




# When numbers matter: Rethinking the role of gene duplication on short evolutionary timescales

Freja Lindstedt  | Qiujie Zhou  | Pascal Milesi 

Department of Ecology and Genetics, Evolutionary Biology Centre, Uppsala University, and SciLifeLab, Uppsala, Sweden

## Correspondence

Pascal Milesi, Department of Ecology and Genetics, Evolutionary Biology Centre, Uppsala University, and SciLifeLab, Uppsala, Sweden.  
 Email: [pascal.milesi@scilifelab.uu.se](mailto:pascal.milesi@scilifelab.uu.se)

## KEYWORDS

adaptation, gene copy number variation, population and quantitative genomics, structural variation

The potential roles of genomic structural variations (SVs) in the control of phenotypic traits and in evolution were suggested as early as the 20th century. However, they were then overshadowed by the emphasis put on single nucleotide polymorphisms (SNPs). Recently, SVs have received renewed attention in evolutionary research due to advancements in sequencing technologies and analytical methods.

At the macroevolutionary scale, plant genomes tend to evolve faster than those of other eukaryotes, due to the prevalence of whole genome duplication events (Wendel et al., 2016). Unlike other types of structural variants, such as inversions, copy number variations (CNVs) result from unbalanced mutations that affect the dosage, or amount, of a DNA sequence. When genes are involved, the number of copies of a gene varies from one individual to another. In plants, gene copy number variations (gCNVs) are likely to be abundant due to events such as mating system shifts (the efficacy of purifying selection is reduced in selfing species), hybridization and subsequent genome rearrangements, and whole genome duplications followed by biased retention (Panchy et al., 2016; Wendel et al., 2016; Van de Peer et al., 2017). For example, in *Arabidopsis thaliana* (L.) Heynh. (Brassicaceae) 10 to 18% of all genes display CNVs (Zmienko et al., 2020; Jaegle et al., 2023). In the genus *Picea* Mill. (Pinaceae), at least 10% of the protein-coding genes display CNVs (*P. abies* (L.) H. Karst and *P. obovata* Ledeb., Q. Zhou et al., 2025; and *P. glauca* (Moench) Voss and *P. mariana* (Mill.) Britton, Sterns & Poggenburg, Prunier et al., 2017).

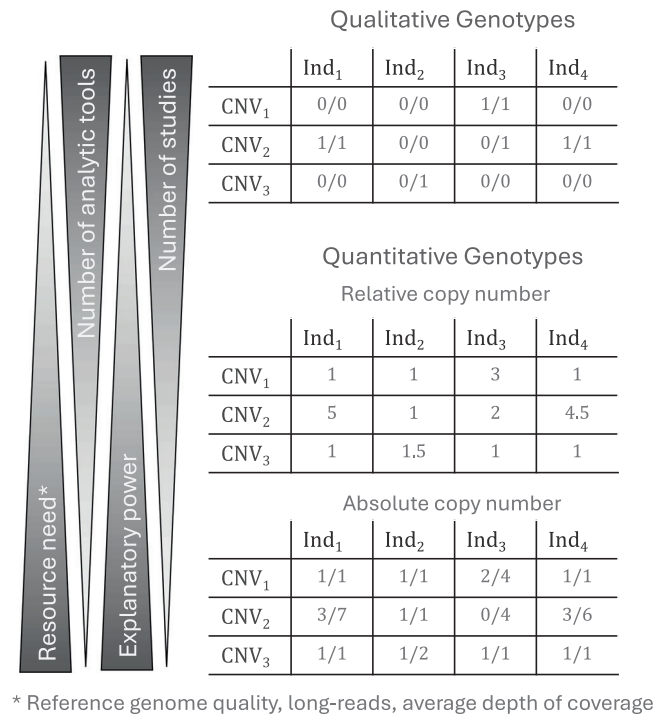
Gene duplications have primarily been studied for their roles in long-term evolution. However, a change in gene dosage

usually results in a change in the amount of gene products, such as RNA or proteins (e.g., Shao et al., 2019). Therefore, gCNVs have a unique, multiallelic, and quantitative nature. Fully apprehending their role in short-term evolutionary processes requires studying them as quantitative genotypes rather than in a presence/absence (or biallelic) manner, as is most often done (Figure 1, top panel). Unlike SNPs, the accuracy and resolution of gCNV genotyping are usually dependent on the platform used. From short-read sequencing data, one can use biased allelic ratios (Figure 2A) and changes in the depth of coverage (DoC) caused by the mis-mapping of reads from duplicated regions to the same locus in the reference genome to identify CNVs (Figure 2). However, short reads often fail to capture the underlying genetic structure of gCNVs, and changes in DoC can only be interpreted as relative copy numbers across homologous chromosomes (Figure 1, middle panel). Long-read sequencing is a promising alternative that allows for the phasing of the various alleles to obtain absolute copy numbers (see Figure 1, bottom panel). However, long-read data can still be biased in assembling repetitive regions (Carvalho et al., 2025 [preprint]), they are more computationally demanding, and likely too costly for extensive population-level genomics studies. Nevertheless, continuous advancements in sequencing technologies and CNV analysis methods open the door to more extensive studies focusing on gCNVs, even in non-model species (Karunarathne et al., 2023). In the following sections, we recognize gCNVs as a largely untapped source of genetic variation and explore their potential for studying short-term evolution in plants. We also discuss the main challenges of incorporating this type of polymorphism into population and

Freja Lindstedt and Qiujie Zhou contributed equally to this study.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2025 The Author(s). *American Journal of Botany* published by Wiley Periodicals LLC on behalf of Botanical Society of America.



**FIGURE 1** From qualitative to quantitative genotyping of gCNVs. The three tables represent genotypes with copy number variation at different loci (CNV<sub>*i*</sub>) in different individuals (Ind<sub>*i*</sub>), analyzed at different resolutions. The top table shows biallelic genotypes, where 0 generally codes for a single copy and 1 codes for more than one copy, as has mainly been done so far. The middle and bottom tables show quantitative genotypes representing different copy numbers in different individuals. These genotypes can be relative, as are those typically obtained from short-read data (middle table), or absolute, with haplotype-resolved copy number (bottom table). The latter could become the standard, particularly with the development of long-read sequencing technologies. Note that different absolute copy-number genotypes are possible for the same relative copy-number genotype.

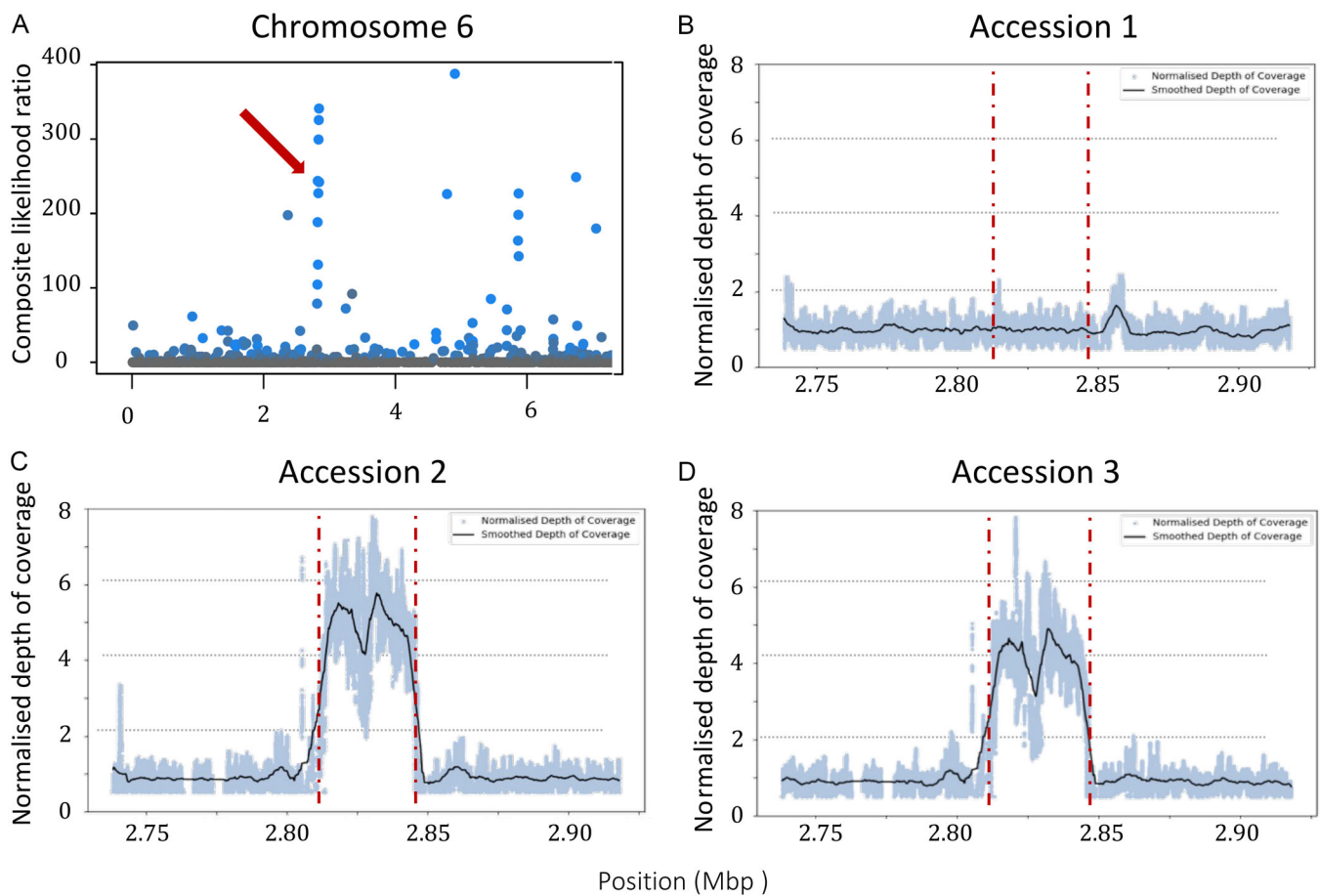
quantitative genomic frameworks, and how plants, as a study system, offer an opportunity to address them.

## UNDERSTANDING THE EVOLUTIONARY SIGNIFICANCE OF GENE COPY NUMBER VARIATIONS

Despite all the evidence that gCNVs are a non-negligible source of polymorphism, it is still unclear to what extent they can be used to address broad questions in evolutionary biology, such as better inferring past demographic events or refining predictions of population responses to global change for conservation strategies. This is partly due to difficulties in estimating key evolutionary parameters, which largely prevent their use in population and quantitative genetic models (Mérot et al., 2020). Gene copy number variation is more likely to occur via low-copy repeat mechanisms (e.g., non-allelic homologous recombination). As a result, duplications at the origin of gCNVs may span multiple genes (as illustrated in Figure 2), thus affecting the apparent gene duplication rate, which has been estimated to be higher than substitution rates (Katju and

Bergthorsson, 2013). The presence of gCNVs induces mismatches during synapsis and can originate from different molecular mechanisms (Hastings et al., 2009). Thus, variable mutation rates in terms of copy number change are expected across loci and mechanisms. Additionally, there may be asymmetric rates of copy number gain and loss, as well as state-dependent mutation rates, where the rate of change in copy number varies with the actual number of copies. Gene copy number variations can affect patterns of recombination and segregation. Therefore, models developed to study the evolution of microsatellites or gene families, such as the stepwise mutation model (SMM) and the birth–death model, respectively, do not accurately predict gCNV evolution. However, these models could produce relevant diversity summary statistics when applied to a large enough number of gCNVs to avoid gene-specific bias. For example, allele size variance (Valdes et al., 1993) and Goldstein's  $\delta\mu^2$  (Goldstein et al., 1995), which are based on relative copy number, and CNV entropy (similar to Shannon's diversity index) and  $R_{ST}$  (Slatkin, 1995), which are based on absolute copy number, could be useful for measuring within-population diversity and between-population divergence of gCNVs, respectively. Comparing these estimates with those obtained from neutral SNPs (e.g., nucleotide diversity and  $F_{ST}$ , respectively) would also inform us about the global distribution of fitness effect of gCNVs.

For structural variation, it is often considered that their phenotypic and fitness effects scales with their size. For gCNVs it means that both could increase with copy number and the apparent copy number may therefore be the result of a trade-off between a high mutation rate and large fitness effects. The highly dynamic nature of plant genome evolution makes them excellent models for addressing the knowledge gaps associated with gCNVs with >1400 reference genomes available (Bernal-Gallardo and de Folter, 2024). For example, the Brassicaceae plant family now contains the highest number of genome and transcriptome sequences for any plant lineage, allowing extensive comparative studies to address some of the challenges mentioned above and separate species-specific effects from general properties of gCNVs. Several genera have closely related species with different mating systems (selfing and outcrossing), which could be used to investigate the main evolutionary forces shaping the evolution of gCNVs while controlling for phylogenetic relationships, as has been done for RNA expression previously (e.g., Zhang et al., 2022). The change in ploidy can also be used to study mutation and recombination rates by comparing patterns of gCNVs between the sub-genomes of auto-polyploids as any differences between the sub-genomes will have been acquired after the polyploidization event. In addition, interspecific hybridization and resynthesized polyploids can be generated experimentally, and many plants are model species for functional genomics (Bernal-Gallardo and de Folter, 2024). The use of natural and synthetic resources, together with state-of-the-art sequencing and gene editing technologies, would allow for direct measurements, for example, of the phenotypic and fitness effects of gCNVs.



**FIGURE 2** Showcase of automated detection and validation of copy number variation in *Capsella rubella* Reut. (Brassicaceae) from short-read data. (A) Chromosome-wise detection of gCNVs from SNP data using the ‘CLrCNV’ method implemented in the rCNV R package. Each point represents the average composite likelihood ratio calculated across all SNPs for a given gene. Candidate gCNVs are shown in blue. (B–D) Architecture of the region indicated by the red arrow in plot A using the software *ArDu* version 1.0.0 (Claret et al., 2025). Accession 1 does not harbor the duplication, while Accession 2 and 3 have five and four copies of a ~20 Kbp region encompassing ten genes, respectively.

## GENE COPY NUMBER VARIATIONS AS KEY PLAYERS IN ADAPTATION PATTERNS

Despite the limitations exposed above, the multiallelic and quantitative nature of gCNVs makes them excellent markers for quantitative genetics as, in many cases, the phenotypic values of traits and the copy number of causal gCNVs show a quantitative relationship, enabling straightforward genotype-to-phenotype mapping. The prevalence of gCNVs in plant genomes makes them natural candidates in the control of quantitative traits and adaptation along environmental gradients.

Forest trees are particularly relevant models for studying such questions, as they often have extensive ranges with populations connected by long-distance gene flow and show strong patterns of local adaptation (Savolainen et al., 2013). In a recent study, we showed that gCNVs are widespread and involved in local adaptation in both *Picea abies* (Norway spruce) and *Picea obovata* (Siberian spruce), two keystone species of the Eurasian boreal forest (Zhou et al., 2025). Importantly, we found no overlap between

candidate genes detected purely from SNP variation and genes whose copy number correlates with environment and/or phenotypic variation. This means that, and in contrast with genomic inversions, the explanatory power of gCNVs is not captured by SNPs, and they must be specifically studied to gain a comprehensive view of the genetic architecture of local adaptation patterns and of phenotypic traits.

Recent studies have shown that considering structural variants in addition to SNPs better explains the heritability of complex traits. However, these studies tend to consider all structural variants together (including gCNVs) in a biallelic manner (Figure 1, top panel). In doing so, information about any quantitative relationship between the number of copies of a gene and the phenotypic trait of interest is lost when there are more than two copy-number states. Gene copy number variations can thus be considered largely untapped genetic variation that may even explain some of the so-called ‘missing’ heritability. Such information would be particularly relevant for plant breeding, where genomic selection and phenotypic prediction are increasingly being used to shorten breeding cycles and reduce phenotyping costs.

## CONCLUSIONS

The recent burst in high-throughput sequencing has revealed a high prevalence of gCNVs in eukaryotes, and the evolutionary significance of this polymorphism largely remains to be determined. We argue that fully comprehending the role of gCNVs in plant evolution requires solving their structure at the genomic level, but even more importantly to study them as quantitative genotypes. The rapid development of long-read sequencing technologies is full of promise for solving their haplotypic structure. However, we also want to emphasize that much can already be done by using short-read sequencing data and dedicated analytical methods as illustrated above. Extensive population-level genomic data have been generated over the past decade—for example for conservation and breeding purposes—and we hope that this paper will serve as an incentive to reanalyze these data with a focus on gCNVs. Finally, plants display valuable features—such as mating system transitions and polyploidization—for designing comparative frameworks to measure population genetics parameters much needed to fully comprehend the role of gCNVs in evolutionary response at short and intermediate timescales.

## AUTHOR CONTRIBUTIONS


F.L., P.M., and Q.Z.: Conceptualization, visualization, writing – original draft; writing – review and editing. P.M.: Funding acquisition, supervision.


## ACKNOWLEDGMENTS

We thank Samia Kousar and Jean-Loup Claret for their help with Figure 2 and Pamela Diggie and an anonymous reviewer for constructive comments on the manuscript. This work was supported by the Nilsson-Ehle Endowments from The Royal Physiographic Society of Lund, Sweden through grants number 43255 to Q.Z. and 45195 to F.L. and by the Lundman's Foundation for Botanical Studies grants from the Swedish Phytogeographic Society awarded to Q.Z.

## ORCID

Freja Lindstedt  <https://orcid.org/0009-0007-6742-3680>

Qiujie Zhou  <https://orcid.org/0000-0001-7351-2371>

Pascal Milesi  <https://orcid.org/0000-0001-8580-4291>

## REFERENCES

- Bernal-Gallardo, J. J., and S. de Folter. 2024. Plant genome information facilitates plant functional genomics. *Planta* 259: 117.
- Carvalho, A. B., B. Y. Kim, and F. Uno. 2025. Strong sequencing bias in Nanopore and PacBio prevents assembly of *Drosophila melanogaster* Y-linked genes. *bioRxiv* In press. <https://doi.org/10.1101/2025.02.23.639762>
- Claret, J.-L., C. Mestre, P. Labbé, and P. Milesi. 2025. Ar(chitecture of)Du (plications), ArDu v1.0.0.0. Website: <https://github.com/ClaretJeanLoup/ArDu/>
- Goldstein, D. B., A. Ruiz Linares, L. L. Cavalli-Sforza, and M. W. Feldman. 1995. Genetic absolute dating based on microsatellites and the origin

- of modern humans. *Proceedings of the National Academy of Sciences, USA* 92: 6723–6727.
- Hastings, P. J., J. R. Lupski, S. M. Rosenberg, and G. Ira. 2009. Mechanisms of change in gene copy number. *Nature Reviews Genetics* 10: 551–564.
- Jaegle, B., R. Pisupati, L. M. Soto-Jiménez, R. Burns, F. A. Rabanal, and M. Nordborg. 2023. Extensive sequence duplication in *Arabidopsis* revealed by pseudo-heterozygosity. *Genome Biology* 24: 44.
- Karunarathne, P., Q. Zhou, K. Schliep, and P. Milesi. 2023. A comprehensive framework for detecting copy number variants from single nucleotide polymorphism data: 'rCNV', a versatile *r* package for paralogue and CNV detection. *Molecular Ecology Resources* 23: 1772–1789.
- Katju, V., and U. Bergthorsson. 2013. Copy-number changes in evolution: Rates, fitness effects and adaptive significance. *Frontiers in Genetics* 4: 273.
- Mérot, C., R. A. Oomen, A. Tigano, and M. Wellenreuther. 2020. A roadmap for understanding the evolutionary significance of structural genomic variation. *Trends in Ecology & Evolution* 7: 561–572.
- Panchy, N., M. Lehti-Shiu, and S.-H. Shiu. 2016. Evolution of gene duplication in plants. *Plant Physiology* 171: 2294–2316.
- Prunier, J., S. Caron, M. Lamothe, S. Blais, J. Bousquet, N. Isabel, and J. MacKay. 2017. Gene copy number variations in adaptive evolution: The genomic distribution of gene copy number variations revealed by genetic mapping and their adaptive role in an undomesticated species, white spruce (*Picea glauca*). *Molecular Ecology* 26: 5989–6001.
- Savolainen, O., M. Lascoux, and J. Merilä. 2013. Ecological genomics of local adaptation. *Nature Reviews Genetics* 14: 807–820.
- Shao, X., N. Lv, J. Liao, J. Long, R. Xue, N. Ai, D. Xu, and X. Fan. 2019. Copy number variation is highly correlated with differential gene expression: A pan-cancer study. *BMC Medical Genetics* 20: 175.
- Slatkin, M. 1995. A measure of population subdivision based on microsatellite allele frequencies. *Genetics* 139: 457–462.
- Valdes, A. M., M. Slatkin, and N. B. Freimer. 1993. Allele frequencies at microsatellite loci: the stepwise mutation model revisited. *Genetics* 133: 737–749.
- Van de Peer, Y., E. Mizrahi, and K. Marchal. 2017. The evolutionary significance of polyploidy. *Nature Reviews Genetics* 18: 411–424.
- Wendel, J. F., S. A. Jackson, B. C. Meyers, and R. A. Wing. 2016. Evolution of plant genome architecture. *Genome Biology* 17: 37.
- Zhang, Z., D. Kryvokhyzha, M. Orsucci, S. Glémin, P. Milesi, and M. Lascoux. 2022. How broad is the selfing syndrome? Insights from convergent evolution of gene expression across species and tissues in the *Capsella* genus. *New Phytologist* 236: 2344–2357.
- Zhou, Q., M. Lascoux, and P. Milesi. 2025. Gene copy number variations are untapped key players in adaptation along environmental gradient, in Q. Zhou, Drivers and components of genetic diversity in boreal forest trees: The role of hybridization and gene copy number variation in the evolution of Norway and Siberian Spruce. PhD. dissertation, Uppsala University, Uppsala, Sweden. Website: <https://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-543181>
- Zmienko, A., M. Marszałek-Zenczak, P. Wojciechowski, A. Samelak-Czajka, M. Luczak, P. Kozłowski, W. M. Karłowski, and M. Figlerowicz. 2020. AthCNV: A Map of DNA copy number variations in the *Arabidopsis* genome. *Plant Cell* 32: 1797–1819.

**How to cite this article:** Lindstedt, F., Q. Zhou, and P. Milesi. 2025. When numbers matter: Rethinking the role of gene duplication on short evolutionary timescales. *American Journal of Botany* 112(7): e70072. <https://doi.org/10.1002/ajb2.70072>