

QUALITY CONTROL OF SNP GENOTYPE DATA IN B&H TOMATO (*SOLANUM LYCOPERSICUM* L.) GERMPLASM: HOW FILTERING THRESHOLDS AND FILTERING ORDER DETERMINE THE NUMBER OF USABLE MARKERS

Mirela Kajkut Zeljković¹, Mirjana Jankulovska²

¹Institute of Genetic Resources of University of Banja Luka, Banja Luka, Bosnia and Herzegovina

²Ss. Cyril and Methodius University in Skopje, Faculty of Agricultural Sciences and Food-Skopje, Skopje, Republic of North Macedonia

*corresponding author: mirela.kajkut@igr.unibl.org

ABSTRACT

Genotyping of gene bank collections is now routine, but the raw output of a genotyping run is never ready for analysis. Failed samples, poorly performing markers and uninformative loci must be removed first, and the choices made at this stage are rarely reported in full. This study quantifies, for a real tomato data set, how much of the final data set is determined by those choices. A collection of 81 tomato accessions from the Gene Bank of the Republic of Srpska, Bosnia and Herzegovina (B&H), was genotyped with a genome-wide SNP panel yielding 5,271 scorable markers. The overall genotype call rate was 69.2%, 13 accessions (16.0%) produced no calls at all, and 149 calls (0.03%) in 19 markers used non-standard genotype codes and were treated as missing. Three sets of missingness thresholds were applied in two different sequences: markers filtered before accessions, and accessions filtered before markers. The number of informative loci retained (minor allele frequency ≥ 0.05) ranged from zero to 1,110 across the six combinations. Under the strictest thresholds, filtering markers first removed the entire panel, because the completely failed accessions imposed a minimum missingness of 0.160 on every marker, while filtering accessions first retained 4,552 markers and 798 informative loci. Removing accessions without any calls as an explicit preliminary step largely eliminated the discrepancy between the two sequences. The results show that the order of filtering operations can matter as much as the threshold values themselves, and that both should be reported.

Keywords: SNP genotyping, quality control, missing data, filtering thresholds, marker retention, tomato, gene bank, reproducibility

INTRODUCTION

Single nucleotide polymorphism (SNP) genotyping has become the standard tool for characterizing crop germplasm. Dedicated panels are available for most major species, and in tomato (*Solanum lycopersicum* L.) genome-wide SNP sets are routinely used to describe collections of cultivars and landraces (Adedze et al., 2025; Xu et al., 2024). Gene banks increasingly genotype whole collections as part of ordinary collection management rather than as individual research projects (Anglin et al., 2025).

Genotyping a collection, however, does not by itself produce a data set that can be analyzed. The raw output of any platform contains failed samples, poorly performing markers, uninformative loci and missing calls, all of which must be removed before any statistic is

computed (Pavan et al., 2020). Quality control is therefore an analytical step with its own parameters, not a technical formality, and dedicated software has been written to standardize it (Gondro et al., 2014). The same recognition has emerged in animal breeding, where quality control of genomic data is now treated as a methodological topic in its own right (Lee et al., 2024).

Filtering decisions are not neutral. The minor allele frequency threshold alone can change the number of genetic clusters inferred from a data set and the apparent relationships among individuals (Linck & Battey, 2019). Broader reviews show that thresholds for missingness, depth and allele frequency interact, so that the same raw data can support different conclusions depending on the combination applied (O'Leary et al., 2018; Hemstrom et al., 2024). The effect extends beyond descriptive statistics: filtering alters which loci appear to be associated with environmental variables (Ahrens et al., 2021), and in small populations it biases estimates of gene flow, with no single parameter set performing best for every question (Nazareno & Knowles, 2021). Retaining more data is not automatically better, because permissive settings admit noise along with signal (Díaz-Arce & Rodríguez-Ezpeleta, 2019).

Despite this, no agreed standard exists for how filtering should be performed or described. Thresholds are often inherited from software defaults or from earlier publications rather than justified for the data set at hand, and the sequence in which they are applied is frequently omitted from the methods section (Hemstrom et al., 2024). This matters for reproducibility, which in genomics depends less on access to raw data than on a complete specification of the steps applied to it (Baykal et al., 2024). Two laboratories using identical threshold values can obtain different analysis-ready data sets if they apply them in a different order, and nothing in the published description would reveal the difference.

One aspect of this problem is particularly relevant to gene bank material and has received little attention. Collections held *ex situ* often yield DNA of uneven quality, and genotyping runs on such material routinely include samples that fail partially or completely (Anglin et al., 2025; Pavan et al., 2020). When a data set contains completely failed samples, every marker carries a floor level of missing data set solely by the number of those samples. If marker call rate is evaluated before the failed samples are removed, that floor is attributed to the markers, and strict marker thresholds may discard a large part of the panel for reasons unrelated to marker performance.

The present study examines this interaction using a single real data set of 81 tomato accessions from a Gene Bank of Republic of Srpska collection. The objectives were to document the genotyping performance of the panel, to quantify how the numbers of retained accessions, retained markers and informative loci respond to three sets of missingness thresholds applied in two different sequences, and to provide the analysis as a short reproducible script. No inference is drawn about the genetic structure or diversity of the collection; the aim is restricted to showing how much of an apparently fixed quantity, the number of usable markers, is in fact determined by processing decisions.

MATERIALS AND METHODS

Plant material and genotyping

Eighty one tomato accessions maintained at the Gene Bank of the Republic of Srpska, B&H, were used. The set comprises predominantly local landraces together with a small number of introduced and heirloom cultivars held in the same collection. Lyophilized young leaf material was submitted to an external service provider for DNA extraction and genotyping with a genome-wide tomato SNP panel. Genotype calls were returned as biallelic codes, with uncalled genotypes flagged by the provider. The returned data set comprised 5,271 scorable markers for the 81 accessions and formed the single input to all analyses reported here.

Definition of a valid call and of missingness

A genotype was treated as a valid call only when it consisted of exactly two single-nucleotide symbols (A, C, G or T) separated by a slash. All other entries were treated as missing. These included the provider missing codes, empty cells, and a small number of non-standard strings in which one or both sides consisted of two bases, such as TT/TT or TG/TG, or of a dash, such as -/-. Such entries numbered 149 (0.03% of all cells) and occurred in 19 of the 5,271 markers. They were not confined to specific accessions and were mixed with ordinary single-base calls within the same markers, so they were treated as unreliable rather than as a distinct class of variant. Missingness was calculated separately for each accession, as the proportion of markers without a valid call, and for each marker, as the proportion of accessions without a valid call. Call rate is reported throughout as one minus missingness.

Filtering scenarios

Three sets of missingness thresholds were defined, expressed as the maximum proportion of missing calls permitted for a marker and for an accession: lenient (0.40 for markers, 0.55 for accessions), moderate (0.20 and 0.30) and strict (0.10 and 0.20). Each set was applied in two sequences. In order A, markers were filtered first and accession missingness was then recalculated on the retained markers before the accession threshold was applied. In order B, accessions were filtered first and marker missingness was then recalculated on the retained accessions. The two sequences use identical threshold values and differ only in the order of operations.

For each of the six resulting data sets, loci were counted as polymorphic when exactly two alleles were observed among the retained accessions. Minor allele frequency was calculated from the observed allele counts, and loci were classified as polymorphic at a minor allele frequency of at least 0.01 and as informative at a minor allele frequency of at least 0.05. Minor allele frequency was used only as a filtering criterion. No allele frequencies at individual loci are reported.

Software

All calculations were performed in R version 4.3.1. Genotype data were imported with the readxl package, and the analytical framework was performed in adegenet (Jombart, 2008) and poppr (Kamvar et al., 2014, 2015) packages for codominant marker data. The complete script that reproduces both tables and the figure is provided as supplementary material.

RESULTS AND DISCUSSION

Genotyping performance

The overall genotype call rate across the 5,271 markers and 81 accessions was 69.2% (Table 1). This figure conceals a strongly bimodal distribution at the accession level (Figure 1a). Thirteen accessions, 16.0% of the collection, produced no valid calls at all, while the median call rate of the remaining 68 accessions was 0.939. Between these two groups lay a small number of accessions with intermediate quality, the poorest of which had a call rate of 0.230. Such a pattern, in which a minority of samples fails outright while most perform well, is characteristic of genotyping runs on gene bank material, where DNA is extracted from stored or regenerated plants of variable physiological condition (Anglin et al., 2025).

The marker-level distribution was different in shape (Figure 1b). Marker call rates were concentrated between 0.63 and 0.82, with a median of 0.728, and the highest marker call rate was 0.840. This ceiling is a direct consequence of the failed accessions: because 13 of 81 accessions carried no data, no marker could have a missingness below $13/81 = 0.160$. A further, much smaller loss arose from the 149 non-standard genotype codes, which affected 19 markers and reduced the overall call rate by 0.03 percentage points. Accordingly, not a single marker passed a missingness threshold of 0.10, 1,648 markers passed a threshold of 0.20, and 4,219

passed a threshold of 0.40 (Table 1). The apparent poor performance of the marker panel is therefore not a property of the markers but an arithmetic consequence of the sample failures.

Table 1. Genotyping performance of the SNP panel in tomato accessions from the Gene Bank of the Republic of Srpska

Metric	Value
Markers assayed	5,271
Accessions genotyped	81
Overall genotype call rate	69.2%
Calls excluded as non-standard genotype codes	149 (0.03%), in 19 markers
Accessions yielding no calls	13 (16.0%)
Accession call rate, median (IQR)	0.912 (0.428–0.967)
Accession call rate, range	0.000–0.978
Accession call rate excluding complete failures (n = 68), median (minimum)	0.939 (0.230)
Accessions with missingness ≤ 0.10	41
Accessions with missingness ≤ 0.20	48
Accessions with missingness ≤ 0.30	53
Accessions with missingness ≤ 0.55	60
Marker call rate, median (IQR)	0.728 (0.630–0.815)
Marker call rate, maximum	0.840
Markers with missingness ≤ 0.10	0
Markers with missingness ≤ 0.20	1,648
Markers with missingness ≤ 0.40	4,219

Effect of filtering thresholds

Applying the three threshold sets produced data sets differing by an order of magnitude in size (Table 2, Figure 1c). Under order B, in which accessions were filtered before markers, the number of informative loci was 1,110 under lenient thresholds, 791 under moderate thresholds and 798 under strict thresholds, while the number of retained accessions fell from 60 to 53 and then to 48. The marker count was comparatively stable in this sequence, between 4,552 and 4,942, because the accession filter removed the samples responsible for most of the missing data before marker quality was assessed.

Under order A, in which markers were filtered first, the same threshold values behaved very differently. The lenient setting retained 4,219 markers and 64 accessions, yielding 940 informative loci. The moderate setting retained only 1,648 markers, and the number of informative loci fell to 398, less than half the value obtained with the same thresholds in the other sequence. The strict setting removed every marker, and therefore the entire data set, for the reason given above.

These magnitudes are consistent with the published literature on filtering sensitivity. Linck and Battey (2019) showed that the minor allele frequency threshold alone can change the number of inferred genetic clusters, and O'Leary et al. (2018) and Hemstrom et al. (2024) report that combinations of missingness and frequency filters routinely alter the size and composition of the retained marker set by large factors. Nazareno and Knowles (2021) reached a similar conclusion for a small plant population, finding that no single filter setting was optimal across analyses. Díaz-Arce and Rodríguez-Ezpeleta (2019) likewise found that the most permissive parameter choice is not necessarily the most informative one. The present data add a plant gene bank example to this body of evidence and show that the effect is large enough to be decisive, rather than marginal.

Table 2. Numbers of markers, accessions and polymorphic loci retained under three sets of missingness thresholds applied in two filtering sequences

Filtering order	Scenario	SNPs retained	Accessions retained	Polymorphic loci (MAF $\geq$ 0.01)	Informative loci (MAF $\geq$ 0.05)
A: markers, then accessions	Lenient	4,219	64	2,148	940
B: accessions, then markers	Lenient	4,942	60	2,441	1,110
A: markers, then accessions	Moderate	1,648	67	886	398
B: accessions, then markers	Moderate	4,659	53	2,179	791
A: markers, then accessions	Strict	0	0	0	0
B: accessions, then markers	Strict	4,552	48	2,185	798

Effect of filtering order

The most consistent finding is that the two sequences did not give the same result at any threshold setting, although they used identical threshold values. The difference was modest under lenient thresholds (940 against 1,110 informative loci, a difference of 18%), substantial under moderate thresholds (398 against 791, a factor of two), and absolute under strict thresholds, where order A returned an empty data set and order B returned 798 informative loci.

The two sequences also differ in what they preserve. Order A retained more accessions under lenient and moderate thresholds (64 and 67, against 60 and 53), because recalculating accession missingness on an already cleaned marker set allows partially failed accessions to pass. Order B retained more markers and more informative loci. The choice between them is therefore a trade-off between sample size and marker number, and the appropriate choice depends on the intended analysis rather than on a general rule. This mirrors the conclusion of Nazareno and Knowles (2021) that filter settings should be matched to the question, and the broader argument of Hemstrom et al. (2024) that filtering parameters should be treated as analytical decisions requiring justification.

A simple preliminary step removes most of the discrepancy. When the 13 accessions without any valid calls were excluded before either threshold was applied, the strict setting under order A retained 2,152 markers, 64 accessions and 517 informative loci instead of nothing at all, and the moderate setting retained 3,436 markers, 60 accessions and 842 informative loci, slightly exceeding the value obtained under order B. Removing complete failures as an explicit first step is uncontroversial, costs no information, and reduces the dependence of the final data set on the order in which the remaining filters are applied.

The practical consequence for reporting is straightforward. A methods section that states only the threshold values is insufficient, because the same values produced between zero and 1,110 informative loci in this data set depending on the sequence used. The sequence of operations, the number of samples and markers removed at each step, the rule defining a valid genotype call, and the treatment of completely failed samples should all be stated explicitly. The last two points are not trivial. In the present data set a permissive definition of a valid call would silently have admitted 149 non-standard codes, and because those codes add further allele symbols at the sites where they occur, the 19 affected markers would then have been

excluded from the polymorphic counts as apparently multi-allelic. This requirement is consistent with current recommendations for reproducibility in genomic analyses, where the specification of processing steps is the limiting factor rather than access to the raw data (Baykal et al., 2024), and with the quality control protocols proposed for crop association studies (Pavan et al., 2020) and for genomic data in animal breeding (Gondro et al., 2014; Lee et al., 2024).

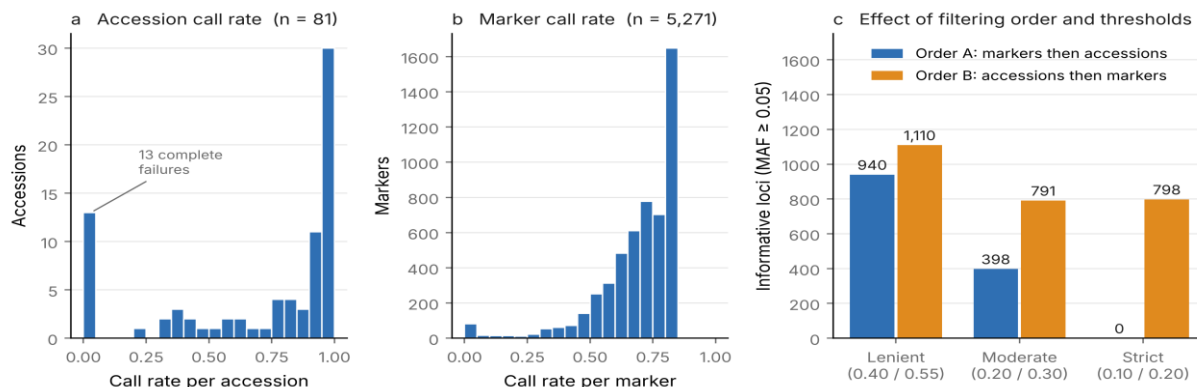


Figure 1. Quality control of the tomato SNP data set. (a) Distribution of call rate across the 81 accessions, (b) Distribution of call rate across the 5,271 markers, (c) Number of informative loci (minor allele frequency ≥ 0.05) retained under three threshold sets applied in two filtering sequences. Thresholds in parentheses are the maximum permitted missingness for markers and for accessions, respectively.

CONCLUSIONS

In a collection of 81 B&H tomato accessions genotyped with a genome-wide SNP panel, the number of usable markers was determined as much by the sequence of quality control operations as by the threshold values applied. Identical thresholds yielded between zero and 1,110 informative loci depending on whether markers or accessions were filtered first. The cause was a group of completely failed accessions, which imposed a fixed minimum level of missing data on every marker and so made strict marker filtering impossible until those accessions were removed. Excluding samples without any valid calls as an explicit preliminary step restored a usable data set under all threshold settings and reduced the difference between the two sequences.

Samples that produce no calls should be identified and removed before marker statistics are calculated. The rule defining a valid genotype call, the threshold values and the order in which they are applied should all be reported, together with the number of samples and markers removed at each step. Where the intended analysis is sensitive to marker number or to sample size, the effect of the chosen settings should be checked against at least one alternative, since the cost of doing so is negligible and the consequences of not doing so may be substantial.

ACKNOWLEDGMENTS

This research was financially supported by the Ministry of Scientific and Technological Development and Higher Education of the Republic of Srpska through the project “Molecular characterization of Tomato (*Solanum lycopersicum* L.) Germplasm in the Republic of Srpska”, grant no. 19.032/961-74/24.

REFERENCES

Adedze, Y. M. N., Xu, Y., Liu, S., Zhao, Y., Mo, C., Zhang, R., Dong, J., Lan, H., Huang, J., Chen, X., Gao, X., Yin, Q., & Zhang, J. (2025). An advanced 10K SNP panel for genotyping

- tomato (*Solanum lycopersicum* L.) via targeted genome sequencing. *Frontiers in Plant Science*, 16. <https://doi.org/10.3389/fpls.2025.1582241>
- Ahrens, C. W., Jordan, R., Bragg, J., Harrison, P. A., Hopley, T., Bothwell, H., Murray, K., Steane, D. A., Whale, J. W., Byrne, M., Andrew, R., & Rymer, P. D. (2021). Regarding the F-word: The effects of data filtering on inferred genotype-environment associations. *Molecular Ecology Resources*, 21(5), 1460–1474. <https://doi.org/10.1111/1755-0998.13351>
- Anglin, N. L., Wenzl, P., Azevedo, V., Lusty, C., Ellis, D., & Gao, D. (2025). Genotyping genebank collections: Strategic approaches and considerations for optimal collection management. *Plants*, 14(2), 252. <https://doi.org/10.3390/plants14020252>
- Baykal, P. I., Łabaj, P. P., Markowetz, F., Schriml, L. M., Stekhoven, D. J., Mangul, S., & Beerenwinkel, N. (2024). Genomic reproducibility in the bioinformatics era. *Genome Biology*, 25(1), 213. <https://doi.org/10.1186/s13059-024-03343-2>
- Díaz-Arce, N., & Rodríguez-Ezpeleta, N. (2019). Selecting RAD-Seq data analysis parameters for population genetics: The more the better? *Frontiers in Genetics*, 10, 533. <https://doi.org/10.3389/fgene.2019.00533>
- Gondro, C., Porto-Neto, L. R., & Lee, S. H. (2014). SNPQC - An R pipeline for quality control of Illumina SNP genotyping array data. *Animal Genetics*, 45(5), 758–761. <https://doi.org/10.1111/age.12198>
- Hemstrom, W., Grummer, J. A., Luikart, G., & Christie, M. R. (2024). Next-generation data filtering in the genomics era. *Nature Reviews Genetics*, 25(11), 750–767. <https://doi.org/10.1038/s41576-024-00738-6>
- Jombart, T. (2008). adegenet: A R package for the multivariate analysis of genetic markers. *Bioinformatics*, 24(11), 1403–1405. <https://doi.org/10.1093/bioinformatics/btn129>
- Kamvar, Z. N., Brooks, J. C., & Grünwald, N. J. (2015). Novel R tools for analysis of genome-wide population genetic data with emphasis on clonality. *Frontiers in Genetics*, 6, 208. <https://doi.org/10.3389/fgene.2015.00208>
- Kamvar, Z. N., Tabima, J. F., & Grünwald, N. J. (2014). Poppr: An R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ*, 2, e281. <https://doi.org/10.7717/peerj.281>
- Lee, J., Jung, J. H., & Oh, S.-H. (2024). Enhancing animal breeding through quality control in genomic data — A review. *Journal of Animal Science and Technology*, 66(6), 1099–1108. <https://doi.org/10.5187/jast.2024.e92>
- Linck, E., & Battey, C. J. (2019). Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. *Molecular Ecology Resources*. <https://doi.org/10.1111/1755-0998.12995>
- Nazareno, A. G., & Knowles, L. L. (2021). There is no 'rule of thumb': Genomic filter settings for a small plant population to obtain unbiased gene flow estimates. *Frontiers in Plant Science*, 12, 677009. <https://doi.org/10.3389/fpls.2021.677009>
- O'Leary, S. J., Puritz, J. B., Willis, S. C., Hollenbeck, C. M., & Portnoy, D. S. (2018). These aren't the loci you're looking for: Principles of effective SNP filtering for molecular ecologists. *Molecular Ecology*, 27(16), 3193–3206. <https://doi.org/10.1111/mec.14792>
- Pavan, S., Delvento, C., Ricciardi, L., Lotti, C., Ciani, E., & D'Agostino, N. (2020). Recommendations for choosing the genotyping method and best practices for quality control in crop genome-wide association studies. *Frontiers in Genetics*, 11, 447. <https://doi.org/10.3389/fgene.2020.00447>
- Xu, W., Gong, C., Mai, P., Li, Z., Sun, B., & Li, T. (2024). Genetic diversity and population structure analysis of 418 tomato cultivars based on single nucleotide polymorphism markers. *Frontiers in Plant Science*, 15. <https://doi.org/10.3389/fpls.2024.1445734>