NAR Genomics and Bioinformatics, 2026, **8**, lqag071
https://doi.org/10.1093/nargab/lqag071
High Throughput Sequencing Methods

OXFORD

The promise of long-read RNA-seq: reducing bias in analyses of allele imbalance

Nadja Nolte [1,2,*], Marko Petek [1], Pablo Angulo Lara [3], Logan Mulroney [3,4,5], Francesco Nicassio [3], Fabio Marroni [6], Lauren McIntyre [7,*]

[1]Department of Biotechnology and Systems Biology, National Institute of Biology, Ljubljana 1000, Slovenia
[2]Jožef Stefan International Postgraduate School, Ljubljana 1000, Slovenia
[3]Center for Genomic Science of IIT@SEMM, Fondazione Istituto Italiano di Tecnologia, Milano 20139, Italy
[4]European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton CB10 1SD, United Kingdom
[5]Epigenetics and Neurobiology Unit, European Molecular Biology Laboratory, Monterotondo 00015, Italy
[6]Department of Agriculture, Food, Environmental and Animal Sciences, University of Udine, Udine 33100, Italy
[7]Department of Molecular Genetics and Microbiology, University of Florida Genetics Institute, University of Florida Cancer Center, University of Florida, Gainesville, FL 32611, United States

*To whom correspondence should be addressed. Email: mcintyre@ufl.edu
Correspondence may also be addressed to Nadja Nolte. Email: nadja.franziska.nolte@nib.si

Abstract

Inaccurate allele and gene expression counts due to map bias and genome ambiguity lead to high false positive and false negative rates in studies of allelic imbalance. We demonstrate that long read RNA sequencing (RNA-seq) and straightforward quality control measures can be used to reduce bias in allele counts in case studies from four species: *Drosophila melanogaster*, a diploid insect; *Solanum tuberosum*, an autopolyploid plant; *Pongo abelii*, a highly heterozygous diploid primate, and *Homo sapiens*. We recommend (i) mapping to a personalized genome to increase the number of allele assignments; (ii) tracking multimapping reads and tuning mapping parameters to ensure accurate allele and gene expression counts; and (iii) evaluating apparent extreme allele bias to identify errors in genome assembly and annotation. We show that these steps can be executed in a straightforward manner and recommend tools for each step.

Introduction

Reference genomes, typically represented as a single haplotype sequence for a species (Fig. 1A), have been a transformative milestone in modern biology. The quantification of allelic expression relies on being able to separate sequencing reads from RNA sequencing (RNA-seq) experiments based on genetic variation, and bias in allele counts can result in high false positive rates [1]. Single nucleotide polymorphisms (SNPs) N-masking is a common technique to reduce bias, however, incorrect assignment of reads to genes and alleles persists [2]. Alternatively, personalized genomes that incorporate allelic variants of an individual (Fig. 1B) have been shown to reduce bias in quantifying allele counts for multiple species [1–14].

The cost of personalized genomes was limiting their use. Genetic reference panels, such as the Drosophila Genetic Reference Panel [15], the Mouse Collaborative Cross [16], and The Arabidopsis Information Resource (TAIR) [17], re-sequenced sets of nearly homozygous genotypes, enabling individual researchers to use stocks of sequenced genotypes. Crosses between these genotypes created heterozygotes with known allelic variation and enabled personalized genomes in studies of allelic imbalance [5, 9, 18–20]. Recent advancements in sequencing technology and genome assembly algorithms have made phased *de novo* reference genomes more accessible and affordable [21, 22]. Haplotype phased cultivar-specific genomes have been reported for agriculturally important plants, such as apple [23], potato [24–26], and rice [27]. Other species, such as honeybee [28] and orangutan [29], have also released phased haplotype references. In humans, major efforts are underway to capture individual haplotype variation in pangenomes [30]. Haplotype-resolved pangenomes reduce bias in read mapping [31] and can be used to create personalized references [13].

In studies of allele imbalance, unbiased estimates of alleles and gene expression are crucial to avoiding large type I error rates [8–11, 20, 32–34]. Accurate estimates of gene expression directly impact allelic imbalance, and there is a relationship between gene expression and the power to detect allelic imbalance [35–37]. Discarding the multimapping reads can lead to biased estimates of gene expression [35, 36, 38–42] by underestimating gene expression in gene families [39, 40, 43], biasing functional enrichment analysis [44], resulting in higher levels of variability between replicates and increased error rates [43] (Fig. 2G).

Despite the demonstrated advantage to using personalized references, studies of allele imbalance often fail to incorporate them, perhaps because of the perceived additional bioinformatic effort and the lack of demonstration of the impact of the personalized genome on bias reduction with third generation, long-read, sequencing technology. Long-read RNA-seq studies of allele imbalance (see Supplemental Table S1 for mapping strategies used with long-read RNA-seq) have used two, apparently divergent, mapping strategies. Mapping parameters are sensitive to the complexity of the reference, suggesting

Received: January 5, 2026. **Revised:** May 11, 2026. **Accepted:** May 25, 2026

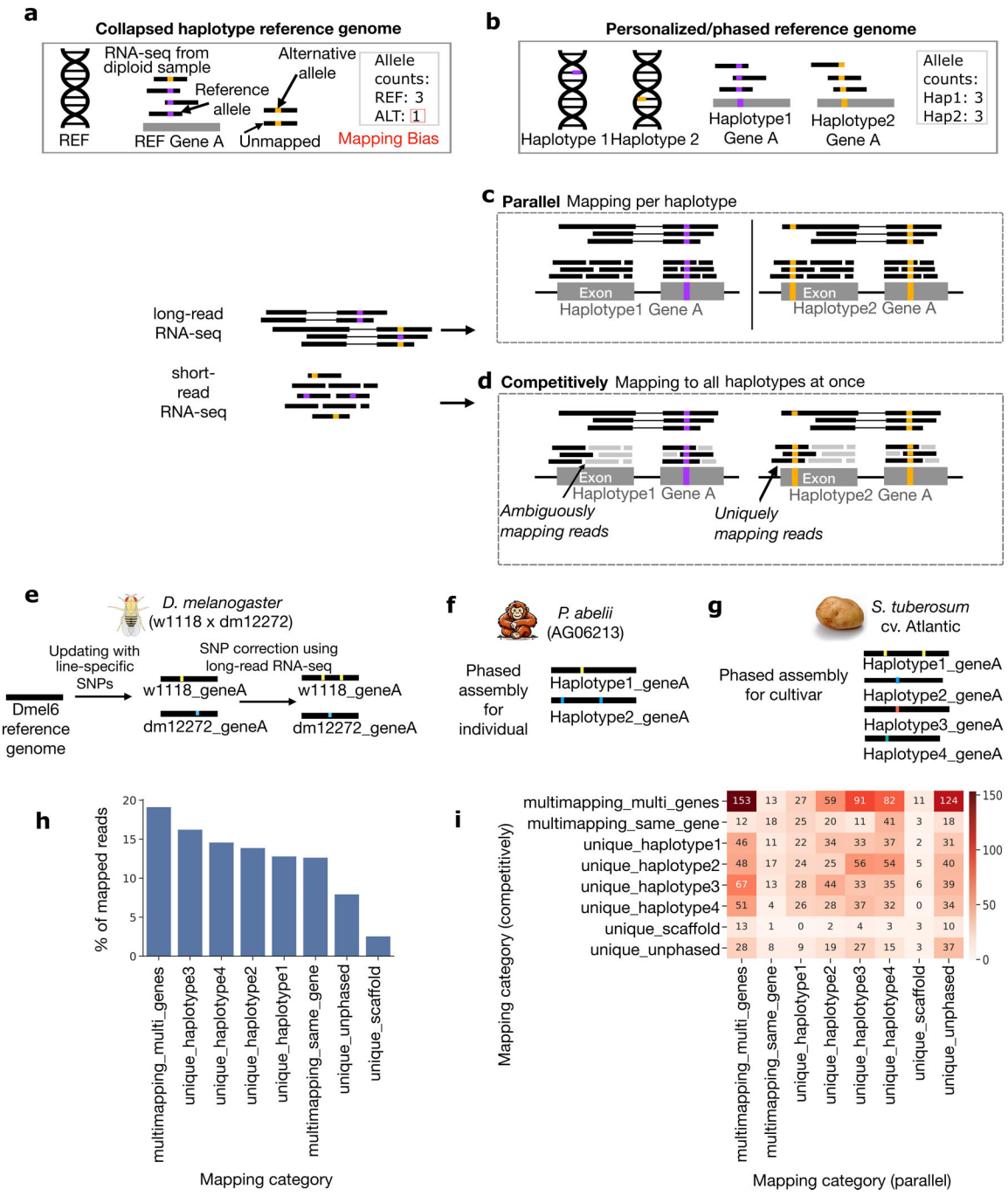

Figure 1. The reference used affects equivalence groups. Tracking multimapping results in equivalent allele and genotype counts regardless of whether mapping is in parallel or competitive. (**a**) Haploid reference genomes and mapping bias. Reads containing the reference allele may be more likely to align than those containing alternative alleles. (**b**) Haplotype-phased personalized reference genomes mitigate mapping bias. RNA-seq data can be mapped either (**c**) in parallel to each haplotype or (**d**) to a combined genotype reference. The proportion of default uniquely mapped reads (black) versus multimapping reads (gray) differs in these two strategies. Tracking the multimapping and equivalently mapping reads separates gene-specific and multigenic reads. Long-read RNA-seq is expected to have less ambiguity and overall fewer multimapping reads compared to short-read RNA-seq. (**e–g**) Case studies for the two mapping strategies applied to (**e**) diploid *D. melanogaster* using a SNP-updated personalized reference, (**f**) diploid *Pongo abelii* with a phased assembly for the individual assayed, and (**g**) tetraploid *Solanum tuberosum* with a phased assembly for the cultivar assayed. (**h**) Barplot of the number of reads in each category with concordant mapping locations in parallel and competitive mapping for *S. tuberosum*. Reads mapping to only one allele are assigned to mapping category "unique_haplotype[1,2,3,4]", reads mapping equally well to multiple alleles of a gene are assigned as "multimapping_same_gene", these are the reads which will be discordant without tracking, and reads mapping equally well to multiple alleles and genes are classified as "multimapping_multi_genes" (**i**) Heatmap showing number of reads with different mapping location in parallel and competitive mapping for *S. tuberosum*.

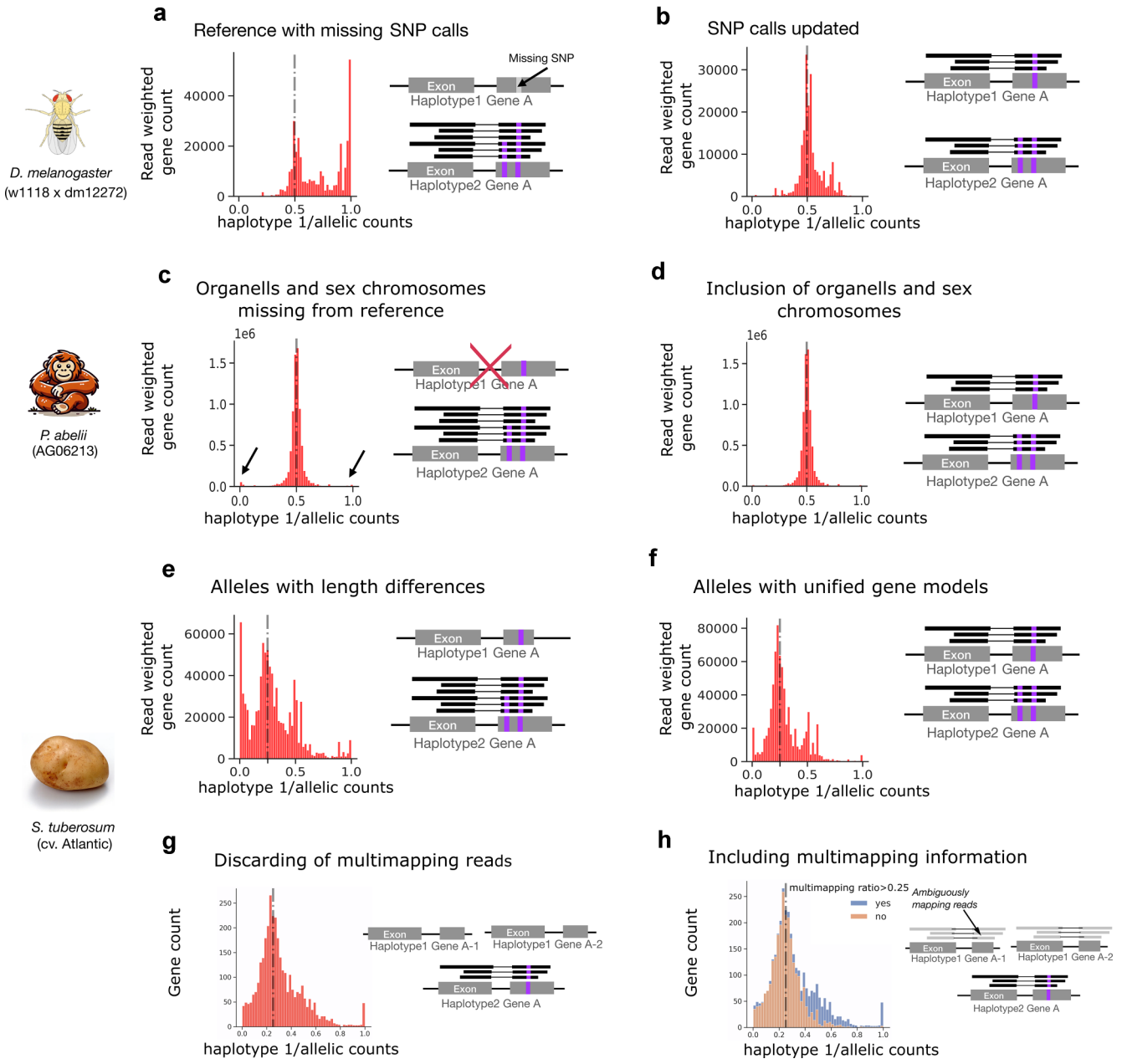

Figure 2. Examples of annotation and assembly errors causing bias that can be identified and corrected with long-read RNA-seq. (**a**) SNP-updated references can introduce bias when there are quality differences in SNP calls between haplotypes. (**b**) Long-read sequencing can address this by correcting for missing SNPs. (**c**) Excluding regions from the reference genome can cause misalignment of reads (arrows), while (**d**) using a complete reference genome resolves such issues. (**e**) Differences in gene lengths between haplotypes can result in reads preferentially aligning to the longer allele, whereas (**f**) unifying annotations across haplotypes using tools like Liftoff reduces this bias. (**g**) Discarding multimapping reads can lead to inaccurate conclusions about allele imbalance, while (**h**) retaining and analyzing multimapping reads provides a straightforward strategy to identify genes affected by ambiguity, such as those with CNV. Dashed lines indicate ratios of allelic counts if all alleles have equal expression.

that default parameters may need to be adjusted if the reference size and complexity changes, as is expected when transitioning from a haploid to a diploid or polyploid reference or moving from a genome to a transcriptome. While tools and pipelines for allele-specific expression analysis with long-read RNA-seq have been developed, such as LORALS [8], IsoPhase [45], IDP-ASE [46], or IsoLaser [47], their use is limited to well annotated and diploid organisms.

We demonstrate using a case study approach of *D. melanogaster*, *Homo sapiens*, *P. abelii*, and *S. tuberosum* how long-read RNA-seq can be used to improve estimates of allele counts and reduce bias. In addition, we provide a schematic overview of the recommended workflow, summarizing the three main steps and highlighting potential sources of bias and their diagnostic interpretation (Supplemental Fig. S1)

Results

Recommendation 1: Use a personalized reference

In our first case study, we evaluate allele counts and gene counts mapping to an updated GRCh38 compared to map-

Table 1. Read assignments based on mapping to the updated GRCh38 or the HG002 assembly

Sample	Read assignment	GRCh38 (%)	HG002 (%)
NA27730	Gene-specific	77.29	71.45
NA26105		79.27	73.31
NA24385		83.73	76.86
NA27730	Allele-specific (hap1)	11.16	14.78
NA26105		10.19	13.86
NA24385		8.14	12.67
NA27730	Allele-specific (hap2)	11.55	13.77
NA26105		10.54	12.84
NA24385		8.13	10.47

The proportion of mapped reads that are haplotype-specific increases for all three samples.

ping to a haplotype-specific reference for HG002 using public data. We used ONT complementary DNA data from three Genome in a Bottle (GIAB) HG002 samples and performed a competitive alignment to the T2T-HG002 assembly and the updated GRCh38 genome [48]. We used Liftoff to map the GENCODE v48 annotations onto each haplotype of the HG002 diploid assembly [49]. We classified genes with only a single annotated copy for both haplotypes of HG002 and GRCh38 as singletons and separated those with CNV. We quantified, for each gene, the number of reads mapping to both haplotypes in the same gene (Gene-specific GS), reads that mapped to a specific haplotype/allele in one gene (Allele-specific hap1/hap2) using the mapping score *ms* to assign the best alignment(s) for a read. While for most genes the differences were small, there were many genes for which there was a dramatic increase in the ability to assign reads to haplotypes (Supplementary Fig. S2). On average 6.2% more of reads could be assigned allele-specifically in HG002 compared to GRCh38 (Table 1).

Recommendation 2: Track multimapping reads, do not filter MapQ = 0

Although it is often implicitly assumed by individual scientists that the genome size and complexity do not influence mapping, as evidenced by the reflexive use of default parameters, complexity of the reference was recognized to be important by mapping algorithm developers [49, 50]. It has been reported that aligning to haplotype-resolved genotype references competitively results in less accurate quantification than aligning in parallel [51, 52]. One possible explanation for the observation is that default filtering of multimapping reads is expected to affect these alignment strategies differently (Fig. 1C and D). Without tracking, reads that map to a specific gene, but not to a specific allele, will appear to be multimapping in the competitive mapping strategy but not in the parallel mapping strategy.

A transcriptome reference further adds complexity as there will be redundant sequence between isoforms, even in relatively simple transcriptomes such as *D. melanogaster,* and filtering multimapping reads is not common practice when aligning to transcriptomes. The use of equivalency classes, rather than filtering, explicitly acknowledges the uncertainty that arises from the use of the more complex reference. Seesaw [14] and Ornaments [3] are frameworks that have been designed to quantify short-read RNA-seq data for analysis of allele imbalance by mapping reads with salmon [53] or kallisto [54] to a diploid-resolved transcriptome and probabilistically assign reads to transcripts and alleles. EMASE uses an expectation maximization algorithm to assign multimappers hierarchically, distinguishing between multiallele, multitranscript, and multigene reads [36]. MMSEQ [55] is another framework using multimapping reads that was proposed for assigning short-read RNA-seq allele-specifically to isoforms using Gibbs sampling.

While probabilistic assignment is an elegant solution, it does not resolve the uncertainty and there is potential for bias in probabilistic assignments [38, 40, 42]. A possible approach to solve this issue is to make use of uncertainty in the quantification output. Alternatively, some researchers have suggested clustering regions with high sequence similarity during gene expression analysis [39, 42, 56] when using transcripts as a reference for mapping [57, 58]. Simulations can provide insights into the impact of the sequence similarity on the uncertainty and bias [20, 59, 60] and the identification of equivalency classes facilitates the evaluation of the reference and annotation for error.

When mapping short-reads, the high proportion of multimapping in repetitive regions makes tracking multimapping reads cumbersome (Fig. 1D). Advances in long-read RNA-seq with the potential to sequence the full transcript, increase the likelihood of capturing single nucleotide and structural variation between haplotypes as the chances of a single read spanning multiple variants are higher (Fig. 1D). We hypothesized that parallel mapping to haplotypes with subsequent read assignment based on the best mapping location and competitive mapping to genotypes should result in comparable estimates of allele-specific and gene-specific expression assuming that: the accuracy of alignment tools remains unaffected by the reference genome's size and similarity; that reads multimapping to alleles or loci are identifiable in both approaches; and that the score function is due to the match quality between the read and the reference.

We use the following designations for each read: (i) unique mapping to a haplotype/allele, (ii) mapping to a gene; or (iii) multimapping to multiple genes. We used Minimap2, one of the most popular alignment tools for long-read mapping [61, 62]. We use this tool here, but the results are general and should not be unduly influenced by this particular choice, although some of the details, as we describe, are particular to the mapping algorithm and we acknowledge that the best alignment may differ based on exactly what function is used. In Minimap2, assignment of allele-specific reads can be determined by alignment score (AS), number of mismatches (NM), or primary alignment score (*ms*). A read that maps equally well to >1 location has equivalent alignment scores (AS and *ms* score), however one mapping location will be chosen as primary alignment and the other one(s) as secondary even if the two alignments are identical [in case –secondary = yes and N (number of secondary alignments reported) is > 0]. As we wish to compare the parallel and competitive alignment strategies, we opt to use the *ms* score (not AS score) to identify the best match as this setting avoids favoring unspliced mapping to pseudogenes [61] (see Supplementary Table S8 for explanation of alignment scores).

Parallel mapping to individual haplotypes

Figure 1C shows the parallel mapping to individual haplotypes. This approach aligns reads to the individual haplotypes separately. allele-specific reads are defined by the haplotype with the highest mapping score for each read. Those reads

with identical mapping scores to multiple reference haplotypes are combined with the allele-specific reads to estimate gene expression [8–11, 20, 32–34]. This parallel mapping approach requires an additional script or tool that decides for each read whether the alignment is better to one haplotype or equally good.

Competitively mapping to genotypes

Figure 1D shows competitively mapping to genotypes. This approach uses a genotype assembly that consists of a compilation of multiple phased haplotypes [63]. The parallel haplotype mapping strategy requires an explicit choice of the criteria to use to consider a read allele-specific , with options for alignment scores (AS or *ms*), as well as number of mismatches. In contrast, in the competitive genotype mapping strategy, the mapping parameters will be what determines if a read has a primary, allele specific, alignment. Reads that map to multiple haplotypes for a single gene (gene-specific reads in the haplotype alignment strategy) will be multimapping to the gene in this strategy. Counting the gene-specific reads is then by necessity a separate step [2, 36, 54, 55]. Reads that are multimapping to a specific gene are expected to be the reads that map uniquely but equivalently to the gene in the parallel mapping strategy. By default, these reads would be filtered in the competitive mapping strategy, leading to an apparent discrepancy.

We mapped long-read RNA-seq data (PacBio Iso-Seq and ONT) to gene regions for personalized haplotype and genotype references in four species: *D. melanogaster* (fruit fly, Fig. 1E, SNP-updated), *P. abelii* (male Sumatran orangutan fibroblast cell line, Fig. 1F, phased assembly with PacBio Hifi and ONT ultra-long-reads), and *S. tuberosum* cv. Atlantic (potato, Fig. 1G, phased assembly with PacBio HiFi, ONT DNA, Hi-C). Details on these genome assemblies and long-read RNA-seq samples are available in Supplemental Table S2. For the four species, we mapped the long-read RNA-seq to the haplotypes in parallel and to the genotypes competitively (see Supplemental Methods for detailed mapping parameters). A read was considered allele-specific if the mapping location with the highest *ms* score reported by Minimap2 was unique to one allele. When a read had multiple maximum *ms* scores, it was labeled as "multimapping gene specific" if all maximum *ms* scores were observed for the same locus or "multimapping multigene" when the maximum *ms* score occurred in multiple loci. Tools developed for long-read RNA-seq quantification can be utilized for allele-specific read assignment. For instance oarfish [64] assigns reads to transcripts based on *ms* scores and outputs read assignment probabilities and raw unique and ambiguous counts, which informs about the uniqueness and certainty of the alignment.

Competitive and parallel mapping strategies are 'identical'

Mapping to haplotypes and mapping to genotypes was identical for >96% of all loci in all the four examples. A total of 19% of reads are multimapping to multiple loci in both strategies, and these reads map to the same set of locations in both strategies (Fig. 1H). The 13% of reads labeled as multimapping same gene are the reads that without tracking would be assigned differently in the two strategies (Fig. 1H). There are some small differences in these two approaches (Fig. 1I). In the 1.9%–3.9% of genes with different read counts (Supplemental Table S4), most differ by a single read. Only 0.52%–0.75% genes have >1 read difference between the two mapping strategies. Overall, only 0.021%–0.033% of all reads have different mapping locations in these two strategies. We hypothesized that these small differences were due to properties of the mapping algorithm and the indexing of the references. Minimap2 is not guaranteed to find the best alignment, the indexing will affect the map location. The impact of the index on the map location could be problematic for large genomes where there are potential areas of sequence similarities, such as in phased haplotypes [61]. In support of this hypothesis, we observe that for *D. melanogaster* (cumulative size of gene regions on haplotypes: 0.2GB) a smaller percentage of reads had a different alignment position between the parallel haplotype mapping and the competitive genotype mapping compared to the larger gene regions of *P. abelii* (cumulative size of gene regions on haplotypes: 3.0 GB; similar to human) and *S. tuberosum* (cumulative size of gene regions on haplotypes: 1.1 GB) (Supplemental Table S4).

Reference complexity affects mapping

Minimap2, like other tools, uses a minimizer index, and parameters for indexing and mapping can influence the primary mapping position found for a read [50, 61]. Genotype references contain more sequence ambiguity, by definition, since multiple alleles are included for each locus. Increasing the threshold for retaining repetitive k-mers can increase the mapping accuracy at cost of longer runtime (see Supplemental Table S4 for the impact of changing default -f 0.0002 to -f 0.000002 for *S. tuberosum*). Using the -a (or -c for .paf output) option in Minimap2 will perform a base-level alignment, which is more accurate than the default reported alignment and we therefore choose to use the base level alignment in our case study [62]. Minimap2 with parameter -P (avoids setting of primary chains, see Supplemental methods) resulted in overall higher concordance of the mapping strategies (Supplemental Table S4). To test our hypothesis that mapping parameters affect disagreement between the competitive and parallel mapping strategies, we repeated the case study with the parameter -N 200 (without -P) recommended for transcript quantification [64, 65]. As predicted, the discordance between the mapping strategies was slightly higher (e.g. loci with different read counts >1 read increased from 0.7% to 2.3% *for P. abelii*) (Supplemental Table S4 and Supplemental Fig. S4A–C). In addition, we also observe an increase in concordance of mapping strategies between minimap2 v 2.24 and v 2.28 (Supplemental Table S4).

Mapping to the transcriptome increases the complexity of the reference

We mapped the reads from one of the *D. melanogaster* genotypes (dm12272) to the haplotype resolved personalized transcriptome using a competitive mapping strategy. We compared mapping to the transcriptome to mapping to the gene regions (described above). As expected, the proportion of multimapping reads increased by 10%, reflecting the increase in the equivalency classes in the transcriptome compared to gene regions (Supplementary Fig. S5).

We examined the impact of mapping to the transcriptome compared to gene-regions on whether the read mapping was allele-specific, haplotype-specific and gene-specific for the *D. melanogaster* data. We found that while there was a large overlap, 77% of the reads were allele-specific in both gene region and transcriptome mapping, there were reads that mapped allele-specifically to gene regions and not the transcriptome and vice versa. The number of reads map-

ping allele-specifically was similar in these two strategies with 1% more reads mapping allele-specifically to the transcriptome (Supplementary Table S6). SQANTI3 categories were computed for each read based on genome alignments [66, 67]. Reads classified as intergenic, novel-in-catalog, novel-not-in-catalog, antisense, and fusion were slightly more likely to map to the gene regions allele-specifically than to the transcriptome; while those reads that were classified as full-splice-match, incomplete-splice-match and genic were slightly more likely to map allele-specifically to the transcriptome (Supplementary Table S7). This suggests that the discrepancies are again due to the difference in the complexity of the reference and underscores the importance of examining the assumption about the reference complexity in choosing parameters for mapping. Likely this discrepancy could be reduced by further optimization of the mapping parameters.

Longer reads increase specificity

The average read length for allele-specific reads was higher than gene-specific and multigene multimapping reads. The average read lengths of multimapping reads were 462 bp in *D. melanogaster*, 2031 bp in *P. abelii*, and 748 bp in S. tuberosum (Supplemental Table S3). In contrast, the corresponding average read lengths for allele-specific reads were 749 bp, 2379 bp, and 838 bp, respectively (Supplemental Table S3). In addition, we observed that for *P. abelii*, only 30% of the reads multimap to the same gene and 68% of the reads mapped allele-specifically, whereas in *D. melanogaster* 75% of the reads multimap to the same gene or multiple genes (Supplemental Fig. S3A and B, and Supplemental Table S2). The result is that longer reads reduce the number of equivalency classes.

Multimapping reads identify genes that share conserved sequences

In *P. abelii* the HOXC4, HOXC6, HOXC8 form a highly conserved gene family. There were 799 reads that multimapped to these loci with an average read length of 1786 bp. A total of 42 reads mapped allele-specifically to HOXC4 and none to HOXC6 and HOXC8. In *D. melanogaster* there were 7581 multimapping reads between loci CG31775 and CG42586 and 16 allele-specific reads. These two genes are annotated paralogs [68]. In both examples, the relatively low number of allele-specific reads compared to the high number of multimapping reads, indicates inferences at these genes need to be treated with caution.

Differences in multimapping between haplotypes can signal CNV

Differences in annotated structural variation between haplotypes can introduce biases [41]. For example, if a gene (or part of a gene) is annotated as duplicated on one haplotype but not the others, multimapping reads may appear to reduce the allele-specific counts on the haplotype with the duplication. In the *S. tuberosum* the allele on haplotype 4 with a single copy for the gene "translationally controlled tumor protein" (Soltu.Atl_v3.01_4G026870.2, see genomic locus representation in Supplemental Fig. S7A) shows 701 allele-specific counts, while other haplotypes' alleles (Soltu.Atl_v3.01_1G026620.2, Soltu.Atl_v3.01_2G031140.2, Soltu.Atl_v3.01_3G024160.4) show no allele-specific counts. However, there are reads multimapping to these loci and their annotated paralogs on haplotypes 1, 2, and 3 (Supplemental Fig. S7A). If multilocus multimapping reads had been discarded (Fig. 2G), the confounding among these genes would not be clear and an incorrect inference of allele imbalance for "translationally controlled tumor protein" would be reported (Supplemental Fig. S7B). Reporting ratios of multimapping reads versus uniquely mapping reads helps to identify genes with high ambiguity (Fig. 2H).

Recommendation 3: Use observed extreme bias to identify errors in assembly and annotation.

In our case studies of *D. melanogaster*, *P. abelii*, and *S. tuberosum* we demonstrate how to identify and correct errors. In normal tissue, we expect most of the genes to be expressed equally from all alleles. Genes that have a high proportion (e.g. >90%) of reads aligning exclusively to one allele, may indicate imprinting or other epigenetic effects, but may also indicate that there are gene annotation or assembly errors in the reference. If the allele imbalance indicates a dominant haplotype this is likely to be due to differences in assembly or annotation quality between the haplotypes rather than reflecting a genuine biological phenomenon [5, 9, 69, 70]. Long-reads make identifying and correcting errors more straightforward [71].

Missing SNPs

For the *D. melanogaster* experiment the F1 genotypes were a cross where SNPs for the two parental lines relative to the haploid reference genome were called independently for the w1118 line [72] and dm12272 line [73]. We updated the reference for each of the parents based on the published SNPs and observed an overall bias toward the dm12272 haplotype (Fig. 2A), perhaps not surprising as there were more SNPs in the dm12272 compared to the w1118. Positions predicted to be heterozygous with a dm12272 SNP relative to the reference and w1118 were evaluated. If there was no evidence for the reference allele, we assumed that the SNP was missing from the w1118 calls and included this SNP in the w1118 variant calls. When we remapped using the corrected w1118 variants, haplotype bias was resolved (Fig. 2B).

Missing genes

Mapping algorithms are "greedy," meaning they will map reads even with high levels of mismatches. Missing genes in the reference assembly can lead to reads mapping erroneously to other alleles (Fig. 2C). If a gene is absent, reads originating from that gene are likely to map to similar genes elsewhere. This may manifest as allele imbalance if the alleles at the missing locus map preferentially to one of the haplotypes. In the *P. abelii*, the gene *PDHA2* had 1040 allele-specific reads aligned to haplotype 1, none to haplotype 2, and 5 reads mapped non-allele-specifically. This high bias towards haplotype 1 suggested an assembly or annotation error. When the allele-specific reads were examined, there was evidence for >2 haplotypes in the alignments suggesting a potential missing gene with similar sequence. The paralog of *PDHA2* known as *PDHA1* is located on the X chromosome. The initial mapping was done for autosomes only and these results suggested that the allele-specific reads for *PDHA2* might originate from the *PDHA1* locus. By expanding the reference to include the X and Y chromosomes, as well as mitochondrial DNA, the bias towards *PDHA2* haplotype 1 was resolved.

All 1040 reads previously aligned to the *PDHA2* haplotype 1 allele now aligned to *PDHA1* on the X chromosome. Indeed, even the five multimapping gene-specific reads for *PDHA2* aligned to *PDHA1*. In total 1507 genes were located on the X, Y and mitochondrial DNA. When we updated the reference to include these chromosomes, 636 996 reads mapped to these genes and of these 71% aligned in the initial mapping, with 57% of these aligning alleles specifically (145 238 to haplotype 1 and 120 488 to haplotype 2; Supplemental Table S5). Including the missing chromosomes decreased the number of loci with apparent extreme bias (Fig. 2D).

Length variation between haplotypes

Incorrect exon–intron structures are common artifacts of genome annotations [74–77]. In the *S. tuberosum* assembly the gene annotation was performed per haplotype, and genes were not explicitly matched between haplotypes during annotation [25]. Alleles with more than a 10% difference between the maximum and minimum lengths of annotated genes showed higher proportions of allele-specific reads mapping to the longest allele (Fig. 2E). For 5099 of the 9130 genes with all four alleles present in the assembly, there was more than a 20% difference in the length of the spliced representative transcript (Supplemental Fig. S6A). We unified the annotation between haplotypes using Liftoff [49], which led to more similar transcript lengths between haplotypes (Supplemental Fig. S6B) and more balanced allelic ratios (Fig. 2F). For example, the number of genes with allelic ratio >0.9 towards haplotype 1 decreased from 136 to 61. A reduction in extreme bias was observed for all haplotypes (Supplemental Fig. S8). Long-read RNA-seq can also be used to improve existing annotations by identifying unannotated transcripts and genes with tools like BAMBU [74]. But to avoid bias due to read misassignment it needs to be ensured that the same transcript models are identified on all haplotypes.

Discussion

Haplotype-resolved telomere-to-telomere genomes are available for diploid individuals [78, 79] to highly polyploid plant cultivars [80, 81]. The next step is the construction of pangenomes. Pangenomes incorporate population level genetic variation [31, 82–84] and tools for pangenome mapping [31, 85] and allele identification and expression estimates for human pangenomes are being developed.

Allelic variation, CNV, ploidy and gene families all contribute to sequence similarity and lead to uncertainty. The degree of uncertainty will vary across the genome and how best to identify and incorporate uncertainty into estimates of allelic and gene expression will take time and effort. Meanwhile, the practice of filtering multimapping reads should be discontinued, even when aligning in parallel to haplotypes. Although read lengths increase specificity of mapping, not all reads will be allele, or gene specific. In our case studies, 30%–75% of reads were gene specific, but not allele specific, and 2%–9% of reads multimapped between loci. Tracking multimapping reads for reads that mapped to multiple loci identified uncertainty and enabled correcting mis-assemblies and mis-annotations and the identification of hidden paralogs. Tracking reads that differ in location, and number of alignments between haplotypes, may also pinpoint hidden polymorphisms in CNV between the haplotypes. CNVs are difficult to disentangle and especially challenging when they are recent expansions.

These considerations are particularly relevant in the context of cancer, where somatic structural variation, aneuploidy, and copy number changes make allele-specific analyses especially challenging [86–91]. Improved assemblies and haplotype-resolved references for cancer genomes would substantially enhance our ability to distinguish true regulatory differences from artifacts introduced by structural complexity. This would have direct implications both for clinical diagnostics, where accurate detection of allelic imbalance may inform patient stratification, and for research projects aimed at understanding tumor evolution and regulatory mechanisms in cancer.

The success of long reads in resolving complex structural polymorphisms in DNA [92–95] presents an opportunity to re-examine pipelines for allele imbalance with long-read RNA-seq. In our case studies we show how personalized references and tracking multimapping of the long-read RNA-seq data can be used to identify, and correct mis-assembly and mis-annotations using relatively straightforward bioinformatic techniques and existing tools, counter to the perception that using personalized reference, and tracking multimapping reads is complicated and time consuming.

Although tools for allele-specific long-read RNA analysis exist, there is currently no universally applicable workflow that can be used for different organisms. Our recommendations apply to diploid and polyploid organisms as well as model and non-model organisms.

Supplementary data

Supplementary data is available at NAR Genomics & Bioinformatics online.

Acknowledgements

We thank the Eichler Lab for their work on the Pongo abelii phased assembly and sharing the gene Liftoff annotation and Alison Morse for work on the Drosophila data. Thanks to all the members of the LongTREC network for discussion.

Author contributions: Nadja Nolte (Conceptualization [equal], Data curation [equal], Formal analysis [equal], Software [equal], Writing—original draft [equal], Writing—review & editing [equal]), Marko Petek (Writing—original draft [equal], Writing—review & editing [equal]), Pablo Angulo Lara (Formal analysis [equal], Software [equal], Validation [equal], Writing—original draft [equal], Writing—review & editing [equal]), Logan Mulroney (Writing—original draft [equal], Writing—review & editing [equal]), Francesco Nicassio (Writing—original draft [equal], Writing—review & editing [equal]), Fabio Marroni (Conceptualization [equal], Writing—original draft [equal], Writing—review & editing [equal]), Lauren McIntyre (Conceptualization [equal], Funding acquisition [equal], Investigation [equal], Supervision [equal], Writing—original draft [equal], Writing—review & editing [equal]).

Conflict of interest

Logan Mulroney has received reimbursement of travel and accommodation expenses to speak at Oxford Nanopore

Technologies (ONT) conferences. All other authors have declared no competing interest.

Funding

This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Actions Doctoral Network "LongTREC" grant agreement no. 101072892; and the Slovenian Research and Innovation Agency grant agreements no. P4-0165 and P4-0431; Associazione Italiana per la Ricerca sul Cancro (AIRC) (IG22851), National Center for Gene Therapy and Drugs based on RNA Technology (CN00000041) supported by European Union – Next Generation EU, Mission 4, Component 2, CUP B93D21010860004; Ministry of Health – bando POS – tr.3 – T3-AN-04 – GENERA; an EMBL ETPOD Fellowship; the NIH funding GM137430 and the department of Molecular Genetics and Microbiology, The University of Florida Genetics Institute, the University of Florida Cancer Center, and the University of Florida Research Computing Center (www.rc.ufl.edu). F.N. is a member of the Human RNome Consortium. Open access publication fees were covered by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Actions Doctoral Network "LongTREC" (grant agreement no. 101072892).

Data availability

The Nextflow [96] pipeline for allele-specific mapping to haplotype and genotype references, mapping comparison, plots, and additional scripts for reference preparation is available at (https://github.com/nadjano/nf-ASE-mapping-comparison and https://doi.org/10.5281/zenodo.20411814). Recommended minimap2 mapping parameters can be found in the README. Liftoff gene annotations for *S. tuberosum* cv Atlantic and *H. sapiens* HG002 have been deposited to Zenodo (https://doi.org/10.5281/zenodo.17279424). The long-read raw data for case study analyses are available at NCBI's Sequence Read Archive under accessions SRR29784420–SRR29784421 (*D. melanogaster*), SRR27438208 (*P. abelii*), and SRR14993892–SRR14993894 (*S. tuberosum* cv Atlantic). For *H. sapiens*, samples NA24385, NA27730, and NA26105 from the Genome in a Bottle HG002 ONT cDNA project were used.

References

1. Degner JF, Marioni JC, Pai AA *et al.* Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* 2009;25:3207–12. https://doi.org/10.1093/bioinformatics/btp579
2. Pandey RV, Franssen SU, Futschik A *et al.* Allelic imbalance metre (A llim), a new tool for measuring allele-specific gene expression with RNA -seq data. *Mol Ecol Resour* 2013;13:740–5. https://doi.org/10.1111/1755-0998.12110
3. Adduri A, Kim S. Ornaments for efficient allele-specific expression estimation with bias correction. *Am J Hum Genet* 2024;111:1770–81. https://doi.org/10.1016/j.ajhg.2024.06.014
4. Castel SE, Levy-Moonshine A, Mohammadi P *et al.* Tools and best practices for data processing in allelic expression analysis. *Genome Biol* 2015;16:195. https://doi.org/10.1186/s13059-015-0762-6
5. Crowley JJ, Zhabotynsky V, Sun W *et al.* Analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance. *Nat Genet* 2015;47:353–60. https://doi.org/10.1038/ng.3222
6. Demirdjian L, Xu Y, Bahrami-Samani E *et al.* Detecting allele-specific alternative splicing from population-scale RNA-seq data. *Am J Hum Genet* 2020;107:461–72. https://doi.org/10.1016/j.ajhg.2020.07.005
7. Dobin A, Davis CA, Schlesinger F *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013;29:15–21. https://doi.org/10.1093/bioinformatics/bts635
8. Glinos DA, Garborcauskas G, Hoffman P *et al.* Transcriptome variation in human tissues revealed by long-read sequencing. *Nature* 2022;608:353–9. https://doi.org/10.1038/s41586-022-05035-y
9. Graze RM, Novelo LL, Amin V *et al.* Allelic imbalance in Drosophila hybrid heads: exons, isoforms, and evolution. *Mol Biol Evol* 2012;29:1521–32. https://doi.org/10.1093/molbev/msr318
10. Munger SC, Raghupathy N, Choi K *et al.* RNA-seq alignment to individualized genomes improves transcript abundance estimates in multiparent populations. *Genetics* 2014;198:59–73. https://doi.org/10.1534/genetics.114.165886
11. Rozowsky J, Abyzov A, Wang J *et al.* AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol Syst Biol* 2011;7:522. https://doi.org/10.1038/msb.2011.54
12. Rozowsky J, Gao J, Borsari B *et al.* The EN-TEx resource of multi-tissue personal epigenomes & variant-impact models. *Cell* 2023;186:1493–511. https://doi.org/10.1016/j.cell.2023.02.018
13. Vaddadi K, Lin M-J, Majidian S *et al.* Minimizing reference bias with an imputed personalized reference. *Genome Res* 2026;36:740–53.
14. Wu EY, Singh NP, Choi K *et al.* SEESAW: detecting isoform-level allelic imbalance accounting for inferential uncertainty. *Genome Biol* 2023;24:165. https://doi.org/10.1186/s13059-023-03003-x
15. Mackay TFC, Richards S, Stone EA *et al.* The Drosophila melanogaster Genetic Reference Panel. *Nature* 2012;482:173–8. https://doi.org/10.1038/nature10811
16. Consortium CC. The genome architecture of the Collaborative Cross Mouse genetic reference population. *Genetics* 2012;190:389–401. https://doi.org/10.1534/genetics.111.132639
17. Rhee SY. The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Res* 2003;31:224–8. https://doi.org/10.1093/nar/gkg076
18. Churchill GA, Gatti DM, Munger SC *et al.* The diversity outbred mouse population. *Mamm Genome* 2012;23:713–8. https://doi.org/10.1007/s00335-012-9414-2
19. Fear JM, León-Novelo LG, Morse AM *et al.* Buffering of genetic regulatory networks in *Drosophila melanogaster*. *Genetics* 2016;203:1177–90. https://doi.org/10.1534/genetics.116.188797
20. León-Novelo LG, McIntyre LM, Fear JM *et al.* A flexible Bayesian method for detecting allelic imbalance in RNA-seq data. *BMC Genomics* 2014;15:920. https://doi.org/10.1186/1471-2164-15-920
21. Miga KH, Eichler EE. Envisioning a new era: complete genetic information from routine, telomere-to-telomere genomes. *Am J Hum Genet* 2023;110:1832–40. https://doi.org/10.1016/j.ajhg.2023.09.011
22. Sarashetti P, Lipovac J, Tomas F *et al.* Evaluating data requirements for high-quality haplotype-resolved genomes for creating robust pangenome references. *Genome Biol* 2024;25:312. https://doi.org/10.1186/s13059-024-03452-y
23. Su Y, Yang X, Wang Y *et al.* Phased telomere-to-telomere reference genome and pangenome reveal an expansion of resistance genes during apple domestication. *Plant Physiol* 2024;195:2799–814. https://doi.org/10.1093/plphys/kiae258
24. Godec T, Beier S, Rodriguez-Granados NY *et al.* Haplotype-resolved genome assembly of the tetraploid potato cultivar Désirée. *Sci Data* 2025;12:1044. https://doi.org/10.1038/s41597-025-05372-3
25. Hoopes G, Meng X, Hamilton JP *et al.* Phased, chromosome-scale genome assemblies of tetraploid potato reveal a complex genome,

transcriptome, and predicted proteome landscape underpinning genetic diversity. *Mol Plant* 2022;15:520–36. https://doi.org/10.1016/j.molp.2022.01.003
26. Sun H, Jiao W-B, Krause K *et al.* Chromosome-scale and haplotype-resolved genome assembly of a tetraploid potato cultivar. *Nat Genet* 2022;54:342–8. https://doi.org/10.1038/s41588-022-01015-0
27. Feng J-W, Lu Y, Shao L *et al.* Phasing analysis of the transcriptome and epigenome in a rice hybrid reveals the inheritance and difference in DNA methylation and allelic transcription regulation. *Plant Commun* 2021;2:100185. https://doi.org/10.1016/j.xplc.2021.100185
28. Bresnahan ST, Lee E, Clark L *et al.* Examining parent-of-origin effects on transcription and RNA methylation in mediating aggressive behavior in honey bees (Apis mellifera). *BMC Genomics* 2023;24:315. https://doi.org/10.1186/s12864-023-09411-4
29. Mao Y, Harvey WT, Porubsky D *et al.* Structurally divergent and recurrently mutated regions of primate genomes. *Cell* 2024;187:1547–62. https://doi.org/10.1016/j.cell.2024.01.052
30. Liao W-W, Asri M, Ebler J *et al.* A draft human pangenome reference. *Nature* 2023;617:312–24. https://doi.org/10.1038/s41586-023-05896-x
31. Sibbesen JA, Eizenga JM, Novak AM *et al.* Haplotype-aware pantranscriptome analyses using spliced pangenome graphs. *Nat Methods* 2023;20:239–47. https://doi.org/10.1038/s41592-022-01731-9
32. Akama S, Shimizu-Inatsugi R, Shimizu KK *et al.* Genome-wide quantification of homeolog expression ratio revealed nonstochastic gene regulation in synthetic allopolyploid *Arabidopsis*. *Nucleic Acids Res* 2014;42:e46. https://doi.org/10.1093/nar/gkt1376
33. Kuo T, Frith MC, Sese J *et al.* EAGLE: explicit alternative genome likelihood evaluator. *BMC Med Genomics* 2018;11:28. https://doi.org/10.1186/s12920-018-0342-1
34. Page JT, Gingle AR, Udall JA. PolyCat: a resource for genome categorization of sequencing reads from allopolyploid organisms. *G3 (Bethesda)* 2013;3:517–25. https://doi.org/10.1534/g3.112.005298
35. León-Novelo L, Gerken AR, Graze RM *et al.* Direct testing for allele-specific expression differences between conditions. *G3 (Bethesda)* 2018;8:447–60. https://doi.org/10.1534/g3.117.300139
36. Raghupathy N, Choi K, Vincent MJ *et al.* Hierarchical analysis of RNA-seq reads improves the accuracy of allele-specific expression. *Bioinformatics* 2018;34:2177–84. https://doi.org/10.1093/bioinformatics/bty078
37. Sherbina K, León-Novelo LG, Nuzhdin SV *et al.* Power calculator for detecting allelic imbalance using hierarchical bayesian model. *BMC Res Notes* 2021;14:436. https://doi.org/10.1186/s13104-021-05851-x
38. Deschamps-Francoeur G, Simoneau J, Scott MS. Handling multi-mapped reads in RNA-seq. *Comput Struct Biotechnol J* 2020;18:1569–76. https://doi.org/10.1016/j.csbj.2020.06.014
39. Robert C, Watson M. Errors in RNA-seq quantification affect genes of relevance to human disease. *Genome Biol* 2015;16:177. https://doi.org/10.1186/s13059-015-0734-x
40. Szabelska-Beresewicz A, Zyprych-Walczak J, Siatkowski I *et al.* Ambiguous genes due to aligners and their impact on RNA-seq data analysis. *Sci Rep* 2023;13:21770. https://doi.org/10.1038/s41598-023-41085-6
41. Van De Geijn B, McVicker G, Gilad Y *et al.* WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat Methods* 2015;12:1061–3.
42. Zytnicki M. mmquant: how to count multi-mapping reads? *BMC Bioinformatics* 2017;18:411. https://doi.org/10.1186/s12859-017-1816-4
43. Choi K, Raghupathy N, Churchill GA. A bayesian mixture model for the analysis of allelic expression in single cells. *Nat Commun* 2019;10:5188. https://doi.org/10.1038/s41467-019-13099-0
44. Almeida Da Paz M, Warger S, Taher L. Disregarding multimappers leads to biases in the functional assessment of NGS data. *BMC Genomics* 2024;25:455. https://doi.org/10.1186/s12864-024-10344-9
45. Wang B, Tseng E, Baybayan P *et al.* Variant phasing and haplotypic expression from long-read sequencing in maize. *Commun Biol* 2020;3:78. https://doi.org/10.1038/s42003-020-0805-8
46. Deonovic B, Wang Y, Weirather J *et al.* IDP-ASE: haplotyping and quantifying allele-specific expression at the gene and gene isoform level by hybrid sequencing. *Nucleic Acids Res* 2017;45:e32. https://doi.org/10.1093/nar/gkw1076
47. Quinones-Valdez G, Amoah K, Xiao X. Long-read RNA-seq demarcates *cis*- and *trans*-directed alternative RNA splicing. *Nat Commun* 2025;16:9603. https://doi.org/10.1038/s41467-025-64605-6
48. Dwarshuis N, Kalra D, McDaniel J *et al.* The GIAB genomic stratifications resource for human reference genomes. *Nat Commun* 2024;15:9029. https://doi.org/10.1038/s41467-024-53260-y
49. Shumate A, Salzberg SL. Liftoff: accurate mapping of gene annotations. *Bioinformatics* 2021;37:1639–43. https://doi.org/10.1093/bioinformatics/btaa1016
50. Kim D, Paggi JM, Park C *et al.* Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* 2019;37:907–15. https://doi.org/10.1038/s41587-019-0201-4
51. Hu G, Grover CE, Arick MA *et al.* Homoeologous gene expression and co-expression network analyses and evolutionary inference in allopolyploids. *Brief Bioinform* 2021;22:1819–35. https://doi.org/10.1093/bib/bbaa035
52. Kuo TCY, Hatakeyama M, Tameshige T *et al.* Homeolog expression quantification methods for allopolyploids. *Brief Bioinform* 2020;21:395–407. https://doi.org/10.1093/bib/bby121
53. Patro R, Duggal G, Love MI *et al.* Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* 2017;14:417–9. https://doi.org/10.1038/nmeth.4197
54. Bray NL, Pimentel H, Melsted P *et al.* Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* 2016;34:525–7. https://doi.org/10.1038/nbt.3519
55. Turro E, Su S-Y, Gonçalves Â *et al.* Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. *Genome Biol* 2011;12:R13. https://doi.org/10.1186/gb-2011-12-2-r13
56. Sarkar H, Srivastava A, Bravo HC *et al.* Terminus enables the discovery of data-driven, robust transcript groups from RNA-seq data. *Bioinformatics* 2020;36:i102–10. https://doi.org/10.1093/bioinformatics/btaa448
57. Davidson NM, Oshlack A. Corset: enabling differential gene expression analysis for de novoassembled transcriptomes. *Genome Biol* 2014;15:410.
58. Turro E, Astle WJ, Tavaré S. Flexible analysis of RNA-seq data using mixed effects models. *Bioinformatics* 2014;30:180–8. https://doi.org/10.1093/bioinformatics/btt624
59. Hafezqorani S, Yang C, Lo T. *et al.* Trans-NanoSim characterizes and simulates nanopore RNA-sequencing data. *GigaScience* 2020;9:giaa061. https://doi.org/10.1093/gigascience/giaa061
60. Lin M, Iyer S, Chen N *et al.* Measuring, visualizing, and diagnosing reference bias with biastools. *Genome Biol* 2024;25:101. https://doi.org/10.1186/s13059-024-03240-8
61. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;34:3094–100. https://doi.org/10.1093/bioinformatics/bty191
62. Li H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* 2021;37:4572–4. https://doi.org/10.1093/bioinformatics/btab705
63. Jallet A, Friedrich A, Schacherer J. Impact of the acquired subgenome on the transcriptional landscape in *Brettanomyces bruxellensis* allopolyploids. *G3 (Bethesda)* 2023;13:jkad115. https://doi.org/10.1093/g3journal/jkad115

64. Zare Jousheghani Z, Singh NP, Patro R. Oarfish : enhanced probabilistic modeling leads to improved accuracy in long read transcriptome quantification. *Bioinformatics* 2025;41:i304–13. https://doi.org/10.1093/bioinformatics/btaf240
65. Ji HJ, Pertea M. Enhancing transcriptome expression quantification through accurate assignment of long RNA sequencing reads with TranSigner. *Genome Biol* 2025;26:257. https://doi.org/10.1186/s13059-025-03723-2
66. Keil N, Monzó C, McIntyre L *et al.* Quality assessment of long read data in multisample lrRNA-seq experiments using SQANTI-reads. *Genome Res* 2025;35:987–98. https://doi.org/10.1101/gr.280021.124
67. Pardo-Palacios FJ, Arzalluz-Luque A, Kondratova L *et al.* SQANTI3: curation of long-read transcriptomes for accurate identification of known and novel isoforms. *Nat Methods* 2024;21:793–7. https://doi.org/10.1038/s41592-024-02229-2
68. Öztürk-Çolak A, Marygold SJ, Antonazzo G *et al.* FlyBase: updates to the *Drosophila* genes and genomes database. *Genetics* 2024;227:iyad211.
69. Boatwright JL, McIntyre LM, Morse AM *et al.* A robust methodology for assessing differential homeolog contributions to the transcriptomes of allopolyploids. *Genetics* 2018;210:883–94. https://doi.org/10.1534/genetics.118.301564
70. Liu C, Wang Y-G. Does one subgenome become dominant in the formation and evolution of a polyploid? *Ann Bot* 2023;131:11–6. https://doi.org/10.1093/aob/mcac024
71. Zhu Y, Watson C, Safonova Y *et al.* CloseRead: a tool for assessing assembly errors in immunoglobulin loci applied to vertebrate long-read genome assemblies. *Genome Biol* 2025;26:131. https://doi.org/10.1186/s13059-025-03594-7
72. Campo D, Lehmann K, Fjeldsted C *et al.* Whole genome sequencing of two North American Drosophila melanogaster populations reveals genetic differentiation and positive selection. *Mol Ecol* 2013;22:5084–97. https://doi.org/10.1111/mec.12468
73. King EG, Merkes CM, McNeil CL *et al.* Genetic dissection of a model complex trait using the Drosophila Synthetic Population Resource. *Genome Res* 2012;22:1558–66. https://doi.org/10.1101/gr.134031.111
74. Chen Y, Sim A, Wan Y *et al.* Context-aware transcript quantification from long-read RNA-seq data with Bambu. *Nat Methods* 2023;20:1187–95. https://doi.org/10.1038/s41592-023-01908-w
75. Mathé C, Dunand C. Automatic prediction and annotation: there are strong biases for multigenic families. *Front Genet* 2021;12:697477.
76. Meyer C, Scalzitti N, Jeannin-Girardon A *et al.* Understanding the causes of errors in eukaryotic protein-coding gene prediction: a case study of primate proteomes. *BMC Bioinformatics* 2020;21:513. https://doi.org/10.1186/s12859-020-03855-1
77. Salzberg SL. Next-generation genome annotation: we still struggle to get it right. *Genome Biol* 2019;20:92. https://doi.org/10.1186/s13059-019-1715-2
78. Han R, Han L, Zhao X *et al.* Haplotype-resolved genome of Sika deer reveals allele-specific gene expression and chromosome evolution. *Genomics Proteomics Bioinformatics* 2023;21:470–82. https://doi.org/10.1016/j.gpb.2022.11.001
79. Yang C, Zhou Y, Song Y *et al.* The complete and fully-phased diploid genome of a male Han Chinese. *Cell Res* 2023;33:745–61. https://doi.org/10.1038/s41422-023-00849-5
80. Healey AL, Garsmeur O, Lovell JT *et al.* The complex polyploid genome architecture of sugarcane. *Nature* 2024;628:804–10. https://doi.org/10.1038/s41586-024-07231-4
81. Zheng Y, Yang D, Rong J *et al.* Allele-aware chromosome-scale assembly of the allopolyploid genome of hexaploid Ma bamboo (Dendrocalamus latiflorus Munro). *J Integr Plant Biol* 2022;64:649–70. https://doi.org/10.1111/jipb.13217
82. Chen N, Solomon B, Mun T *et al.* Reference flow: reducing reference bias using multiple population genomes. *Genome Biol* 2021;22:8. https://doi.org/10.1186/s13059-020-02229-3
83. Coombes B, Lux T, Akhunov E *et al.* Introgressions lead to reference bias in wheat RNA-seq analysis. *BMC Bioinformatics* 2024;22:56. https://doi.org/10.1186/s12915-024-01853-w
84. Yuan S, Johnston HR, Zhang G *et al.* One size doesn't fit all - RefEditor: building personalized diploid reference genome to improve read mapping and genotype calling in next generation sequencing studies. *PLoS Comput Biol* 2015;11:e1004448. https://doi.org/10.1371/journal.pcbi.1004448
85. Chandra G, Gibney D, Jain C. Haplotype-aware sequence alignment to pangenome graphs. *Genome Res* 2024;34:1265–75. https://doi.org/10.1101/gr.279143.124
86. Baker TM, Waise S, Tarabichi M *et al.* Aneuploidy and complex genomic rearrangements in cancer evolution. *Nat Cancer* 2024;5:228–39. https://doi.org/10.1038/s43018-023-00711-y
87. Cortés-Ciriano I, Lee JJ-K, Xi R *et al.* Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. *Nat Genet* 2020;52:331–41. https://doi.org/10.1038/s41588-019-0576-7
88. Drews RM, Hernando B, Tarabichi M *et al.* A pan-cancer compendium of chromosomal instability. *Nature* 2022;606:976–83. https://doi.org/10.1038/s41586-022-04789-9
89. Garribba L, De Feudis G, Martis V *et al.* Short-term molecular consequences of chromosome mis-segregation for genome stability. *Nat Commun* 2023;14:1353. https://doi.org/10.1038/s41467-023-37095-7
90. Klockner TC, Campbell CS. Selection forces underlying aneuploidy patterns in cancer. *Mol Cell Oncol* 2024;11:2369388. https://doi.org/10.1080/23723556.2024.2369388
91. Network TCGA. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* 2015;517:576–82. https://doi.org/10.1038/nature14129
92. Chang C-H, Gregory LE, Gordon KE *et al.* Unique structure and positive selection promote the rapid divergence of Drosophila Y chromosomes. *eLife* 2022;11:e75795. https://doi.org/10.7554/eLife.75795
93. Courret C, Larracuente AM. High levels of intra-strain structural variation in *Drosophila simulans* X pericentric heterochromatin. *Genetics* 2023;225:iyad176. https://doi.org/10.1093/genetics/iyad176
94. Logsdon GA, Rozanski AN, Ryabov F *et al.* The variation and evolution of complete human centromeres. *Nature* 2024;629:136–45. https://doi.org/10.1038/s41586-024-07278-3
95. Porubsky D, Eichler EE. A 25-year odyssey of genomic technology advances and structural variant discovery. *Cell* 2024;187:1024–37. https://doi.org/10.1016/j.cell.2024.01.002
96. Di Tommaso P, Chatzou M, Floden EW *et al.* Nextflow enables reproducible computational workflows. *Nat Biotechnol* 2017;35:316–9.

Received: January 5, 2026. **Revised:** May 11, 2026. **Accepted:** May 25, 2026