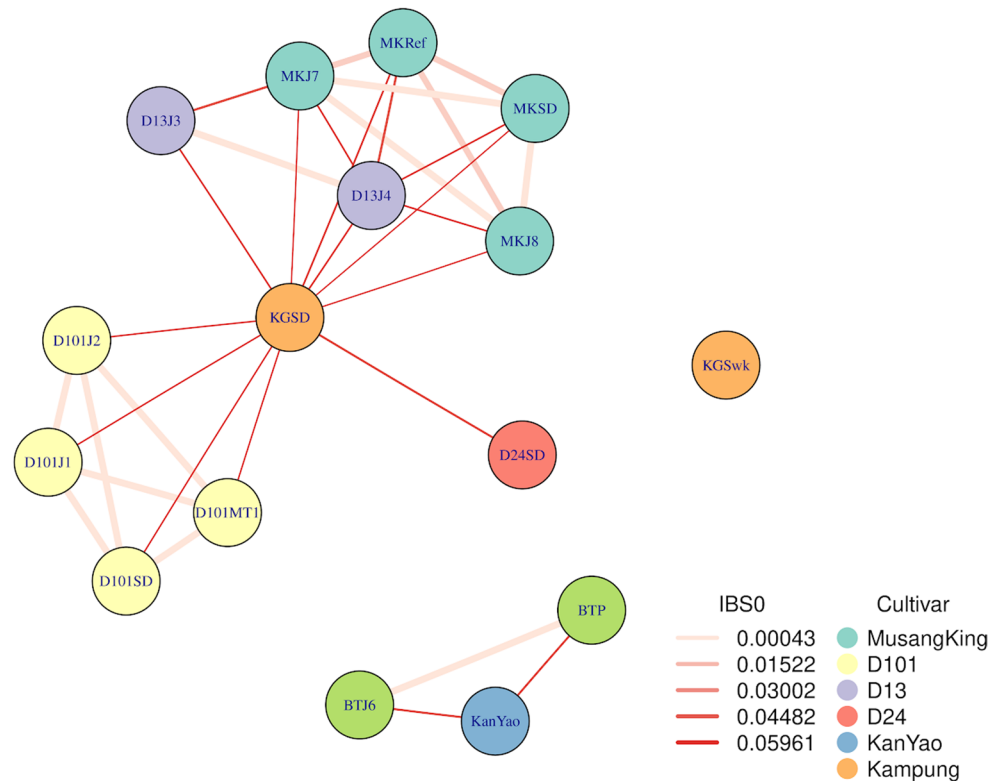


Fig. 5 Presence/absence heatmap of ITS1-5.8S-ITS2 haplotypes across durian samples. Green indicates haplotype presence, and white indicates absence. Both samples and haplotypes were subjected to hierarchical clustering. A haplotype was considered present if its frequency exceeded 5% in a sample

content (Ng et al. 2025), represents a promising advance for this species. The fact that flesh tissue yields whole-genome data of sufficient quality and depth across both nuclear and plastid genomes means that cultivar authenticity can be verified directly from the flesh sold at stalls or packaged for retail, offering a convenient and non-destructive alternative to leaf sampling that typically requires access to farms. Such an approach has strong implications for molecular authentication and traceability down to the level of individual fruit portions. Interestingly, our sequencing data revealed a paradoxically higher proportion of chloroplast reads in the non-green flesh compared to leaf tissue. This unexpected enrichment likely stems from a tissue-specific bias during DNA extraction. Unlike leaf chloroplasts, which are mechanically stabilized by an intricate internal thylakoid network (Kirchhoff 2018), fruit plastids lose this architecture during ripening and dedifferentiate into highly fragile and vesicle-filled sacs (Egea et al. 2010). Consequently, standard mechanical and chemical lysis parameters preferentially rupture these destabilized fruit plastid envelopes thus increasing the proportion of plastid DNA in the lysate compared to the more resilient chloroplasts found in leaf tissue.

A key limitation of the present study is the small sample size (14 samples) and the uneven representation across

cultivars, with some cultivar groups represented by one or two individuals. This limits strong inference on within-cultivar diversity, clonality and wider population structure. While increasing the sampling breadth would be desirable, it is acknowledged that this would be beyond the scope of the current study. Such limitations should be explicitly taken into account when considering the generality of the findings. The stratification observed in the principal component analysis indicates that Malaysian durian cultivars have been isolated long enough to develop cultivar-specific genomic signatures (Husin et al. 2018). These patterns reflect the unique breeding histories and regional selection pressures that shaped each variety. More broadly, the observed pattern is consistent with a perennial crop system that is currently dominated by clonal propagation practice and repeated selection of elite lineages, where within-cultivar variation is reduced while between-cultivar differentiation is maintained (Siew et al. 2018). Differentiation among certain lineages became more apparent in later principal components, indicating additional layers of genetic variation beyond the primary axes (Peloso and Lunetta 2011; Zhong et al. 2025). The clustering of D24SD (D24 cultivar) with KGSD (Kampung cultivar) suggests a localized evolutionary relationship, consistent with their shared geographic origin. However, given that this observation is based on a single individual, these results should be interpreted with caution. Because Kampung durians represent open-pollinated, semi-wild seedling trees, two hypotheses emerge: (1) the local KGSD individual may be an uncultivated relative of the original parental stock from which the premium D24 clone was selected, or conversely, (2) it may represent a feral seedling directly descended from local D24 cultivation (Besnard et al. 2018). It is also important to note that some of the samples used in this study were obtained directly from commercial sellers where the identity of the cultivar is based on market labeling and may not always be independently verified. In this regard, genome-wide clustering provides an additional level of validation for cultivar assignment in most of the samples, although the possibility of mislabeling in a few samples cannot be completely ruled out and might be responsible for unexpected clustering patterns.

The Musang King reference genome (MKRef), collected from Raub in Pahang (Teh et al. 2017), exhibited a somewhat distinct genetic profile compared with the other Musang King samples analyzed in this study. Although this variation was observed across multiple analyses, it does not diminish the utility of MKRef as a reference for the *Durio zibethinus* species. Rather, it suggests that the assembly captures a specific genomic variant within what is a slightly heterogeneous cultivar, rather than a single uniform lineage. Given the high commercial value of Musang King cultivar (Husin et al. 2018; Thorogood et al. 2022), understanding

these minor genomic nuances is important for precise authentication and germplasm management. Interestingly, our sequenced samples maintained high genetic consistency across diverse geographic locations, indicating strong overall cultivar integrity. One possible explanation for this pattern is that the reference Musang King (Teh et al. 2017) may still retain residual heterozygosity from the original breeding material and/or subtle genomic variation accumulated during long-term vegetative propagation (Myles et al. 2011). In widely cultivated and commercially propagated durian cultivars, in particular the Musang king cultivar, the extensive dissemination and repeated clonal multiplication may have led to the co-existence of closely related but not strictly identical lineages under a single cultivar designation. Expanded genomic sampling of this cultivar from more known localities and reputable farms will be instructive in delineating the boundaries of normal biological variation within this Musang King lineage.

A key contribution of this work is the inclusion of two whole genome sequencing datasets for the Black Thorn cultivar, which currently commands a premium price higher than Musang King (Husin et al. 2018). This genetic link aligns well with its local oral history whereby the Black Thorn mother tree in southern Penang is anecdotally reported to have been grown from a seed brought back directly from Thailand (Gasik 2018). Our genomic data provide the first empirical evidence supporting this anecdote. Future comparative analyses that include additional Thai cultivars may provide deeper insight into these regional relationships and help map out the broader domestication transitions across the Thai-Malay peninsula. It is worth noting that although three additional Thai durian genomes have recently been published, the raw sequencing data were not made available (Nawae et al. 2023) and therefore could not be incorporated into our population-level analyses. This data gap underscores a persistent challenge in durian genomics, where the lack of shared raw data hinders the resolution of fine-scale, cross-border evolutionary histories. Nevertheless, our results highlight the distinctiveness of Black Thorn, which also possesses a unique ITS1-5.8S-ITS2 haplotype profile. Because the nuclear ribosomal ITS region is highly stable and easily amplified, this distinct haplotype could potentially serve as a preliminary, cost-effective screening tool for PCR-based diagnostic assays.

The considerable variation observed among ITS haplotypes across cultivars highlights the dynamic evolution of rDNA loci in durian. Although ITS variation did not always align with genome-wide SNP patterns, this diversity is consistent with the known heterogeneity of multicopy rDNA regions (Wang et al. 2023; Xu et al. 2017). Such discordance is expected given incomplete concerted evolution and paralogous sequence variation within rDNA arrays. In

clonally propagated crops such as durian, nuclear genome-wide variation is expected to be relatively constrained within cultivars, whereas multicopy loci may retain additional layers of intragenomic diversity. By reconstructing haplotypes directly from whole-genome sequencing reads, instead of from consensus assemblies, our approach captures a richer representation of ITS diversity that would otherwise be obscured. However, low frequency variants should be interpreted with caution because of the biases that can be introduced by sequencing errors and unequal read coverage. The 5% frequency threshold used here is a conservative filter to reduce noise, but lower frequency variants may also be true biological signals due to multicopy rDNA heterogeneity and should be further investigated in future studies. Overall, the ITS data should be interpreted as a complementary marker system rather than a reflection of genome-wide population structure.

Acknowledgements We extend our sincere thanks to Datuk Leow Cheok Kiang for providing the Black Thorn leaf samples and for facilitating access to his farm for sample collection.

Author Contributions HHC, LYL, LWKL and HMG wrote the main manuscript text. ABMF, LYL and HMG performed sample collection and lab validation. HMG and HHC performed data analysis and interpretation. HHC, LWKL and HMG and LYL prepared figures. All authors reviewed the manuscript.

Funding This work was supported by the AgriNXT 2023 Technology Challenge (Problem Statement AGRI 1: “Improving the accuracy and efficiency of durian variant identification”), funded by the Malaysian Communications and Multimedia Commission (MCMC). The authors also acknowledge the support of Bioeconomy Corporation and the Ministry of Agriculture and Food Security (KPKM) as organizing partners of the AgriNXT program.

Data Availability Raw paired-end FASTQ sequencing data for all durian samples generated in this study have been made publicly available in NCBI under BioProject PRJNA1175461. Variant calling data, ITS1-5.8S-ITS2 haplotype sequences and KING kinship analysis output have been uploaded to Zenodo under the DOI (<https://doi.org/10.5281/zenodo.17479799>).

Declarations

Competing interests The authors declare no competing interests.

References

- Alexander DH, Lange K (2011) Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics* 12:246. <https://doi.org/10.1186/1471-2105-12-246>
- Bengtsson-Palme J, Ryberg M, Hartmann M, Branco S, Wang Z, Godhe A, De Wit P, Sánchez-García M, Ebersberger I, de Sousa F, Amend A, Jumpponen A, Unterseher M, Kristiansson E, Abarenkov K, Bertrand YJK, Sanli K, Eriksson KM, Vik U, Veldre V, Nilsson RH (2013) Improved software detection and extraction of ITS1 and ITS2 from ribosomal ITS sequences of fungi and other eukaryotes for analysis of environmental sequencing data. *Methods Ecol Evol* 4:914–919. <https://doi.org/10.1111/2041-210X.12073>
- Besnard G, Terral J-F, Cornille A (2018) On the origins and domestication of the olive: a review and perspectives. *Ann Bot* 121:385–403. <https://doi.org/10.1093/aob/mcx145>
- Chang C, Chow C, Laurent CAM, Tellier, Vattikuti S, Shaun M Purcell, Lee J (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *Giga Science* 4(1). <https://doi.org/10.1186/s13742-015-0047-8>
- Chen S, Zhou Y, Chen Y, Gu J (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34:i884–i890. <https://doi.org/10.1093/bioinformatics/bty560>
- Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H (2021) Twelve years of SAMtools and BCFtools. *GigaScience* 10:giab008. <https://doi.org/10.1093/gigascience/giab008>
- Egea I, Barsan C, Bian W, Purgatto E, Latché A, Chervin C, Bouzayen M, Pech J-C (2010) Chromoplast Differentiation: Current Status and Perspectives. *Plant Cell Physiol* 51:1601–1611. <https://doi.org/10.1093/pcp/pcq136>
- Fu L, Niu B, Zhu Z, Wu S, Li W (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28:3150–3152. <https://doi.org/10.1093/bioinformatics/bts565>
- Gasik L (2018) The Whole Black Thorn Durian Story. Year Durian. URL <https://www.yearofthedurian.com/2018/05/black-thorn-durian-story.html>. Accessed 6.12.26
- Husin NA, Rahman S, Karunakaran R, Bhore SJ (2018) A review on the nutritional, medicinal, molecular and genome attributes of Durian (*Durio zibethinus* L.), the King of fruits in Malaysia. *Bioinformation* 14:265–270. <https://doi.org/10.6026/97320630014265>
- Inglis PW, Pappas M, de Resende CR, Grattapaglia LV, D (2018) Fast and inexpensive protocols for consistent extraction of high quality DNA and RNA from challenging plant and fungal samples for high-throughput SNP genotyping and sequencing applications. *PLoS ONE* 13:e0206085. <https://doi.org/10.1371/journal.pone.0206085>
- Ji X, Zhong Y, Zheng D, Xie S, Shi M, Wang X, Liu F, Feng X, Wang H (2025) Chromosome-scale haploid genome assembly of *Durio zibethinus* KanYao. *Sci Data* 12:384. <https://doi.org/10.1038/s41597-025-04656-y>
- Khoo GC (2025) Durian matters. *Continuum* 39:211–217. <https://doi.org/10.1080/10304312.2024.2421883>
- Kirchhoff H (2018) Structure-function relationships in photosynthetic membranes: Challenges and emerging fields. *Plant Sci* 266:76–82. <https://doi.org/10.1016/j.plantsci.2017.09.021>
- Knyazev S, Tsyvina V, Shankar A, Melnyk A, Artyomenko A, Malygina T, Porozov YB, Campbell EM, Switzer WM, Skums P, Mangul S, Zelikovsky A (2021) Accurate assembly of minority viral haplotypes from next-generation sequencing through efficient noise reduction. *Nucleic Acids Res* 49:e102. <https://doi.org/10.1093/nar/gkab576>
- Lefouili M, Nam K (2022) The evaluation of Bcftools mpileup and GATK HaplotypeCaller for variant calling in non-human species. *Sci Rep* 12:11331. <https://doi.org/10.1038/s41598-022-15563-2>
- Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*.
- Li D, Liu C-M, Luo R, Sadakane K, Lam T-W (2015) MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 31:1674–1676. <https://doi.org/10.1093/bioinformatics/btv033>
- Liu J, Shen Q, Bao H (2022) Comparison of seven SNP calling pipelines for the next-generation sequencing data of chickens. *PLoS ONE* 17:e0262574. <https://doi.org/10.1371/journal.pone.0262574>
- Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM (2010) Robust relationship inference in genome-wide association studies. *Bioinformatics* 26(22):2867–73.

- Myles S, Boyko AR, Owens CL, Brown PJ, Grassi F, Aradhya MK, Prins B, Reynolds A, Chia J-M, Ware D, Bustamante CD, Buckler ES (2011) Genetic structure and domestication history of the grape. *Proc Natl Acad Sci* 108:3530–3535. <https://doi.org/10.1073/pnas.1009363108>
- Nawae W, Naktang C, Charoensri S, U-thoomporn S, Narong N, Chusri O, Tangphatsornruang S, Pootakham W (2023) Resequencing of durian genomes reveals large genetic variations among different cultivars. *Front Plant Sci* 14. <https://doi.org/10.3389/fpls.2023.1137077>
- Ng WL, Wong XJ, Ramaiya SD, Sabran S, Hung BM, Lee SY (2025) Extraction of useful DNA from different parts of the durian (*Durio zibethinus*) fruit. *Pertanika J Trop Agric Sci* 48. <https://doi.org/10.47836/pjtas.48.4.03>
- Pedersen BS, Quinlan AR (2018) Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* 34:867–868. <https://doi.org/10.1093/bioinformatics/btx699>
- Peloso GM, Lunetta KL (2011) Choice of population structure informative principal components for adjustment in a case-control study. *BMC Genet* 12:64. <https://doi.org/10.1186/1471-2156-12-64>
- Shearman JR, Sonthirod C, Naktang C, Sangrakru D, Yoocha T, Chatbanyong R, Vorakuldumrongchai S, Chusri O, Tangphatsornruang S, Pootakham W (2020) Assembly of the durian chloroplast genome using long PacBio reads. *Sci Rep* 10:15980. <https://doi.org/10.1038/s41598-020-73549-4>
- Siew GY, Ng WL, Tan SW, Alitheen NB, Tan SG, Yeap SK (2018) Genetic variation and DNA fingerprinting of durian types in Malaysia using simple sequence repeat (SSR) markers. *PeerJ* 6:e4266. <https://doi.org/10.7717/peerj.4266>
- Sokolov EP (2000) An improved method for DNA isolation from mucopolysaccharide-rich molluscan tissues. *J Molluscan Stud* 66:573–575. <https://doi.org/10.1093/mollus/66.4.573>
- Teh BT, Lim K, Yong CH, Ng CCY, Rao SR, Rajasegaran V, Lim WK, Ong CK, Chan K, Cheng VKY, Soh PS, Swarup S, Rozen SG, Nagarajan N, Tan P (2017) The draft genome of tropical fruit durian (*Durio zibethinus*). *Nat Genet* 49:1633–1641. <https://doi.org/10.1038/ng.3972>
- Thorogood CJ, Ghazalli MN, Siti-Munirah MY, Nikong D, Kusuma YWC, Sudarmono S, Witono JR (2022) The king of fruits. *PLANTS PEOPLE PLANET* 4:538–547. <https://doi.org/10.1002/ppp3.10288>
- Wang W, Zhang X, Garcia S, Leitch AR, Kovarik A (2023) Intragenomic rDNA variation - the product of concerted evolution, mutation, or something in between? *Heredity* 131:179–188. <https://doi.org/10.1038/s41437-023-00634-5>
- Xu B, Zeng X-M, Gao X-F, Jin D-P, Zhang L-B (2017) ITS non-concerted evolution and rampant hybridization in the legume genus *Lespedeza* (Fabaceae). *Sci Rep* 7:40057. <https://doi.org/10.1038/srep40057>
- Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS (2012) A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* 28:3326–3328. <https://doi.org/10.1093/bioinformatics/bts606>
- Zhong Y, Feng L, Deng H, Ji X, Zhang J, Sun Y, Lin P, Qiao Y, Xie S, Wang H, Guo L, Feng X (2025) Genomic resequencing reveals genetic diversity, population structure, and core collection of durian germplasm. *Commun Biol* 8:1273. <https://doi.org/10.1038/s42003-025-08715-3>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.