nature plants

Letter

https://doi.org/10.1038/s41477-025-01979-x

Re-analysis of mobile mRNA datasets raises questions about the extent of long-distance mRNA communication

Received: 2 October 2024

Accepted: 10 March 2025

Published online: 16 April 2025

Check for updates

Pirita Paajanen[1,7] ✉, Melissa Tomkins[1,7], Franziska Hoerbst[1,7], Ruth Veevers[1], Michelle Heeney[2], Hannah Rae Thomas[3], Federico Apelt[4], Eleftheria Saplaoura[4], Saurabh Gupta[4,6], Margaret Frank[2], Dirk Walther[4], Christine Faulkner[3], Julia Kehr[5], Friedrich Kragler[4] & Richard J. Morris[1] ✉

Short-read RNA-seq studies of grafted plants have led to the proposal that thousands of messenger RNAs (mRNAs) move over long distances between plant tissues[1–7], potentially acting as signals[8–12]. Transport of mRNAs between cells and tissues has been shown to play a role in several physiological and developmental processes in plants, such as tuberization[13], leaf development[14] and meristem maintenance[15]; yet for most mobile mRNAs, the biological relevance of transport remains to be determined[16–19]. Here we perform a meta-analysis of existing mobile mRNA datasets and examine the associated bioinformatic pipelines. Taking technological noise, biological variation, potential contamination and incomplete genome assemblies into account, we find that a high percentage of currently annotated graft-mobile transcripts are left without statistical support from available RNA-seq data. This meta-analysis challenges the findings of previous studies and current views on mRNA communication.

A key step in mobile mRNA studies is the assignment of RNA-seq reads to different genotypes. One way of identifying the genotype is based on single nucleotide polymorphisms (SNPs) (Fig. 1). Typically, a requirement is made for a defined number of RNA-seq reads to have a SNP that corresponds to the alternative allele for a transcript to be assigned to a foreign genotype. Published criteria are: ≥1 RNA-seq read covering at least two SNPs[3], ≥2 reads[3], ≥3 reads[20] or >3 reads[2] covering a single SNP. When these criteria are met, the corresponding transcript is defined as mobile.

As previously reported[21], criteria based on absolute numbers of reads, such as those above, exhibit a read-depth dependency (Extended Data Fig. 1). This is a consequence of sequencing noise.

Illumina sequencing machines produce base-calling errors at a rate of ~0.1–1% per base[22,23]. Sequencing providers often provide a quality assurance, for instance, that 85% of the reads have a Phred quality score of at least Q30 (that is, a base-calling error of less than $10^{-3} = 0.1\%$). However, base-calling inaccuracies are not the only source of error. Before sequencing, reverse transcriptases can introduce base changes with an error rate of ~0.001–0.01%; the reverse transcription reaction error may exhibit a nucleotide bias, for instance, 'G' to 'A'[24,25], and a range of other artefacts[26]. On average, 6.4 ± 1.24% of sequences are mutated[22]. The average error rate of next-generation sequencing technologies has been estimated as 0.24 ± 0.06% per base[22,27], with RNA-seq errors tending to be higher[27].

[1]Computational and Systems Biology, John Innes Centre, Norwich, UK. [2]School of Integrative Plant Science, Cornell University, Ithaca, NY, USA. [3]Cell and Developmental Biology, John Innes Centre, Norwich, UK. [4]Department II, Max Planck Institute of Molecular Plant Physiology, Potsdam-Golm, Germany. [5]Department of Biology, Institute for Plant Sciences and Microbiology, University of Hamburg, Hamburg, Germany. [6]Present address: Curtin Medical School, Curtin Health Innovation Research Institute (CHIRI), Curtin University, Perth, Western Australia, Australia. [7]These authors contributed equally: Pirita Paajanen, Melissa Tomkins, Franziska Hoerbst. ✉e-mail: Pirita.Paajanen@jic.ac.uk; Richard.Morris@jic.ac.uk

https://doi.org/10.1038/s41477-025-01979-x

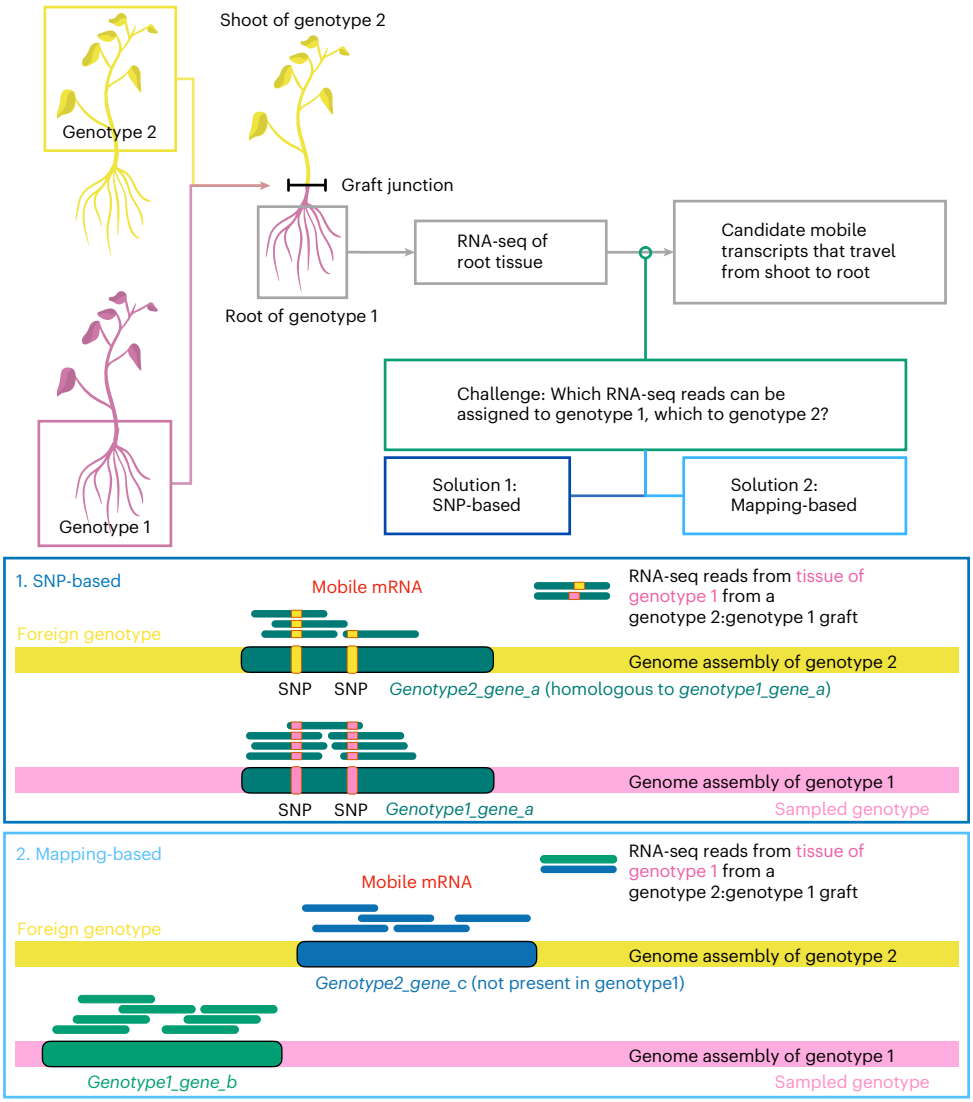


Fig. 1 | Grafting coupled with RNA-seq to identify transcripts that move from tissue of one genotype/species/ecotype/cultivar into tissue of another across the graft junction. Grafting-based strategy for identification of mRNAs that move from shoot (scion) to root (stock), from genotype 2 to genotype 1, using a scion:stock = genotype 2:genotype 1 heterograft. The same strategy can be used to identify transcripts that move from shoot to root from genotype 1 to genotype 2 using a genotype 1:genotype 2 graft. Transcripts that move root to shoot can be identified by analysing mRNAs in shoot tissue. Natural grafts, such as those established between the parasitic dodder plant and its host plants, can be used in place of artificial grafts. A key challenge in all such approaches is how to assign transcripts to each genotype; methods for doing so are based on (**1**) SNP identification or on (**2**) the alignment to different reference genomes. For grafts from the same species, or similar genotypes, SNPs can be used to distinguish between genotypes and thus identify the source genotype of each transcript (**1**). For grafts between different species, mapping (**2**) each RNA-seq read to the genome assemblies can be an effective method for determining which transcripts are specific to one species.

We therefore investigated whether noise in RNA-seq may influence the identification of mobile mRNAs. Figure 2a lists how many reported mobile mRNAs have numbers of reads with SNP occurrences that are consistent with an assumed error rate[21,28]. As an example, for an accuracy of SNP calling of 99.97% (that is, 0.03% sequencing noise, Phred score Q35, and an error probability for the alternative allele of ~0.01%), the evidence for 1,086 out of 2,006 (54%) and 384 out of 1,130 (34%) previously identified mobile mRNAs[2,3] is in line with what would be expected from sequencing noise (Fig. 2a).

One way to increase the accuracy of detecting foreign transcripts is to consider multiple SNPs per read. If SNPs are located closely together, then a single RNA-seq read may cover more than one SNP. Accounting for co-occurring SNPs on the same read leads to the multiplication of their probabilities, resulting in higher accuracy (less likely to occur by chance), less pronounced read-depth dependence than single SNP criteria (Extended Data Fig. 1) and greater confidence in these reads being from a foreign genotype. We therefore examined reads over co-occurring SNPs (Extended Data Fig. 3 and Supplementary Table 1). In the *Arabidopsis* homograft datasets[2], we found a total of 1,753,179 reads covering more than 1 SNP in the root and 1,977,539 in the shoot of Col-0; of these 1,675 (0.10%) and 1,797 (0.091%), respectively, had reads supporting the alternative allele for at least 1 but not all SNPs. These inconsistent calls are in line with the notion that sequencing noise may confound the identification of mobile mRNAs. We found 29 reads (1.6×10^{-3}%) in the root and 2 reads (1.0×10^{-4}%) in the shoot for which all SNPs supported the alternative allele. Interestingly in Ped-0 homograft data, the proportion of reads with full support for the alternative allele was significantly higher (0.038% in the root, 0.12% in the shoot). Investigating these co-occurring SNPs revealed another confounding factor in the identification of mobile mRNA; several loci showed apparent heterozygosity in the Ped-0 ecotype (Extended Data Fig. 4).

https://doi.org/10.1038/s41477-025-01979-x

a

	# reported	# consistent with: 0.01%	0.1%	1.0%
A. thaliana	2,006 (100%)	1,086 (54%)	1,455 (73%)	1,790 (90%)
V. girdiana	1,130 (100%)	384 (34%)	945 (84%)	1,069 (95%)

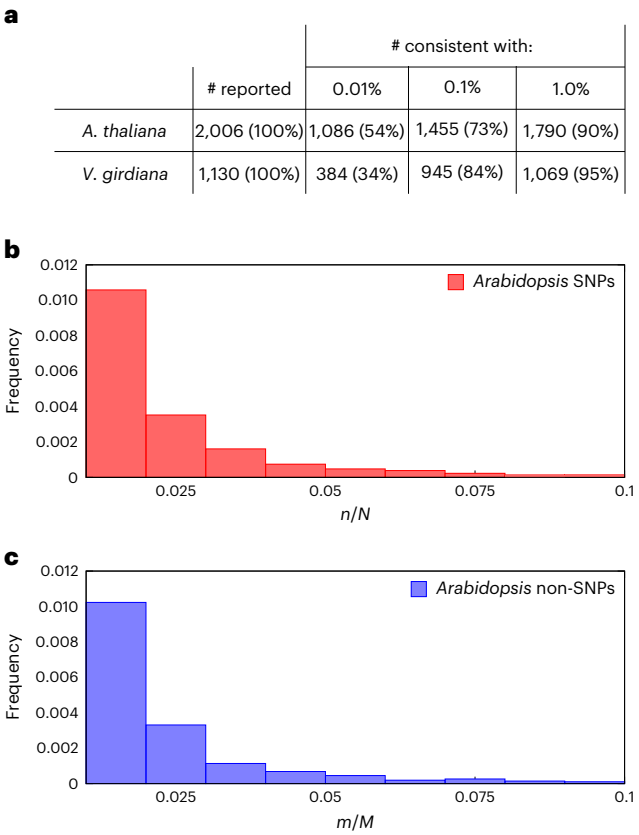


Fig. 2 | Alternative interpretations for the evidence for mobile mRNAs. **a**, Total numbers of reported mobile mRNAs in *Arabidopsis thaliana*[2] and *Vitis girdiana*[3] that can be explained by expected sequencing noise. Two values for the probability of the sequenced nucleotide at a SNP position being assigned to an alternative allele are given: 0.01% and 0.1%. **b**,**c**, The distributions of nucleotides at SNP and other positions ('non-SNP') can be informative for evaluating the evidence for the alternative allele. **b**, Histograms of the ratio of the number reads that match the alternative allele, *n*, over the number of reads of local and foreign reads, *N*, for each SNP position in the mobile population on examples from *Arabidopsis*[2]. Several SNPs have reads that match the alternative allele ($n/N > 0$). **c**, Histograms of the ratio of the number reads that match the second most frequent nucleotide, *m*, over the sum of the number of reads over the most frequent and second most frequent nucleotide, *M*, for neighbouring positions to SNPs. An exact two-sample Kolmogorov–Smirnov test does not find significant differences in the distributions over SNPs and other positions ($D = 0.089302$, $P = 0.3575$). A Welch two-sample *t*-test ($P = 0.6421$), Wilcox rank sum test ($P = 0.6388$) or an exact two-sample Kolmogorov–Smirnov test ($P = 0.4065$) does not support the values for SNPs being higher than other positions. **d**,**e**, Of the 2,006 previously identified mobile mRNAs[2], 953 unique mobile mRNAs were found in only 1 replicate in different organs of an adult Ped-0:Col-0 (root:shoot) graft. Such high numbers are not consistent with our hypothesis of sequencing noise and biological variation. Investigating the reciprocal relationship between root alleles that were detected in the rosette (**d**) and vice versa (**e**), in the root (1,373 mRNAs/867 unique) and rosette (577 mRNAs/151 unique) samples, identifies a strong linear correlation ($P = 2 \times 10^{-16}$) between expression in the source tissue and potential mobility. Interestingly, those SNPs that lie towards the lower read depth in each plot deviate the most from the linear relationship. However, these transcripts have low read numbers only in the 'source' tissue, whereas they have high read numbers in the sampled tissue and reads over SNPs that are consistent with sequencing errors. These plots thus show two effects: sequencing noise + either non-selective transport across the whole transcriptome or contamination.

Such apparent heterozygosity could be caused by a lack of introgression or gene copy-number variation; it has been estimated that 10% of the annotated genes in *Arabidopsis* have copy-number variation[29,30]. Differences in gene copy numbers can lead to reads not mapping correctly, which gives rise to pseudo-SNPs and pseudo-heterozygosity[29,30]. Of the 2,570 genes assigned as pseudo-heterozygous[30], we found 188 mobile transcripts[2] (Extended Data Figs. 2 and 4). We identified 19 transcripts in the Ped-0 samples that are likely caused by mismapping; interestingly, these include transcripts that frequently fulfill the criteria for being classified as mobile (Supplementary Table 3). Thus, in addition to technological noise, there are also biological causes that could be falsely interpreted as SNPs of an alternative allele. As a consequence, it becomes important to not rely solely on Phred scores for estimating errors in SNP assignments. We next sought to estimate this background noise level, that is, the frequency for finding the alternative allele when the alternative allele is not actually present. This value can be estimated from available *Arabidopsis* homograft data[2]. We counted the number of RNA-seq reads in the homograft with a SNP that matched the foreign genotype. For *Arabidopsis* homograft datasets (ecotypes Col-0 and Ped-0), these background noise levels were 0.084% (Col-0:Col-0 root), 0.082% (Col-0:Col-0 shoot), 0.68% (Ped-0:Ped-0 root) and 0.51% (Ped-0:Ped-0 shoot). The higher background error rate in Ped-0 is consistent with more Col-0 transcripts being identified as mobile in sampled Ped-0 tissue[2]. For an average background error rate of 0.34%, we find that over 1,455 out of 2,006 (>73%) and over 945 out of 1,130 (>84%) of annotated mobile mRNAs would not be distinguishable from expected errors (Fig. 2a). Consistent with this, poor overlap between experiments has been noted[18,31], orthologues in closely related species exhibit conflicting mobility, and reported low ratios of mobile to endogenous mRNAs[3,5,7] are in line with the level of noise.

Another way to distinguish noise from potential evidence for the alternative allele is to investigate the differences in nucleotide distributions at SNP positions compared to other positions in the sequence (non-SNP positions). If a second genotype were present, we would expect the distribution of nucleotides at any SNP position to be enriched in the nucleotide that supports the alternative

https://doi.org/10.1038/s41477-025-01979-x

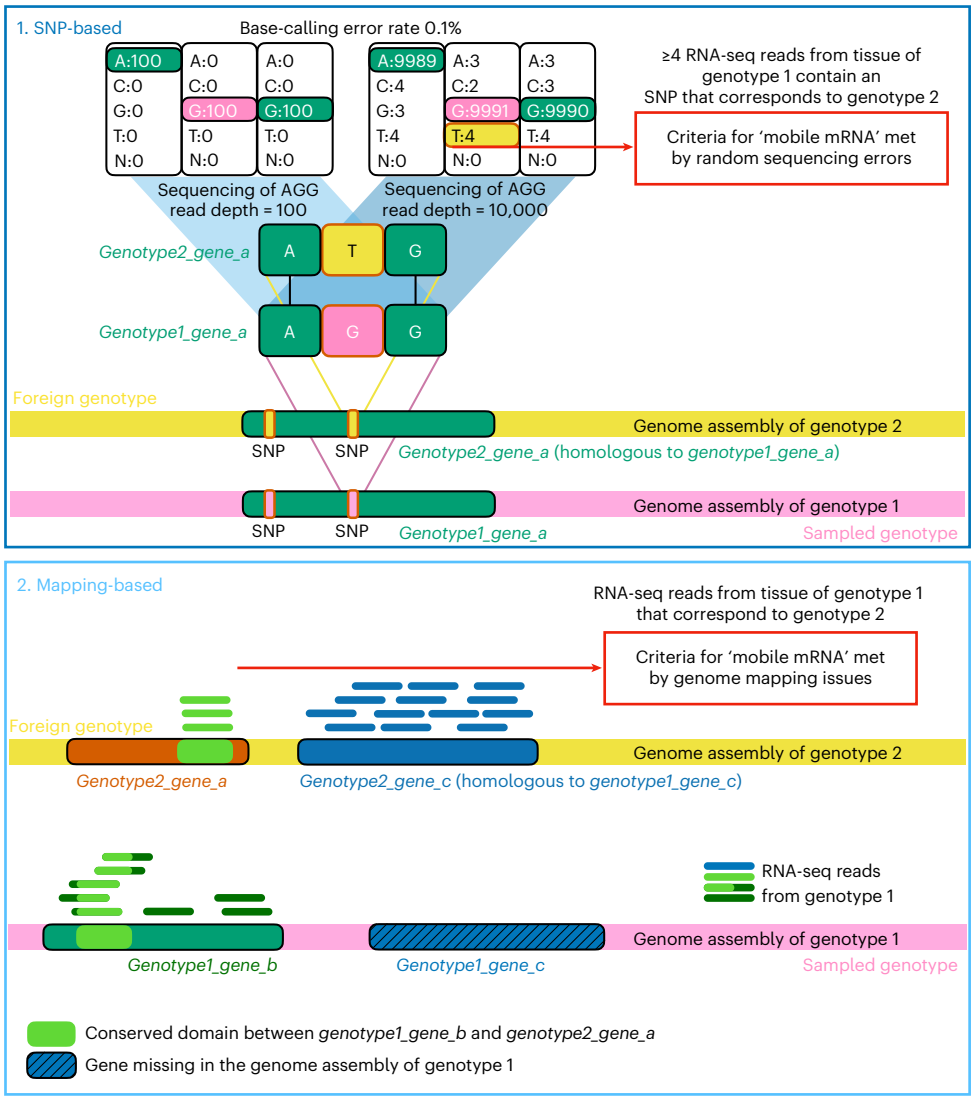


Fig. 3 | Mobile mRNA identification is not without challenges. (**1**) Technological noise can lead to challenges in the assignment of RNA-seq reads. In SNP-based methods of mobile mRNA detection, it is important to be able to differentiate between sequencing-associated errors and genuine SNPs. In the above case, an RNA-seq read with a 'T' at the SNP position would be indicative of the read having come from the alternative allele, genotype 2. However, every position has an error rate and the higher the read depth, the more incorrect base calls are to be expected. Base changes could arise for reverse transcriptase or amplification steps, although their error rate is typically orders of magnitude lower than sequencing errors. Conserved regions in gene families can give rise to similar challenges in distinguishing mapping ambiguities from genuine SNPs. Defining an mRNA as being mobile based on thresholds of RNA-seq reads that contain a SNP can result in base-calling errors and mapping ambiguities biasing the interpretation. To reduce the risk of such events occurring, further stringent filters can be applied (for instance, using only SNPs that are bi-allelic[2]) or applying rigorous statistical comparisons (for instance, estimating the allele calling frequencies and comparing them between homograft and heterograft[21]). (**2**) Genome complexity and genome quality can lead to mapping challenges. Orthologous sequences (light green) can result in some RNA-seq reads aligning to a different gene and different genotype. Genome assemblies that are not complete (telomere to telomere) from exactly the same genotype as used for grafting can result in potential mismappings. The shaded blue gene in genotype 1 is missing in the reference genome assembly, resulting in RNA-seq reads from this transcript being mapped to genotype 2.

allele. Furthermore, mRNAs that are transported to cells with low endogenous level (potential signals) would have a value of n/N close to 1, where n is the number of reads that match the alternative allele and N the total number of reads (endogenous + foreign). We investigated the distribution of n/N for each SNP in the mobile population of *Arabidopsis*[2]. While we do not find evidence for n/N values close to 1, there are non-zero values of n/N that seem to support the presence of the alternative allele (Fig. 2b). However, looking at all neighbouring positions of SNPs and computing the number of reads with the second most frequent nucleotide, m, over the sum of the most frequent and second most frequent nucleotides, M, we find no support for the SNP positions being different ($P = 0.3575$) (Fig. 2c). Thus, the expected shift in the distribution towards higher n/N values, that is $n/N > m/M$, is not observed. Given the low prevalence, it is important to note that this analysis does not exclude there being instances, potentially even thousands, of reads with SNPs associated with mobile mRNAs in the data, but if so we cannot distinguish them from noise.

Interestingly, two samples from *Arabidopsis*[2] do contain numbers of foreign reads that exceed expected noise levels. Investigating further, we find that these samples exhibit a strong linear correlation between the read counts of the grafted tissues (Fig. 2d,e). Similarly, *Arabidopsis* transcripts found in *Cuscuta pentagona* correlate with the expression levels in the host genotype[1]. Finding constant proportions of a whole transcriptome is indicative of contamination. Another explanation is that the whole transcriptome is transported, with detection

being proportional to read depth. Given the available data, we cannot distinguish between these possibilities.

Approaches that do not rely on SNPs, such as for cross-species studies, might avoid some of the above issues. A typical pipeline for analysing between-species grafts first maps reads to the reference genome of the sampled tissue (genotype 1 in Fig. 1). Unmapped reads are then compared to the reference genome of the potential source tissue (genotype 2 in Fig. 1). The success of this approach depends on the quality of the genome assembly. Supplementary Table 2 lists some genome completeness estimates for assemblies that were used in previous mobile mRNA studies. For instance, at the time of the study that investigated the movement of transcripts from a *Nicotiana benthamiana* scion to a *Solanum lycopersicum* (tomato) rootstock[6], ~15% of the genome was not yet assembled (Extended Data Fig. 6). The authors therefore collected RNA-seq data and applied stringent mapping criteria to mitigate effects of using an incomplete assembly. However, repeating their procedure, we found that many reads that did not map to the tomato genome all aligned to small regions of the *N. benthamiana* genome, and that coverage was highly uneven over exons (Extended Data Figs. 5 and 6). Furthermore, blasting the reads identified as being from *N. benthamiana* against the whole NCBI nucleotide database resulted in 100% matches to highly conserved sequences contained within many genomes, including *N. benthamiana* and other Solanaceae species, in particular to 18S ribosomal RNA genes, which accounted for 97.7% of the blast hits to *N. benthamiana* (Extended Data Fig. 8). To test for false negatives, we mapped the heterograft reads directly to the *N. benthamiana* genome and found 16 short transcripts that could not be distinguished between genomes (Supplementary Table 4).

In addition to genome assembly quality, read depth can also bias the interpretation of RNA-seq data from grafts between different species (Extended Data Fig. 7). For instance, ~30% of the *Arabidopsis thaliana* transcriptome was reported to move into *Cuscuta pentagona*, while only 9% of the tomato transcriptome moves to *Cuscuta*[1]. However, there is a large discrepancy in the amount of RNA-seq data between tomato (6 Mb) and *Arabidopsis* experiments (2 Gb). Greater coverage would be expected to lead to more transcripts being detected[32–34], thus explaining the reported bias in mobility between species.

Overall, our study raises questions about published numbers of mobile mRNAs. The experimental evidence for movement of a small number of mRNAs over long distances in plants is compelling[5,6,11,15,17,35,36]. However, on the basis of RNA-seq studies, several thousand mobile transcripts have been reported[1–4,6,7]. Here we question this extrapolation from tens of validated cases to the published vast numbers of potential long-distance signalling agents.

Recommendations

We described several challenges in identifying mobile mRNAs from short-read RNA-seq data (Fig. 3). While we do not present solutions, we suggest checks that can be performed to reduce the risk of false positives. We thus end with a list of recommendations. We assume that experimental issues have been taken care of, such as checking the samples for cross-contamination, verifying that graft junctions form functional vascular connections, and every effort has been made to use high-quality genome assemblies.

1. SNP reliability. A genome mapping visualization tool such as IGV[37] can be used to check for pseudo-heterozygosity and contamination in the samples. Observing the distribution of nucleotides at potential SNP positions and comparing to other positions can provide confidence in the SNPs and the alternative allele calls. These distributions should be compared to those from homograft data.
2. Co-occurring SNPs. RNA-seq reads that cover multiple SNPs can be used to check whether the SNPs that are associated with a certain genotype co-occur in such reads. Long-read and direct RNA sequencing have higher error rates but would allow the full transcript with all SNPs to be assessed. Sequencing protocols that barcode individual molecules by using adapters with unique molecular identifiers (UMIs) can be used to determine the error rates and check whether all reads from the same molecule are consistent in terms of their genotype assignment.
3. Accuracy of experimental and computational procedures for identifying foreign RNA-seq reads. Calculating the ratio of the number of RNA-seq reads assigned to an alternative allele (foreign reads) over the total number of mapped RNA-seq reads for an experiment (foreign + endogenous reads) is a useful metric. This value should be computed for homografts and compared to the value calculated from heterograft data.
4. Reproducibility and consistency of putative mobile transcripts. Independent biological replicates should be used to characterize the inherent variability in the identification of candidate mobile transcripts. Reciprocal grafting is recommended to evaluate whether mobile mRNA and their orthologues are consistently mobile (if mobility motifs are inherent to transcripts, then near-identical sequences would be expected to also be mobile) and, if not, potentially pinpoint determinants of mobility.
5. Alternative hypotheses. Definitions for mobile mRNAs using non-validated criteria are best avoided. It is important to test different hypotheses (for example, SNP vs sequencing noise; read from a foreign genotype vs mapping error; transport vs contamination; signalling molecules vs leftovers from differentiating cells) to explain the data. The plausibility of associated mechanisms can lend weight to different hypotheses.

Methods

All code and scripts are freely available from our GitHub repository at https://github.com/mtomtom/reanalysis-mobile-mrna/tree/main (ref. 38).

RNA-seq data processing

The raw reads were mapped to the references using hisat2 (v.2.1.0)[39],

```
hisat2 -x genome -1 read1 -2 read2 > mapping.sam
```

and processed using samtools (v1.9)[40],

```
samtools sort -o mapping.bam mapping.sam
samtools index mapping.bam
```

Expression level quantification

The expression levels were quantified with Stringtie (v.1.3.5)[41] using

```
stringtie mapping.bam -e -G genes.gff -o output.gtf
-A output.abundance.txt
```

Quantification of raw counts of all nucleotides

The raw counts were quantified with bcftools (v.1.10.2)[40] using

```
bcftools mpileup -A -q 0 -Q 0 -B -d 500000
--annotate FORMAT/AD, FORMAT/ADF, FORMAT/ADR,
FORMAT/DP, FORMAT/SP, INFO/AD,
INFO/ADF, INFO/ADR
```

These flags were chosen to compare the raw error rates between the homograft and heterograft to catch all nucleotides. Note that the bcftools mpileup default sequencing depth is 8,000, but the most highly expressed genes have up to 200,000 reads covering a locus within the datasets we considered.

Blast search

The NCBI nucleotide database was downloaded on 21 October 2022 and blast+ (v.2.9.0)[42] was utilized for alignments using

```
blastn -db nt -query unmapped.fasta -max_target_seqs 10
-max_hsps 1-evalue 1e-25
-outfmt '6 qseqid sseqid pident evalue staxids sscinames
scomnames sskingdoms stitle'
```

Estimating the accuracy of mobile mRNA detection

If we are only interested in the number of reads that contain a SNP that corresponds to the alternative allele, we can use a binomial distribution (q is the probability the SNP matches the alternative allele, $1-q$ is the probability that it does not) to evaluate the probability of this event occurring by chance[21]. The probabilities of errors occurring by chance were calculated from a standard cumulative binomial distribution, $P(k \geq m|N) = 1 - P(k < m|N)$, which accounts for the requirement of having k reads, where k is at least m, out of N. Considering replicates can be handled in the same way (the probability of each SNP is computed from the cumulative binomial function and the requirement for a defined number of replicates can likewise be computed from a cumulative binomial function). Multiple SNPs per read results in a multiplication of probabilities. Cumulative binomial function values were computed using standard available functions in Python and R.

Assessing how many SNPs can be explained by sequencing-associated errors

Rather than 'defining' a transcript as mobile, we evaluated the probability of the data being consistent with expected noise against the probability of the data being best explained by the presence of two genotypes (and therefore potential candidates for mobile transcripts)[21]. Essentially, this means that if we find 10 out of 100 reads that match the alternative allele, we compute how likely this would occur by chance for a defined error rate. The implicit but rarely checked assumption in all SNP-based mobile mRNA detection pipelines is that the occurrence of reads that support the alternative allele in the heterograft data is larger than in homograft data. The uncertainty in the inferred error rate depends on the amount of data. We capture this uncertainty through probability distributions to inform inferences drawn from the data[21]. This ratio of the statistical evidence of one hypothesis over another is known as the Bayes factor[43]. The classifications in Fig. 2a are based on the commonly used value of log Bayes factor greater than 1 (refs. 21,43). The statistical comparison of error rates was performed using baymobil[28].

Statistics for comparing nucleotide distribution as SNP positions vs other positions

To compare the full distributions of n/N and m/M values for different positions of RNA-seq reads, we used an exact two-sample Kolmogorov–Smirnov test, ks.test, available in R[44]. To evaluate whether the data supported the SNP distributions having higher values of n/N than other positions (m/M), we used an asymptotic two-sample Kolmogorov–Smirnov test. These tests were carried out for histograms with 100 bins.

Pseudo-heterozygosity

We downloaded the pseudo-heterozygous data from https://zenodo.org/records/6025134 (ref. 30). From the vcf-file we extracted all heterozygous calls for accession 9947 (Ped-0) and obtained 6,303 heterozygous SNPs. We compared these SNPs against the MATRIX_GWAS_raw_position.txt (from https://doi.org/10.5281/zenodo.5702395). We intersected these potential duplicate genes with the list of mobile genes[2] and found 19 duplicate genes. These are given in Supplementary Table 3.

Genome assembly completeness estimation

We downloaded all the assemblies mentioned in the original papers and estimated their completeness with Abyss (v.1.9.0) using the command 'abyss-fac'.

Contamination analysis

We analysed the samples of the root (1,373 mRNAs/867 unique) and rosette (577 mRNAs/151 unique), and reciprocally inspected the relationship between root alleles that were detected in the rosette and vice versa. We took the raw sequencing depth for 48,934 previously identified SNPs. For each SNP, we plotted the number of reads with a rosette allele (Col-0) found in the root sample (Ped-0) against the number of reads with the same SNP in the rosette sample. Similarly, we plotted the number of reads with the root allele (Ped-0) found in the rosette sample (Col) against the number of reads with the endogenous SNP (Ped-0) in the root sample. The linear fit was performed within gnuplot[45].

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

We used the following published datasets and the archived reads from NCBI: *Cuscuta pentagona*[1] (PRJNA257158; this dataset was incomplete and partly corrupt); *Vitis vinifera*[3] (SRP058158 and SRP058157); *Solanum lycopersicum, Nicotiana benthamiana*[6] (SRP111187); *Arabidopsis thaliana*[2] (PRJNA271927). We used deposited supplementary datasets of the associated publications to obtain the numbers of identified mRNAs. For each of the graft studies, we downloaded the reference genome sequence that matched the one that was used in the original paper with the same annotations; most are publicly available in Ensembl plants[46].

Code availability

We used largely available software packages as stated in the Methods. All code and scripts are freely available on GitHub at https://github.com/mtomtom/reanalysis-mobile-mrna/tree/main (ref. 38).

References

1. Kim, G., LeBlanc, M. L., Wafula, E. K., dePamphilis, C. W. & Westwood, J. H. Genomic-scale exchange of mRNA between a parasitic plant and its hosts. *Science* **345**, 808–811 (2014).
2. Thieme, C. J. et al. Endogenous *Arabidopsis* messenger RNAs transported to distant tissues. *Nat. Plants* **1**, 15025 (2015).
3. Yang, Y. et al. Messenger RNA exchange between scions and rootstocks in grafted grapevines. *BMC Plant Biol.* **15**, 251 (2015).
4. Wang, Y. et al. A universal pipeline for mobile mRNA detection and insights into heterografting advantages under chilling stress. *Hortic. Res.* **7**, 13 (2020).
5. Notaguchi, M., Higashiyama, T. & Suzuki, T. Identification of mRNAs that move over long distances using an RNA-seq analysis of *Arabidopsis*/*Nicotiana benthamiana* heterografts. *Plant Cell Physiol.* **56**, 311–321 (2014).
6. Xia, C. et al. Elucidation of the mechanisms of long-distance mRNA movement in a *Nicotiana benthamiana*/tomato heterograft system. *Plant Physiol.* **177**, 745–758 (2018).
7. Zhang, Z. et al. Vascular-mediated signalling involved in early phosphate stress response in plants. *Nat. Plants* **2**, 16033 (2016).
8. Lucas, W. J., Yoo, B.-C. & Kragler, F. RNA as a long-distance information macromolecule in plants. *Nat. Rev. Mol. Cell Biol.* **2**, 849–857 (2001).
9. Jorgensen, R. A., Atkinson, R. G., Forster, R. L. S. & Lucas, W. J. An RNA-based information superhighway in plants. *Science* **279**, 1486–1487 (1998).

10. Spiegelman, Z., Golan, G. & Wolf, S. Don't kill the messenger: long-distance trafficking of mRNA molecules. *Plant Sci.* **213**, 1–8 (2013).
11. Winter, N. & Kragler, F. Conceptual and methodological considerations on mRNA and proteins as intercellular and long-distance signals. *Plant Cell Physiol.* **59**, 1700–1713 (2018).
12. Ham, B.-K. & Lucas, W. J. Phloem-mobile RNAs as systemic signaling agents. *Annu. Rev. Plant Biol.* **68**, 173–195 (2017).
13. Hannapel, D. J. & Banerjee, A. K. Multiple mobile mRNA signals regulate tuber development in potato. *Plants* **6**, 8 (2017).
14. Kim, M., Canio, W., Kessler, S. & Sinha, N. Developmental changes due to long-distance movement of a homeobox fusion transcript in tomato. *Science* **293**, 287–289 (2001).
15. Kitagawa, M., Wu, P., Balkunde, R., Cunniff, P. & Jackson, D. An RNA exosome subunit mediates cell-to-cell trafficking of a homeobox mRNA via plasmodesmata. *Science* **375**, 177–182 (2022).
16. Notaguchi, M. Identification of phloem-mobile mRNA. *J. Plant Res.* **128**, 27–35 (2015).
17. Kehr, J. & Kragler, F. Long distance RNA movement. *New Phytol.* **218**, 29–40 (2018).
18. Kehr, J., Morris, R. J. & Kragler, F. Long-distance transported RNAs: from identity to function. *Annu. Rev. Plant Biol.* **73**, 457–474 (2022).
19. Heeney, M. & Frank, M. H. The mRNA mobileome: challenges and opportunities for deciphering signals from the noise. *Plant Cell* **35**, 1817–1833 (2023).
20. Wang, T. et al. Movement of ACC oxidase 3 mRNA from seeds to flesh promotes fruit ripening in apple. *Mol. Plant* **17**, 1221–1235 (2024).
21. Tomkins, M. et al. Exact Bayesian inference for the detection of graft-mobile transcripts from sequencing data. *J. R. Soc. Interface* **19**, 20220644 (2022).
22. Pfeiffer, F. et al. Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Sci. Rep.* **8**, 10950 (2018).
23. Loman, N. J. et al. Performance comparison of benchtop high-throughput sequencing platforms. *Nat. Biotechnol.* **30**, 434–439 (2012).
24. Fungtammasan, A. et al. Reverse transcription errors and RNA–DNA differences at short tandem repeats. *Mol. Biol. Evol.* **33**, 2744–2758 (2016).
25. Li, W. & Lynch, M. Universally high transcript error rates in bacteria. *eLife* **9**, e54898 (2020).
26. Verwilt, J., Mestdagh, P. & Vandesompele, J. Artifacts and biases of the reverse transcription reaction in RNA sequencing. *RNA* **29**, 889–897 (2023).
27. Stoler, N. & Nekrutenko, A. Sequencing error profiles of Illumina sequencing instruments. *NAR Genom. Bioinform.* **3**, lqab019 (2021).
28. Hoerbst, F., Morris, R. J. & Tomkins, M. baymobil: a Python package for detection of graft-mobile mRNA using exact Bayesian inference on RNA-seq data. Preprint at *Res. Square* https://doi.org/10.21203/rs.3.rs-2520491/v1 (2023).
29. Zapata, L. et al. Chromosome-level assembly of *Arabidopsis thaliana* Ler reveals the extent of translocation and inversion polymorphisms. *Proc. Natl Acad. Sci. USA* **113**, E4052–E4060 (2016).
30. Jaegle, B. et al. Extensive sequence duplication in *Arabidopsis* revealed by pseudo-heterozygosity. *Genome Biol.* **24**, 44 (2023).
31. Morris, R. J. On the selectivity, specificity and signalling potential of the long-distance movement of messenger RNA. *Curr. Opin. Plant Biol.* **43**, 1–7 (2018).
32. Wang, Z., Gerstein, M. & Snyder, M. RNA-seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2009).
33. Conesa, A. et al. A survey of best practices for RNA-seq data analysis. *Genome Biol.* **17**, 13 (2016).
34. Stark, R., Grzelak, M. & Hadfield, J. RNA sequencing: the teenage years. *Nat. Rev. Genet.* **20**, 631–656 (2019).
35. Zhang, W. et al. tRNA-related sequences trigger systemic mRNA transport in plants. *Plant Cell* **28**, 1237–1249 (2016).
36. Yang, L., Machin, F., Wang, S., Saplaoura, E. & Kragler, F. Heritable transgene-free genome editing in plants by grafting of wild-type shoots to transgenic donor rootstocks. *Nat. Biotechnol.* **41**, 958–967 (2023).
37. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
38. mtomtom. mtomtom/reanalysis-mobile-mrna: Re-analysis of mobile mRNA datasets raises questions about the extent of long-distance mRNA communication. *Zenodo* https://zenodo.org/records/15150276 (2025).
39. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
40. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
41. Pertea, M., Kim, D., Pertea, G. M., Leek, J. T. & Salzberg, S. L. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc.* **11**, 1650–1667 (2016).
42. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
43. Jaynes, E. T. *Probability Theory: the Logic of Science* (Cambridge Univ. Press, 2003).
44. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2021); https://www.R-project.org/
45. Williams, T. et al. Gnuplot 4.6: an interactive plotting program. http://gnuplot.sourceforge.net/ (2013).
46. Yates, A. et al. Ensembl Genomes 2022: an expanding genome resource for non-vertebrates. *Nucleic Acids Res.* **50**, D996–D1003 (2021).

Acknowledgements

We thank D. Staiger (Bielefeld University), W. Haerty (Earlham Institute), C. Dean (JIC), K. Schneeberger (LMU Munich), M. Mayer (QIB), C. Abreu-Goodger (University of Edinburgh), B. Zagrovic and A. Polyansky (Max Perutz Labs, Vienna) for discussions, insightful comments and constructive feedback on previous versions of the manuscript. The presented reanalysis and the insights derived from it would thus not have been possible without the availability of raw data; we thank all authors who deposited their data, meta-data and methods in public repositories. R.J.M. gratefully acknowledges support from the Biotechnology and Biological Science Research Council Institute Strategic Programme 'Building Resilience in Crops' (BB/X01102X/1). M.F. and M.H. acknowledge support from National Science Foundation Grants DGE-2139899, DBI-2019674 and IOS-1942437. C.F. and H.R.T. acknowledge support from Biotechnology and Biological Science Research Council Grants BB/X010996/1, BB/X007685/1, BB/X016056/1 and BB/Y008782/1. J.K. acknowledges support from the Deutsche Forschungsgemeinschaft (DFG; Project No. 433194101, Research Unit 5116). This Article is part of a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 810131) to F.K., J.K. and R.J.M.

Author contributions

P.P., M.T., F.H., C.F., J.K., F.K. and R.J.M. conceptualized the project. P.P., M.T., F.H., R.V. and R.J.M. designed the methodology. P.P. and M.T. conducted investigation, and with R.J.M. performed formal analysis. P.P., F.H. and R.J.M. performed visualization. M.T. and F.H. designed

software. M.F., F.K. and R.J.M. supervised the project. J.K., F.K. and R.J.M. acquired funding. P.P. and R.J.M. wrote the original draft, and all authors reviewed and edited the manuscript.

Competing interests
The authors declare no competing interests.

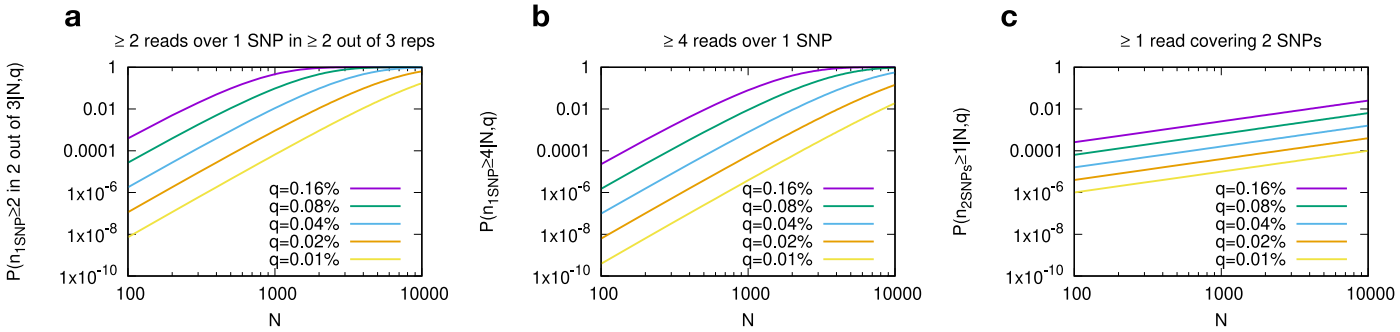
Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41477-025-01979-x.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41477-025-01979-x.

Correspondence and requests for materials should be addressed to Pirita Paajanen or Richard J. Morris.

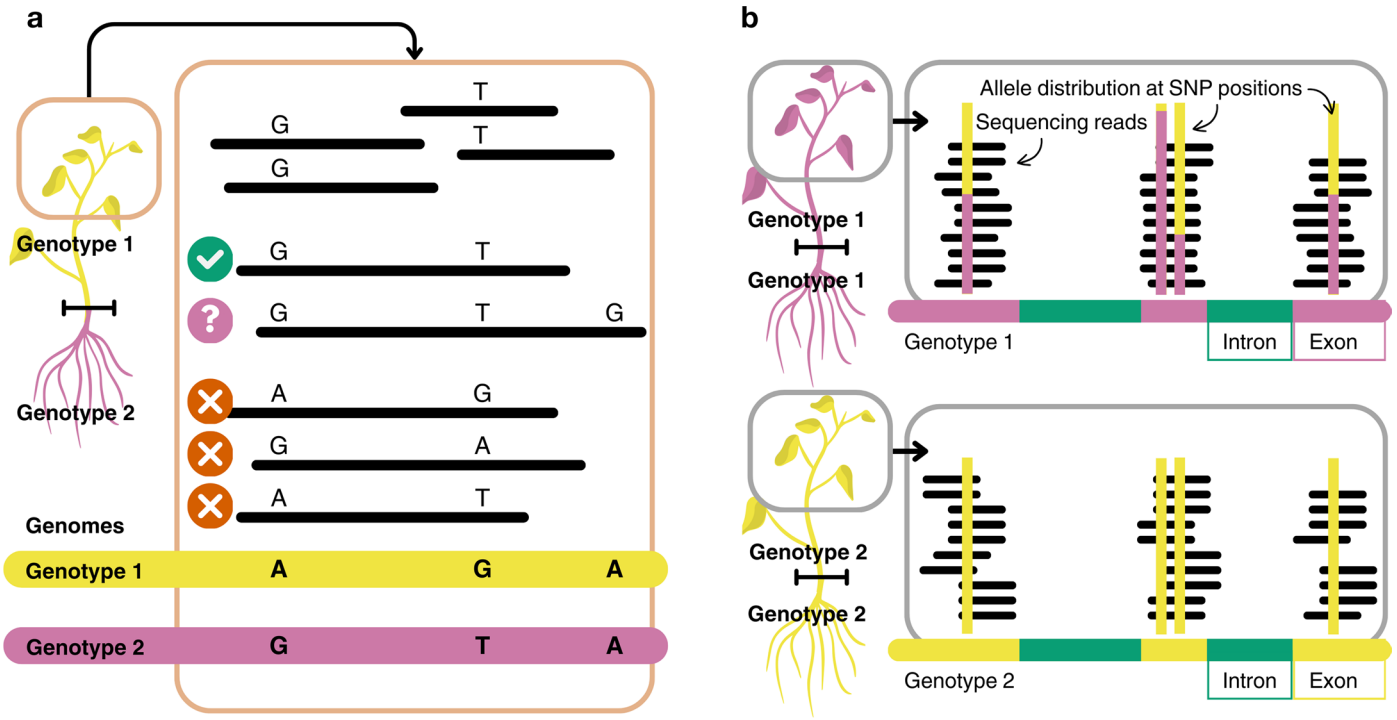
Peer review information *Nature Plants* thanks Marco Catoni, Cankui Zhang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

https://doi.org/10.1038/s41477-025-01979-x



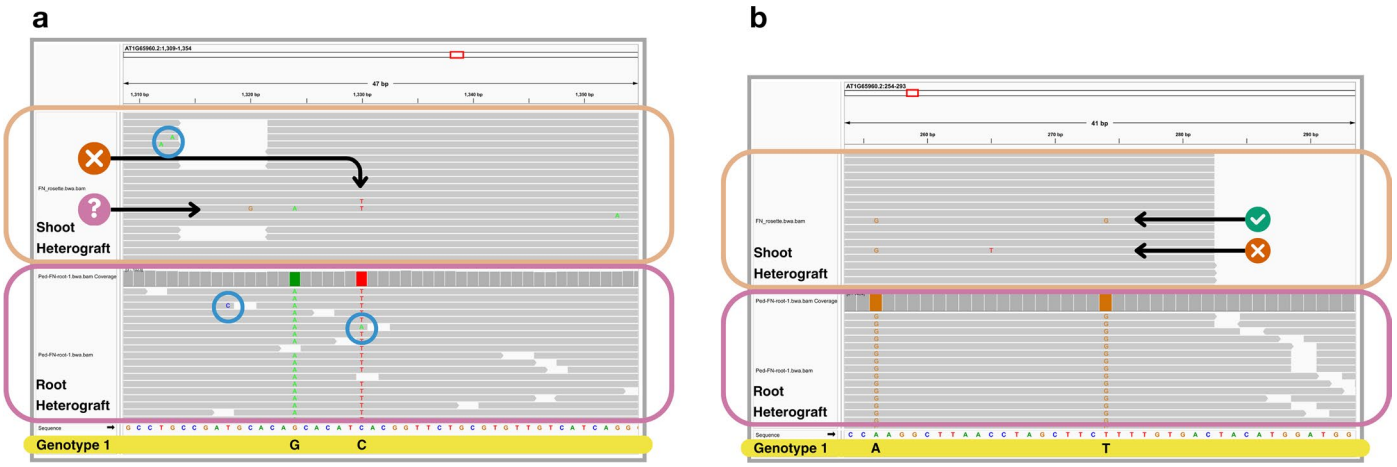
Extended Data Fig. 1 | Published criteria for defining mobile mRNAs based on absolute read counts suffer from read-depth dependencies. These plots show the probabilities of transcripts being defined as mobile by chance. Three different mobile mRNA definitions (**a, b, c**) and their dependence on read-depth (N) and on the rate of a SNP matching to the alternate allele (q) are depicted. The number of read counts over one SNP that correspond to the alternate allele is denoted by n1SNP, over two SNPs by n2SNPs. The probabilities were calculated using a cumulative binomial distribution, that is we account only for the nucleotides that correspond to the two alleles of interest. Note that both axes are on a log-scale. The requirement for co-occurring SNPs on one read (**c**) is more stringent and less likely to occur by chance at higher read-depths. For low values of q, these criteria are robust up to moderately high (several hundred) read-depths and would be unlikely to occur by chance.

https://doi.org/10.1038/s41477-025-01979-x



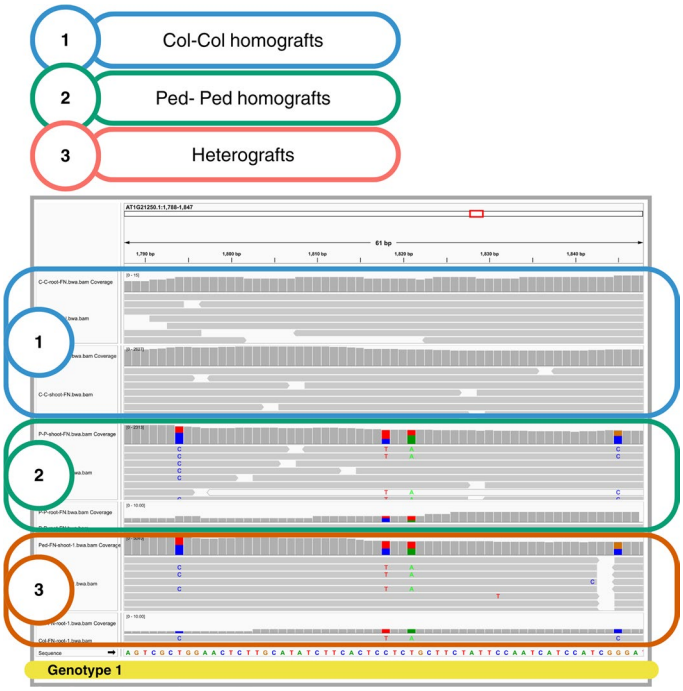
Extended Data Fig. 2 | Allelic differences in multiple SNPs per read and the appearance of heterozygosity (in homozygous species) can be used to check the viability of SNPs and exclude potentially problematic transcripts from the analysis. a, SNPs can be in close proximity, and therefore it can happen that several SNPs are recorded in the same RNA-Seq read. In this example, genotype 1 has three SNPs very close to each other: A, G and A (yellow bar). In genotype 2, we find G, T, A (magenta bar) in those positions. In this schematic example, reads from the shoot of Genotype 1 are mapped to Genotype 2. If all covered loci carry the allele of Genotype 2, we are observing evidence for the read being from Genotype 2 and the associated transcript being potentially mobile. On the other hand, if only one loci carries the allele of Genotype 2, the outcome is inconclusive, as it may indicate sequencing errors. **b**, A. thaliana is a selfing species, so we expect homozygocity at all positions for all reads mapping to the genome at all positions. However, for duplicated genes (magenta) in Genotype 1, which may be single copy genes in Genotype 2 (yellow), short read sequencing and mapping to Genotype 2, can give rise to what appears to be heterozygocity. When there are two alleles present in the homograft data (magenta and yellow), we may be observing pseudo-heterozygocity. See also Extended Data Figure 3.

https://doi.org/10.1038/s41477-025-01979-x



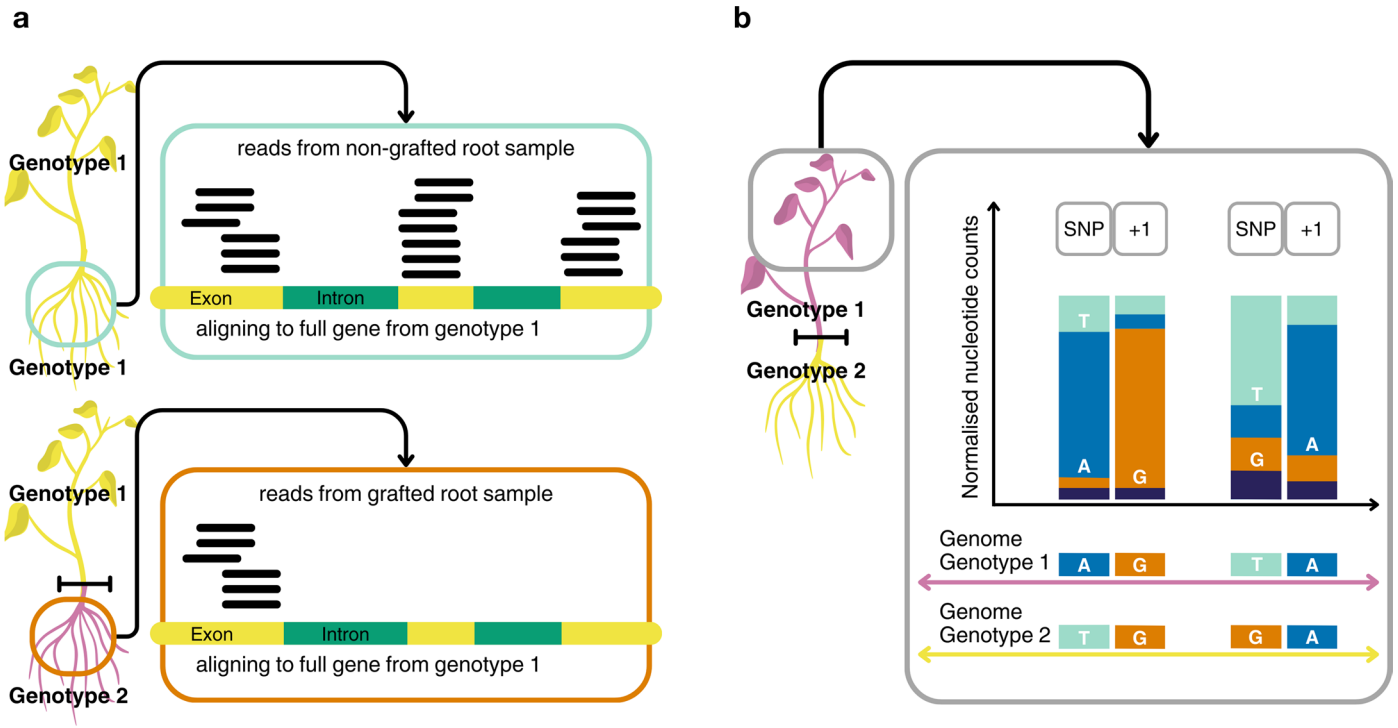
Extended Data Fig. 3 | Examples of co-occurring SNPs. a, There are two SNPs, G and C in Genotype 1 (Col) and A and T in Genotype 2 (Ped). These are likely two sequencing errors in the root sample, a C at a non-SNP position and an A at a SNP position (both highlighted in blue circles). In the shoot sample we see potential evidence for mobility at the SNP level but in one case the second SNP is not present and in the other case another sequencing error has occurred (G). Three further sequencing errors (two As on the top left, one A on the right) are also present in the shoot. **b**, This example shows two positions, A and T in Genotype 1 (Col) and G and G in Genotype 2 (Ped), for which some reads support the alternate allele (green tick), whereas others are likely sequencing errors (red cross). In the latter case, one G is in the correct position but the other G is not present and a further mismatch (T) has occurred. See Figure 2 for further explanations. The images are annotated screenshots taken in IGV[37]. Data taken from[2].

https://doi.org/10.1038/s41477-025-01979-x



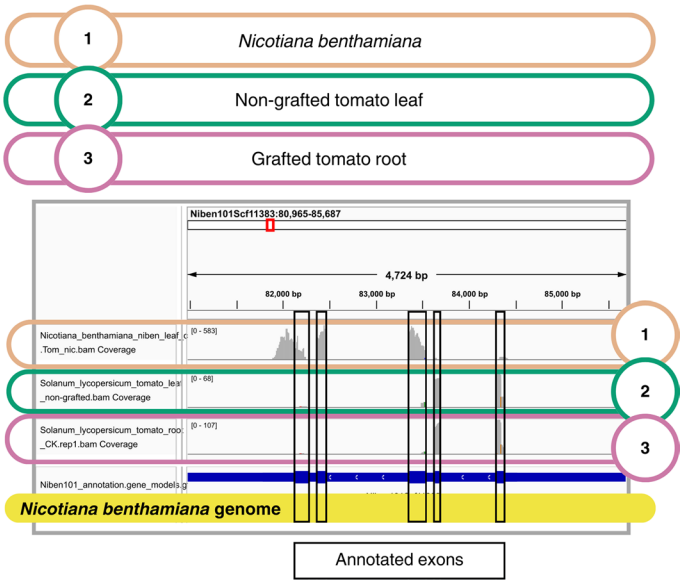
Extended Data Fig. 4 | A number of genes in Ped-0 show apparent heterozygosity, both in the homograft (2) and heterograft (3) datasets. At the highlighted positions there are distinct populations of alleles (depicted as red/blue and red/green bars). This is possibly due to the gene being duplicated in Ped, resulting in pseudoheterozygocity.

https://doi.org/10.1038/s41477-025-01979-x



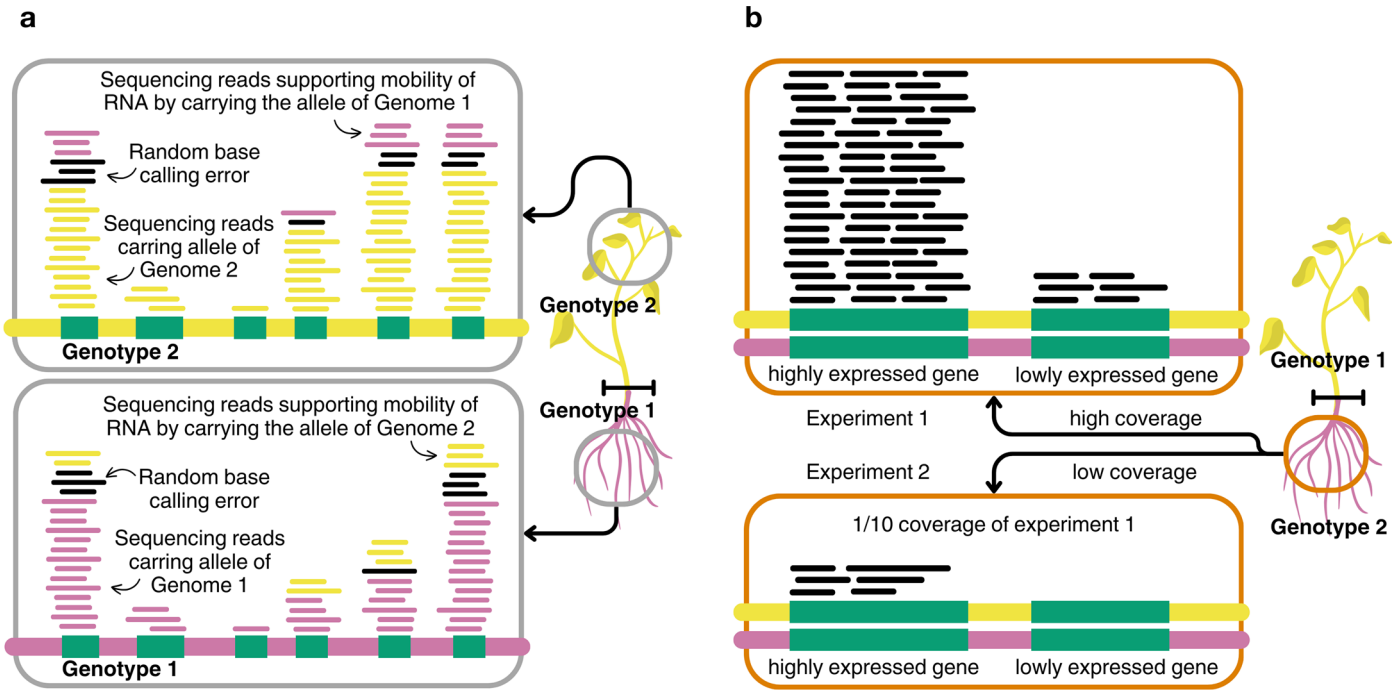
Extended Data Fig. 5 | Full-length transcript coverage and differences in the distribution of nucleotides between SNPs and other positions enhance the evidence for the presence of a foreign transcript in the sampled tissue. **a**, Sequenced transcripts would ideally have RNA-Seq reads covering most of the sequence, that is that all exons of the mRNA are approximately equally covered by sequenc- ing reads (top left). Reads covering all exons in the sample from Genotype 2 provide support for the whole transcript having moved from Genotype 1 to Genotype 2 across the graft junction. Transcripts with coverage only for a subsequence (bottom left) do not support full-length presence of the for- eign transcript. **b**, Neighbouring positions to SNPs can be used as a negative control to evaluate the strength of the signal at SNP positions. Shown here are neighbouring positions of the identified SNPs at the next nucleotide (SNP position +1). If the neighbouring position shows similar levels of alternative nucleotides as the SNP position, the these are likely sequencing errors, rather than evidence for the alternate allele. If the SNP positions have a different frequency of Genotype 2 allele than the neighbouring position has errors, then there is evidence for the alternate allele. Analysing the fre- quencies of nucleotides at known SNP positions and their neighbours can aid data interpretation.

https://doi.org/10.1038/s41477-025-01979-x



Extended Data Fig. 6 | An example of poor coverage for a candidate mobile mRNA. In the Nico- tiana benthamiana annotation of the depicted gene (Niben101Scf11383g00015.1) we find 5 annotated exons of which all are populated with reads at different levels (grey histograms). In the samples from tomato, non-grafted or grafted we see that not all annotated exons are populated with reads and that the exons with coverage are populated in both grafted and non-grafted samples. Coverage over the full length of the mRNA may help reduce the risk of reads mapping to isolated regions being potentially misinterpreted, Extended Data Figure 5. This is a screenshot taken in IGV[37].

https://doi.org/10.1038/s41477-025-01979-x



Extended Data Fig. 7 | Challenges in identifying non-selective mobility versus contamination in high-throughput mobile mRNA detection using RNA-seq data in within-species grafts and cross-species grafts. (**a**) The presence of Genotype 1 reads in Genotype 2 samples and vice versa, across the whole of genome, especially in genes expressed in both tissues is consistent both with non- selective transport and contamination. (**b**) The two genes presented in this schematic figure have different relative expression levels. In Experiment 2 the sequencing depth is insufficient to detect lowly expressed genes.

https://doi.org/10.1038/s41477-025-01979-x

Extended Data Fig. 8 | A bar plot of the blast results of unmapped reads against the NCBI database that matched Nicotiana benthamiana.

nature portfolio

Corresponding author(s): Richard J Morris

Last updated by author(s): Dec 19, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☒	☐ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	We used the only existing published datasets in this study
Data analysis	All code and scripts are freely available from the github repository https://github.com/mtomtom/reanalysis-mobile-mrna/tree/main. We have used largely available software packages as stated in the Methods. There are small pieces of additional custom code - all code to be made free available from our github repository upon publication. Only freely available tools were used. The raw reads were mapped to the references using hisat2 (v.2.1.0), and processed by samtools (v1.9), the expression levels were quantified with Stringtie (v1.3.5). The variants were called with bcftools. The binomial errors were computed in R. The statistical comparison of error rates was performed using baymobil. Distributions were compared using the Kolmogorov-Smirnov test, available in R (47), ks.test.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

We used the following existing published datasets in this study: Cuscuta pentagona (1): PR- JNA257158. Vitis vinifera (3): SRP058158 and SRP058157. Solanum lycopersicum, Nicotiana benthamiana (6): SRP111187. Arabidopsis thaliana (2): PRJNA271927. Citrullus lanatus L. (4): PRJNA553072.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	*Use the terms sex (biological attribute) and gender (shaped by social and cultural circumstances) carefully in order to avoid confusing both terms. Indicate if findings apply to only one sex or gender; describe whether sex and gender were considered in study design; whether sex and/or gender was determined based on self-reporting or assigned and methods used.* *Provide in the source data disaggregated sex and gender data, where this information has been collected, and if consent has been obtained for sharing of individual-level data; provide overall numbers in this Reporting Summary. Please state if this information has not been collected.* *Report sex- and gender-based analyses where performed, justify reasons for lack of sex- and gender-based analysis.*
Reporting on race, ethnicity, or other socially relevant groupings	*Please specify the socially constructed or socially relevant categorization variable(s) used in your manuscript and explain why they were used. Please note that such variables should not be used as proxies for other socially constructed/relevant variables (for example, race or ethnicity should not be used as a proxy for socioeconomic status).* *Provide clear definitions of the relevant terms used, how they were provided (by the participants/respondents, the researchers, or third parties), and the method(s) used to classify people into the different categories (e.g. self-report, census or administrative data, social media data, etc.)* *Please provide details about how you controlled for confounding variables in your analyses.*
Population characteristics	*Describe the covariate-relevant population characteristics of the human research participants (e.g. age, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."*
Recruitment	*Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.*
Ethics oversight	*Identify the organization(s) that approved the study protocol.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	n/n
Data exclusions	n/a
Replication	n/a
Randomization	n/a
Blinding	n/a

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.