

Quantification of organelle genome contamination in public *Silene latifolia* RAD-seq datasets

Ryo Ootsuki

ootsuki@komazawa-u.ac.jp

Komazawa University

Short Report

Keywords: RAD-seq, Organelle contamination, *Silene latifolia*, Chloroplast, Mitochondria, Data quality control

Posted Date: January 16th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-8606738/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Objective: Restriction site-associated DNA sequencing (RAD-seq) is widely used for linkage mapping and population/phylogenetic inference, yet off-target reads originating from chloroplast and mitochondrial genomes may bias downstream analyses. We quantified organelle-derived reads in publicly available *Silene latifolia* RAD-seq libraries and compared them with whole-genome sequencing (WGS) and cDNA (RNA-seq) datasets.

Results: We analyzed 42 libraries from public repositories (26 RAD-seq, 9 WGS, 7 cDNA). Organelle-derived reads were detected in every library. RAD-seq libraries were typically low in organelle reads (23/26 libraries <5%), but three RAD-seq runs showed extreme organelle carry-over (26.4-32.9%). Across RAD-seq libraries, organelle-mapped reads were predominantly mitochondrial rather than chloroplast. These results highlight that organelle carry-over can vary markedly among RAD-seq libraries and may benefit from routine quantification and filtering prior to SNP calling to reduce potential downstream bias.

Conclusions: Organelle-derived reads are detectable in all examined *S. latifolia* libraries and may be extreme in a subset of RAD-seq runs; routine quantification and removal of organelle-mapped reads may be a useful component of RAD-seq quality control to minimize potential downstream bias.

Introduction

Reduced-representation sequencing approaches such as RAD-seq have become standard tools for high-throughput SNP discovery and genotyping, enabling cost-effective linkage mapping, population genomics, and phylogenetic analysis [1-5]. However, RAD-seq libraries prepared from green tissue may unintentionally include a non-trivial fraction of organelle DNA (chloroplast and mitochondria), owing to high organelle copy numbers and variability in DNA extraction and size-selection efficiency. Because organelle genomes differ from the nuclear genome in inheritance patterns (often uniparental), effective population size, and evolutionary dynamics, the presence of organelle-derived reads may influence heterozygosity estimates, allele-frequency spectra, and phylogenetic signal if not removed [6].

Silene latifolia is a dioecious model system for sex chromosome evolution [7]. Extensive RAD-seq and transcriptome datasets for *S. latifolia* are publicly available, yet the extent of organelle read carry-over in these datasets has not been summarized in a concise, reusable form for secondary users. Here, we provide a practical quantification of organelle-derived reads across public *S. latifolia* RAD-seq libraries and place these observations in context using WGS and cDNA (RNA-seq) libraries as contrasts, because these non-reduced-representation datasets provide a baseline for typical organelle-mapping fractions.

Methods

Public sequencing datasets. We analyzed three library types from public archives: 26 RAD-seq libraries (SRR2050365-SRR2050390), 9 WGS libraries (SRR2102972, SRR2103365, SRR2103367, SRR2103430-

SRR2103432, SRR2103434, SRR2104765-SRR2104766), and 7 cDNA (RNA-seq) libraries (ERR020158-ERR020159, ERR020162, ERR037835-ERR037836, ERR037898-ERR037899) [12-14].

Reference organelle genomes. We used the complete chloroplast genome (NC_016730) and mitochondrial genome (NC_014487) of *S. latifolia* as mapping references.

Read mapping and counting. For each library, reads were aligned separately to the chloroplast and mitochondrial references using the Burrows-Wheeler Aligner (BWA) [8]. Mapped reads were counted using SAMtools [9] with the default mapping quality threshold (MAPQ $\geq$ 0), without suppressing multimapping reads and without duplicate removal; counts were per read (FASTQ record) and included cases where only one mate in a pair mapped. Organelle-derived reads were defined as reads with primary alignments to the chloroplast and/or mitochondrial reference. We defined the organelle-derived fraction as:

$$\text{Organelle-derived (\%)} = 100 \times (\text{N}_{\text{cp_or_mt}} / \text{N}_{\text{total}})$$

where N_{total} is the total number of sequencing reads in the library (FASTQ records; in paired-end libraries, mates are counted separately) and $\text{N}_{\text{cp_or_mt}}$ is the number of reads with primary alignments mapping to either organelle reference (union of reads mapping to the chloroplast and/or mitochondrion). To distinguish cpDNA-only, mtDNA-only, and reads mapping to both references, we extracted read identifiers from each alignment and computed set union and intersection (Supplementary Methods; Additional file 2). Reads mapping to both organelles were rare in the RAD-seq libraries (Table S1).

Summary statistics. For each library type, we summarized the distribution of organelle-derived fractions using the minimum, median, and maximum values and identified RAD-seq outliers with >25% organelle-derived reads.

Results and Discussion

Organelle reads were detected in all analyzed libraries, with organelle-derived fractions ranging from 0.5% to 32.9% across RAD-seq runs (Table S1) and from 1.1% to 16.4% in WGS runs. Among cDNA libraries, organelle-derived fractions ranged from 1.1% to 24.4%.

RAD-seq libraries were generally low in organelle reads: 23 of 26 RAD-seq runs contained <5% organelle-derived reads. Nevertheless, three RAD-seq runs showed extreme organelle carry-over (SRR2050365: 28.4%; SRR2050377: 32.9%; SRR2050386: 26.4%), indicating substantial variability among libraries (Table S1).

A key qualitative pattern was the organelle composition of mapped reads. In RAD-seq libraries, organelle-mapped reads were predominantly mitochondrial rather than chloroplast (Table S1), whereas several WGS and cDNA libraries contained more chloroplast-mapped reads in absolute terms. Reads that mapped to both organelle references were negligible in RAD-seq datasets, suggesting that the majority of organelle-like reads may be assigned unambiguously to one organelle reference.

Practical implications for secondary analyses. Even when organelle-derived reads are a minority, their presence may be consequential because organelle loci may contribute abundant, high-coverage reads. In extreme cases (>25%), the inclusion of organelle-derived reads in de novo clustering or SNP calling pipelines may inflate apparent locus depth, potentially alter allele-frequency estimates, and introduce non-nuclear signal into phylogenetic or population-genetic summaries. Therefore, incorporating an explicit organelle-filtering step prior to SNP calling may help minimize these potential effects, for example by mapping reads to organelle references and removing mapped reads, or by screening assembled loci against organelle genomes using similarity searches such as BLAST [10].

Outlook. Because RAD-seq data are often reused long after generation, reporting simple, reproducible QC metrics (such as organelle-derived fractions) may substantially improve interpretability and reusability. The workflow provided here is intended as a lightweight QC layer that can be ported to other plant RAD-seq datasets whenever organelle references are available.

Table 1. Summary of organelle-derived read fractions across library types

Library type	n	Min (%)	Median (%)	Max (%)	No. of outliers (>25%)
RAD-seq	26	0.5	3.75	32.9	3*
WGS	9	1.1	7.6	16.4	0
cDNA (RNA-seq)	7	1.1	2.3	24.4	0

Notes: Values are computed from Table S1 (Additional file 1). Outliers are defined as libraries with >25% organelle-derived reads.

*Outlier RAD-seq runs: SRR2050365, SRR2050377, SRR2050386.

Limitations

- This study is a retrospective analysis of public data; detailed wet-lab metadata (extraction protocol, tissue state, and library kit) were not consistently available, limiting inference about the causes of outlier contamination.
- Mapping-based quantification captures reads that align to organelle references. Some reads may originate from nuclear insertions of organelle DNA (NUMTs/NUPTs), so the estimated fraction should be interpreted as organelle-like rather than strictly organelle-origin reads.
- We did not re-run downstream RAD-seq clustering/SNP calling to measure the quantitative impact of organelle filtering on specific analyses (e.g., linkage maps or phylogenetic reconstructions).

Abbreviations

RAD-seq, restriction site-associated DNA sequencing; WGS, whole-genome sequencing; RNA-seq, RNA sequencing; cDNA, complementary DNA; cp, chloroplast; mt, mitochondrion; SRA, Sequence Read Archive; ENA, European Nucleotide Archive; NCBI, National Center for Biotechnology Information; BWA, Burrows–Wheeler Aligner; MAPQ, mapping quality; NUMTs, nuclear mitochondrial DNA segments; NUPTs, nuclear plastid DNA segments.

Declarations

Ethics approval and consent to participate

Not applicable. This study analyzed publicly available, de-identified sequencing datasets.

Consent for publication

Not applicable.

Availability of data and materials

All datasets analyzed in this study are publicly available. RAD-seq runs: SRR2050365–SRR2050390; WGS runs: SRR2103367, SRR2103430–SRR2103432, SRR2103434, SRR2104765–SRR2104766; RNA-seq runs: ERR020158–ERR020159, ERR020162, ERR037835–ERR037836, ERR037898–ERR037899. Per-run organelle-mapping fractions and summary statistics are provided as Table S1 (Additional file 1). A reproducible command-line workflow is provided as Additional file 2.

Competing interests

The author declares that he has no competing interests.

Funding

This work was supported by Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Number JP26650146.

Authors' contributions

RO conceived the study, performed the bioinformatic analysis of public datasets, prepared tables and supplementary files, and wrote the manuscript.

Acknowledgements

The author thanks Dr. Kazuharu Misawa and Dr. Ryosuke Imai for valuable comments on earlier versions of this manuscript.

Author details

1 Department of Natural Sciences, Faculty of Arts and Sciences, Komazawa University, Tokyo, Japan.

2 Faculty of Chemical and Biological Sciences, Japan Women's University, Tokyo, Japan.

References

1. Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, et al. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One. 2008;3(10):e3376. doi:10.1371/journal.pone.0003376.
2. Huang X, Feng Q, Qian Q, Zhao Q, Wang L, Wang A, et al. High-throughput genotyping by whole-genome resequencing. Genome Res. 2009;19(6):1068-1076. doi:10.1101/gr.089516.108.
3. Andolfatto P, Davison D, Erezyilmaz D, Hu TT, Mast J, Sunayama-Morita T, et al. Multiplexed shotgun genotyping for rapid and efficient genetic mapping. Genome Res. 2011;21(4):610-617. doi:10.1101/gr.115402.110.
4. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One. 2011;6(5):e19379. doi:10.1371/journal.pone.0019379.
5. Eaton DAR. PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. Bioinformatics. 2014;30(13):1844-1849. doi:10.1093/bioinformatics/btu121.
6. Birky CW Jr. The inheritance of genes in mitochondria and chloroplasts: laws, mechanisms, and models. Annu Rev Genet. 2001;35:125-148. doi:10.1146/annurev.genet.35.102401.090231.
7. Vyskot B, Hobza R. The genomics of plant sex chromosomes. Plant Sci. 2015;236:126-135. doi:10.1016/j.plantsci.2015.03.019.
8. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-1760. doi:10.1093/bioinformatics/btp324.
9. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-2079. doi:10.1093/bioinformatics/btp352.
10. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403-410. doi:10.1016/S0022-2836(05)80360-2.
11. National Center for Biotechnology Information (NCBI). RefSeq: *Silene latifolia* chloroplast genome (NC_016730) [Internet]. Bethesda (MD): NCBI; [cited 15 Jan 2026]. Available from: https://www.ncbi.nlm.nih.gov/nuccore/NC_016730

12. National Center for Biotechnology Information (NCBI). Sequence Read Archive (SRA): *Silene latifolia* RAD-seq libraries (SRR2050365-SRR2050390) [Internet]. Bethesda (MD): NCBI; [cited 15 Jan 2026]. Available from: <https://www.ncbi.nlm.nih.gov/sra>
13. National Center for Biotechnology Information (NCBI). Sequence Read Archive (SRA): *Silene latifolia* whole-genome sequencing libraries (SRR2103367, SRR2103430-SRR2103432, SRR2103434, SRR2104765-SRR2104766) [Internet]. Bethesda (MD): NCBI; [cited 15 Jan 2026]. Available from: <https://www.ncbi.nlm.nih.gov/sra>
14. European Nucleotide Archive (ENA). *Silene latifolia* RNA-seq datasets (ERR020158-ERR020159, ERR020162, ERR037835-ERR037836, ERR037898-ERR037899) [Internet]. Hinxton, UK: EMBL-EBI; [cited 15 Jan 2026]. Available from: <https://www.ebi.ac.uk/ena>
15. National Center for Biotechnology Information (NCBI). RefSeq: *Silene latifolia* mitochondrial genome (NC_014487) [Internet]. Bethesda (MD): NCBI; [cited 15 Jan 2026]. Available from: https://www.ncbi.nlm.nih.gov/nuccore/NC_014487

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Additionalfile1TableS1.xlsx](#)
- [Additionalfile2SupplementaryMethods.txt](#)