Al-Farabi Kazakh National University (KazNU)

Faculty of Biology and Biotechnology

Department of Molecular Biology and Genetics

Discipline: «Modern Problems of Plant Genetics»

Lecture 2

Challenging Features of Plant Genomes

Genome size variation (C-value paradox), LTR retrotransposons, autopolyploidy vs allopolyploidy, bread wheat evolution, heterozygosity, T2T assembly & pangenomics

Author: Amirova Aigul Kuzembayevna, Candidate of Biological Sciences, Associate Professor
E-mail: aigul_amir@mail.ru

Aim and Structure of the Lecture

Aim: understand why plant genomes are structurally among the most complex eukaryotic genomes and how genomics overcomes these challenges

KEY FOCUS

Lecture Plan & Modules

Key sections:
1. Tripartite Plant Genome
2. Genome Size & C-Value Paradox
3. Transposable Elements (TEs)
4. Polyploidy & Wheat Evolution
5. Heterozygosity & Repeats
6. T2T & Plant Pangenomics

- **Tripartite Plant Genome:** Coordinated organization of nuclear, plastid (chloroplast), and mitochondrial DNA
- **Genome Size & C-Value Paradox:** 2,400-fold size variation from Genlisea (~61 Mb) to Paris japonica (~149 Gb)
- **Transposable Elements (TEs):** Class I LTR retrotransposons (Copia, Gypsy) and Class II DNA transposons shaping genome architecture
- **Polyploidy & Wheat Evolution:** Autopolyploidy vs allopolyploidy, diploidization, and origins of hexaploid wheat (AABBDD)
- **Heterozygosity & Repeats:** Assembly bottlenecks in outcrossing trees, clonal crops, and heterozygous haplotypes
- **T2T & Plant Pangenomics:** Telomere-to-telomere phasing and core vs dispensable genomes for crop improvement

Tripartite Organization of the Plant Cell Genetic System

Unlike animal cells (2 genomes), plant cells coordinate three distinct DNA compartments with endosymbiotic transfer

Part I: Core Mechanisms

Linear chromosomes (61 Mb – 149 Gb)

- Houses 25,000–100,000+ protein-coding genes, vast repetitive TE landscapes, and complex chromatin domains
- Biparental Mendelian inheritance with extensive inter-organellar gene transfers over evolutionary time

Circular dsDNA (~120–160 kb)

- **Quadripartite structure:** Large Single Copy (LSC), Small Single Copy (SSC), and two Inverted Repeats (IRa/IRb)
- Encodes ~110–130 conserved photosynthesis genes (rbcL, matK); high copy number (up to 10,000/cell) & maternal inheritance

Part II: Applications & Balance

Dynamic Multipartite Network (200 kb – 11 Mb)

- Unlike compact animal mtDNA (16.5 kb), plant mtDNA is huge, rich in non-coding introns, and constantly recombines
- **Encodes chimeric ORFs causing Cytoplasmic Male Sterility (CMS) —** vital for commercial F1 hybrid seed production

Endosymbiotic Gene Transfer & Compatibility

- Nuclear PPR (pentatricopeptide repeat) Rf genes evolve to restore fertility by silencing mitochondrial CMS transcripts
- Somatic cybrids and wide crosses must maintain compatible nuclear–organellar protein complexes (e.g., Rubisco, ATP synthase)

The C-Value Paradox: 2,400-Fold Genome Size Diversity in Angiosperms

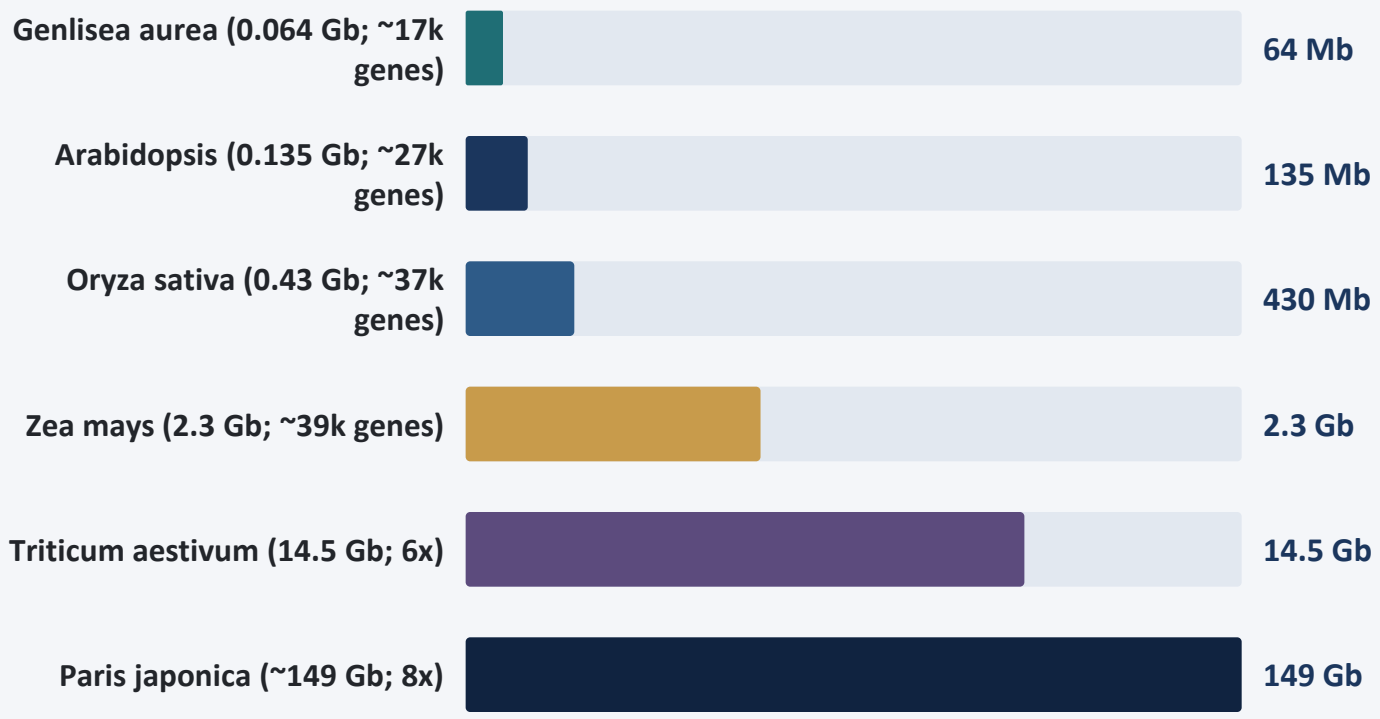
Haploid nuclear DNA content (1C, Mbp/Gbp) is uncorrelated with organismal complexity or gene count

~2,400x
Range (Genlisea to Paris)

61 Mb
Genlisea tuberosa (smallest)

~149 Gb
Paris japonica (octoploid)

Comparison of 1C Haploid Genome Sizes (Gbp, illustrative scale) & Gene Counts



Why the Paradox Exists

- Arabidopsis (135 Mb) and maize (2.3 Gb) possess almost the same basic gene repertoire (~27–39k genes) per diploid set
- Genome obesity is driven primarily by bursts of LTR retrotransposon proliferation and whole-genome duplications (WGD)
- Large genomes carry higher metabolic costs for nucleotide replication (P/N demand) and longer cell cycle times

The «Dark Matter» of Plant Genomes: Transposable Elements and Repeats

Over 50–85% of cereal and conifer genomes consists of mobile genetic elements and tandem satellite arrays

Repeat Class	Transposition Mechanism	Key Plant Superfamilies	Impact on Genome Structure & Breeding
Class I: LTR Retrotransposons	Copy-and-paste via RNA intermediate & reverse transcriptase	Ty1-copia (even dispersion) & Ty3-gypsy (pericentromeric)	Up to 80% of maize & wheat DNA; drive rapid genome expansion and heterochromatin
Class I: Non-LTR Elements	RNA intermediate without long terminal repeats	LINEs and SINEs	Insert near promoters/genes, creating novel splice variants or gene silencing
Class II: DNA Transposons	Cut-and-paste via transposase; MITEs amplify in copy number	Mutator (Mu), hAT (Ac/Ds), CACTA, Stowaway/Tourist MITEs	McClintock's Ac/Ds discovery; MITEs sit inside gene promoters modifying stress response
Tandem Repeats & rDNA	Unequal crossing-over and replication slippage	Centromeric satellites, telomeres (TTTAGGG)n, 45S/5S rDNA	Essential for kinetochore attachment; primary cause of gaps in pre-T2T genome assemblies

From «Junk DNA» to Regulatory Innovation: Positive Roles of Transposons

Barbara McClintock's «controlling elements» supply fresh cis-regulatory variation for domestication and stress adaptation

KEY FOCUS

Core Analytical Focus

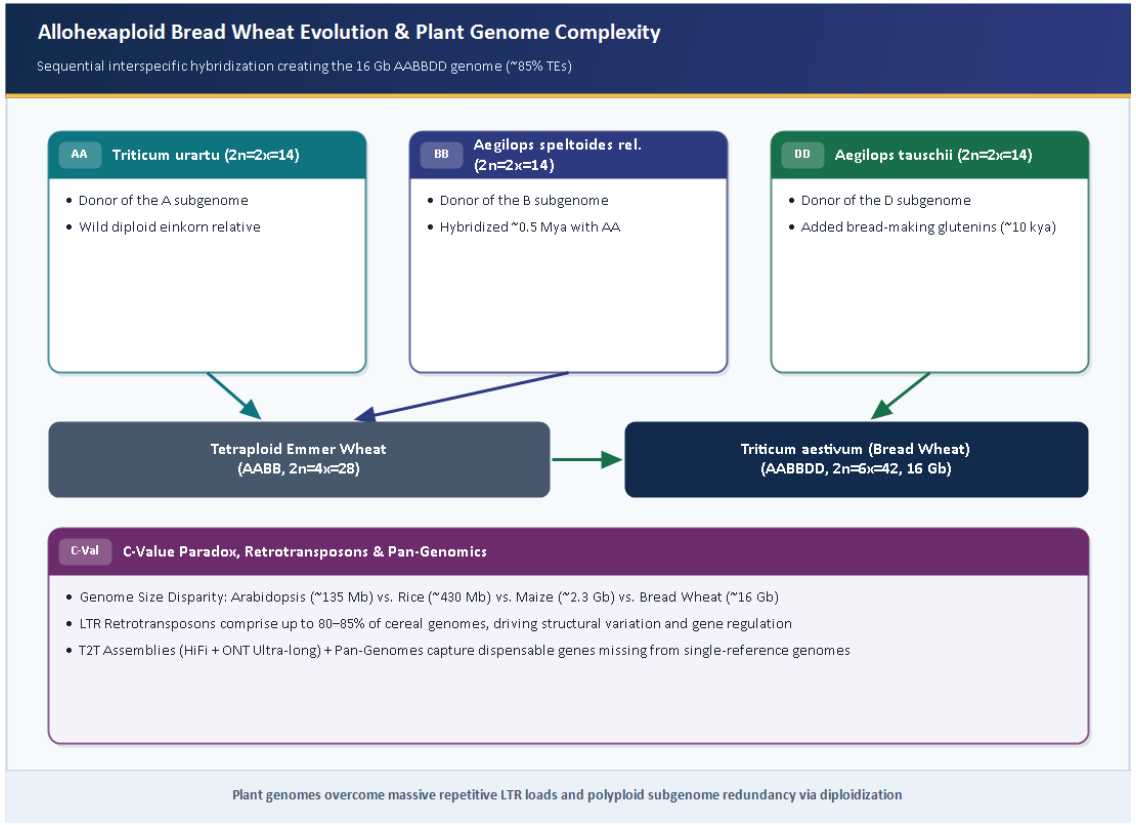
Key sections:
1. tb1 (Teosinte Branched 1)
2. MYB Anthocyanin Regulators
3. Cold-Inducible Tcs1 Retrotransposon
4. RNA-Directed DNA Methylation (24-nt siRNA)

- **tb1 (Teosinte Branched 1):** Insertion of a Hopscotch retrotransposon 60 kb upstream of tb1 acts as an enhancer elevating gene expression; Transformed multi-stemmed bushy wild teosinte into unbranched, single-stalked domesticated maize
- **MYB Anthocyanin Regulators:** Retrotransposon insertions inside or upstream of VvMYBA1 (grape) and MdMYB10 (red-fleshed apple) alter fruit pigmentation; Explains white- vs red-berry sports and rapid phenotypic variation in clonally propagated cultivars
- **Cold-Inducible Tcs1 Retrotransposon:** A cold-activated Copia element inserted upstream of the Ruby transcriptional activator turns on anthocyanin synthesis after frost; Demonstrates how stress bursts unleash TE-mediated host gene expression
- **RNA-Directed DNA Methylation (24-nt siRNA):** Plants deploy 24-nt siRNAs, CMT2/CMT3, and DRM2 methyltransferases to keep TEs transcriptionally silenced; TE methylation frequently spreads to neighboring agronomic genes, creating heritable epialleles

Polyploidy (Whole-Genome Duplication): Autopolyploidy vs Allopolyploidy

Virtually all flowering plants are ancient paleopolyploids; ~35% of vascular plants are recent polyploids

- **Autopolyploidy (Within One Species:** AAAA): Originates from somatic doubling or fusion of unreduced 2n gametes within a single biological species
- Possesses multiple homologous sets (4 or more copies per locus) that form multivalents (quadrivalents) during meiosis
- Follows complex polysomic inheritance (tetrasomic ratios such as 35:1 instead of 3:1), complicating QTL mapping and selection
- **Agronomic examples:** autotetraploid potato (Solanum tuberosum, 2n=4x=48), alfalfa (Medicago sativa), seedless triploid watermelon (3x)
- **Allopolyploidy (Hybridization Between Species:** AABB / AABBDD): Originates from interspecific hybridization followed by chromosome doubling (hybrid fertility restoration)
- Distinct divergent subgenomes (homeologs) pair strictly as homologous bivalents in meiosis via pairing-control loci (like Ph1 in wheat)



Bread wheat allohexaploidy (AABBDD), autopolyploidy & TE dynamics

Two-Step Allopolyploid Evolution of Bread Wheat (2n = 6x = 42, AABBDD)

How three wild diploid Triticeae grasses (~7 chromosomes each) combined into the world's most complex major food crop genome

1 Step 1 (~0.5–0.8 MYA)

Triticum urartu (AA, 2n=14) × wild Aegilops speltoides-like (BB, 2n=14)

2 Tetraploid Emmer / Durum

Chromosome doubling yielded Triticum turgidum (AABB, 2n=28) domesticated in Fertile Crescent

3 Step 2 (~8,000–10,000 BP)

Cultivated emmer wheat (AABB) naturally hybridized with goatgrass Aegilops tauschii (DD, 2n=14)

4 Bread Wheat (6x = 42)

Spontaneous doubling formed Triticum aestivum (AABBDD), adding baking gluten & cold hardiness

5 Ph1 Pairing Control

The Ph1 locus on 5BL suppresses homeologous pairing, ensuring stable disomic bivalent inheritance

The ~14.5 Gb bread wheat genome (~5× human genome) contains >85% repetitive elements and ~107,891 high-confidence homeologous genes in triplets (A, B, D).

Genomic Aftermath of Polyploidy: Fractionation and Subgenome Dominance

Following whole-genome duplication, nascent polyploids undergo rapid structural, genetic, and epigenetic rewiring

01 **SUBGENOME DOMINANCE — Asymmetric Homeolog Retention & Expression:** One parental subgenome frequently retains more functional genes and higher mRNA expression across tissues; Driven by lower TE density near genes in the dominant subgenome (observed in Brassica, maize, and strawberry)

02 **GENE FRACTIONATION — Biased Loss of Redundant Paralogs:** Duplicated genes face pseudogenization (non-functionalization) unless preserved by dosage balance or divergence; Homeolog triad redundancy in wheat (A+B+D) masks recessive mutations during mutagenesis — requiring triple knockouts by CRISPR!

03 **NEOFUNCTIONALIZATION —** Evolution of Novel Traits & Subfunctionalization: Relaxed selection allows one copy to acquire new tissue expression or enzymatic roles (neofunctionalization); Or ancestral functions divide reciprocally between duplicates (subfunctionalization — e.g., organ-specific starch synthases)

04 **DIPLOIDIZATION — Chromosome Fusion & Meiotic Stabilization:** Over millions of years, nested chromosome fusions and rearrangements restore diploid-like meiosis and chromosome counts; Ancient paleopolyploidy waves (e.g., core eudicot γ triplication) underlie floral and fruit diversity

Heterozygosity, Outcrossing, and the Challenge of Phased Genome Assembly

Why heterozygous perennial crops (apple, grape, potato) are harder to assemble than inbred rice or wheat

- **Self-Pollinating / Inbred Genomes (Low Heterozygosity):** Repeated selfing drives homozygosity (>99%) across homologous chromosomes (rice, Arabidopsis, bread wheat lines)
- Maternal and paternal haplotypes are nearly identical, producing a single clean haploid consensus sequence
- Relatively straightforward contig scaffolding, though large polyploid homeologs (A vs B vs D) still require long reads to avoid chimeric collapses
- **Outcrossing / Clonal Perennial Crops (1–5% Heterozygosity):** Self-incompatibility or vegetative propagation maintains extreme structural and SNP divergence between homologous chromosomes (apple Malus sieversii, potato, grapevine)
- Early short-read assemblers collapsed divergent alleles or produced fragmented «bubble» assemblies with inflated duplicated contigs
- **Modern solution:** Trio-binning, Hi-C phasing + PacBio HiFi reads resolve complete diploid or autotetraploid haplotype-resolved assemblies (e.g., phased 4-haplotype potato genome)



Long-read sequencing overcoming heterozygosity in phased genome assembly

Anatomy of a Modern Telomere-to-Telomere (T2T) Plant Genome Assembly

Combining accurate long reads, ultralong reads, and 3D chromatin proximity maps eliminates historic assembly gaps

1 1. PacBio HiFi (15–25 kb)

Q30+ accuracy cleanly distinguishes >99% similar LTR retrotransposons and wheat subgenomes

2 2. ONT Ultralong (>100 kb)

Spans intact megabase centromeric satellite repeats, NOR rDNA arrays, and large structural inversions

3 3. Hi-C Chromatin Conformation

Physical proximity ligation anchors contigs to chromosome arms and separates parental haplotypes

4 4. Telomere & Gap Polishing

Verify Arabidopsis-type (TTTAGGG)n telomeric caps on every chromosome end (0 Ns remaining)

5 5. Iso-Seq & Functional Annotation

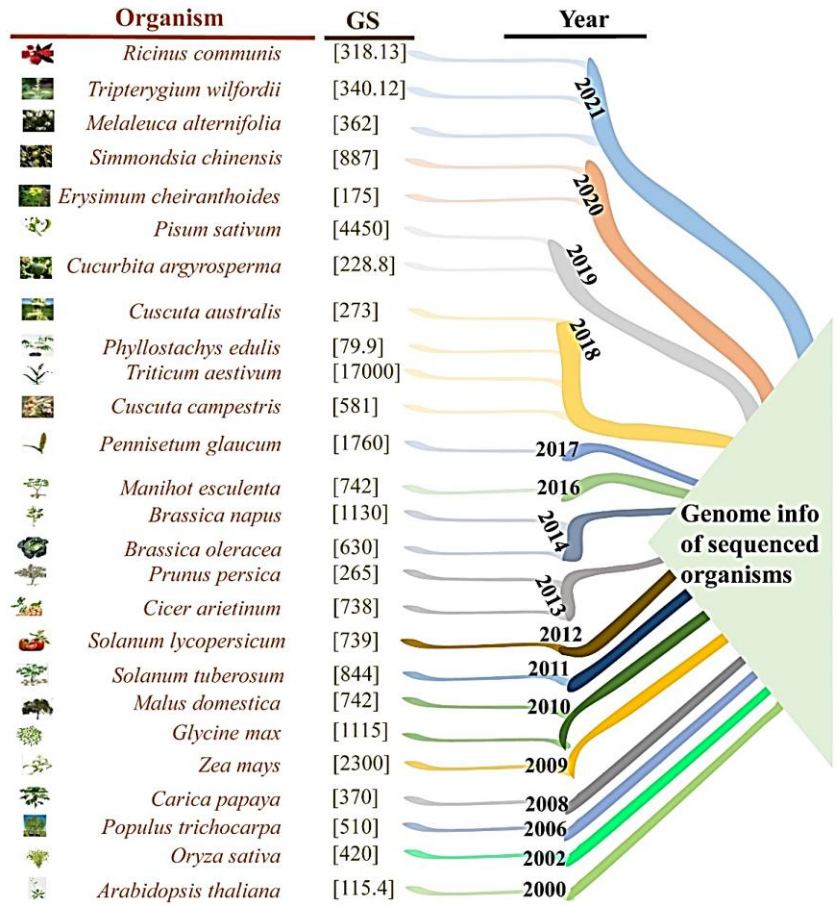
Full-length cDNA long reads map alternative splicing isoforms and non-coding RNAs accurately

T2T gapless genomes of rice, maize, soybean, apple, and Chinese Spring wheat (2024–2025) revealed previously invisible centromeric evolution and disease resistance clusters.

From One Reference Genome to the Plant Pangenome Paradigm

A single reference cultivar omits up to 20–40% of the gene families and structural variants present across a crop species

- **Core Genome vs Dispensable (Accessory) Genome:** Core genome (~60–75% of gene families): present in every accession of the species; enriched in housekeeping, photosynthesis, meiosis, and central metabolism
- **Shell & Cloud (Dispensable) genome (~25–40%):** present in only a subset or single accession via Presence–Absence Variants (PAVs) and Copy Number Variants (CNVs)
- Dispensable genes are heavily enriched in NLR immune receptors, secondary metabolism, and abiotic stress response — key targets for climate resilience!
- **Super-Pangenomes & Structural Variants (SVs):** Super-pangenomes extend comparisons across multiple wild relatives within a genus (e.g., Oryza, Solanum, Triticum, Malus)
- Graph-based pangenomes represent SNPs, InDels, large inversions, and PAV insertions simultaneously without reference mapping bias
- **Example:** tomato and soybean graph pangenomes discovered hidden SVs driving fruit flavor, seed weight, and flowering adaptation missed by SNP-only GWAS



Evolution of sequenced plant genomes toward the pangenome paradigm

Comparative Genomics Across Reference Plant Species

Illustrating the interplay of chromosome number, ploidy, repetitive content, and gene count

Species (Crop)	Ploidy (2n)	1C Size (Mb)	Repetitive DNA (%)	Annotated Genes
Arabidopsis thaliana	2x = 10	135 Mb	~20% (compact)	~27,650
Oryza sativa (Rice)	2x = 24	380–430 Mb	~35–40%	~37,500
Zea mays (Maize)	2x = 20 (paleo-4x)	2,100–2,400 Mb	~85% (mostly LTR)	~39,000
Malus domestica / sieversii (Apple)	2x = 34 (paleo-4x)	650–700 Mb	~55–60% (heterozygous)	~45,000
Triticum aestivum (Bread Wheat)	6x = 42 (AABBDD)	14,500 Mb	>85% repetitive	~107,891

Practical Implications of Plant Genome Features for Applied Biotechnology

Translating structural genomics into precise gene editing and breeding strategies

Part I: Core Mechanisms

Triple-Allele Editing Requirement

- In hexaploid wheat (A, B, D), knocking out a susceptibility gene requires simultaneous mutation of all 6 alleles across 3 homeologs
- Conserved sgRNA design targets identical exon regions shared among A, B, and D subgenomes (e.g., TaMLO-A1/B1/D1)

Pericentromeric Linkage Drag

- Crossovers are heavily suppressed across massive heterochromatic pericentromeric blocks in barley and wheat (>50% of chromosome)
- Useful resistance genes inside low-recombination zones suffer severe linkage drag during conventional backcrossing

Part II: Applications & Balance

Choosing Open Chromatin Targets

- Heavy H3K9me2 and DNA methylation around TEs reduce CRISPR/Cas9 accessibility and transgene expression
- Epigenomic ATAC-seq and methylation maps guide optimal safe-harbor transgene integration

New Frontiers in Plant Biotechnology

- Understanding centromeric CENH3 repeats enables engineered haploid inducers (Lecture 3)
- Mitochondrial TALENs (mitoTALENs) knock out CMS-causing ORFs directly in crop mtDNA

Control Questions and Key Takeaways

Consolidating Lecture 2 concepts

? Control & Review Questions

- 1. Describe the tripartite genome of plant cells and highlight unique features of plant mitochondrial DNA (CMS).
- 2. What is the C-value paradox in plants, and what molecular mechanisms drive 2,400-fold size variation?
- 3. Compare Class I LTR retrotransposons (Ty1-copia vs Ty3-gypsy) and Class II DNA transposons, giving an example of TE-driven domestication.
- 4. Distinguish autopolyploidy from allopolyploidy and trace the two hybridization events that formed hexaploid bread wheat (AABBDD).
- 5. Why do single reference genomes fail to represent crop diversity, and how do HiFi/ONT T2T assemblies and pangenomes solve this?

✓ Key Takeaways

- Plant genomes combine nuclear, plastid, and dynamic recombinogenic mitochondrial DNA compartments.
- Retrotransposon bursts and recurrent polyploidization explain genome expansion without proportional increases in core gene count.
- Allopolyploid bread wheat (2n=6x=42, AABBDD) relies on the Ph1 locus to maintain regular bivalent pairing and fertility.
- T2T long-read assemblies and graph pangenomes capture dispensable stress-resistance genes hidden from single references.

Recommended Literature and Genomic Resources

Sources on plant genome organization, polyploidy, T2T assembly, and pangenomics

1 Main Literature

- 1. Bennett M.D., Leitch I.J. Nuclear DNA amounts in angiosperms and Plant DNA C-values database // Annals of Botany. — 2012–2024.
- 2. Wendel J.F., Jackson S.A., Meyers B.C., Wing R.A. Evolution of plant genome architecture // Genome Biology. — 2016. — Vol. 17:37.
- 3. IWGSC. Shifting the limits in wheat research and breeding using a fully annotated reference genome // Science. — 2018. — Vol. 361.
- 4. Della Coletta R. et al. How the pan-genome is changing crop genomics and improvement // Genome Biology. — 2021. — Vol. 22:3.

2 Additional Literature & Databases

- 1. Lisch D. How important are transposons for plant evolution? // Nature Reviews Genetics. — 2013. — Vol. 14. — P. 49–61.
- **2. Plant DNA C-values Database (Royal Botanic Gardens, Kew):** https://cvalues.science.kew.org
- **3. GrainGenes (Triticeae & Wheat/Barley Genomics Portal):** https://wheat.pw.usda.gov
- 4. T2T Crop Genome Consortium Publications (Nature Genetics / Molecular Plant, 2024–2025).

Thank you for your attention! • Department of Molecular Biology and Genetics, KazNU