BMC Genomics

RESEARCH

Open Access

Disregarding multimappers leads to biases in the functional assessment of NGS data

Michelle Almeida da Paz[1], Sarah Warger[1] and Leila Taher[1*]

Abstract

Background Standard ChIP-seq and RNA-seq processing pipelines typically disregard sequencing reads whose origin is ambiguous ("multimappers"). This usual practice has potentially important consequences for the functional interpretation of the data: genomic elements belonging to clusters composed of highly similar members are left unexplored.

Results In particular, disregarding multimappers leads to the underrepresentation in epigenetic studies of recently active transposable elements, such as AluYa5, L1HS and SVAs. Furthermore, this common strategy also has implications for transcriptomic analysis: members of repetitive gene families, such the ones including major histocompatibility complex (MHC) class I and II genes, are under-quantified.

Conclusion Revealing inherent biases that permeate routine tasks such as functional enrichment analysis, our results underscore the urgency of broadly adopting multimapper-aware bioinformatic pipelines –currently restricted to specific contexts or communities– to ensure the reliability of genomic and transcriptomic studies.

Keywords Next-generation sequencing (NGS), ChIP-seq, RNA-seq, Multimappers, Functional analysis

Background

Next-generation sequencing (NGS) technologies such as Chromatin Immunoprecipitation followed by sequencing (ChIP-seq) [1] and RNA-seq [2, 3] have emerged as the state of the art for obtaining insights into gene regulatory processes. ChIP-seq and RNA-seq sequencing reads are typically short, with customary protocols recommending 1×50 bp and 2×75 bp, respectively [4, 5]. Such read lengths are insufficient to completely span many of the repetitive elements that abound in complex eukaryotic genomes. Consequently, standard analysis pipelines struggle to unambiguously trace the locus from which the reads have arisen and fail to quantify closely related sequences of the genome.

*Correspondence:
Leila Taher
leila.taher@tugraz.at
[1] Institute of Biomedical Informatics, Graz University of Technology, Graz, Austria

The challenge of assigning reads which map equally well to multiple loci in the genome has been discussed for over a decade. Already in the early days of NGS, Chung et al. [6] acknowledged that in ChIP-seq data 32% of human STAT1 and 74% of mouse GATA1 binding sites ("peaks") were unlikely to be detected when ambiguously mapping reads ("multimappers") were discarded from the analysis. Similarly, while trying to determine the range of detection of RNA-seq, Mortazavi et al. [7] found that 13–24% of the 25 bp-long reads obtained after sequencing transcriptomic libraries from mouse brain, liver and skeletal muscle tissues were multimappers, and suggested that discarding multimappers would result in a severe underestimation of genes with closely related paralogs, such as the members of the ubiquitin B family (97% of the reads that map to members of this family are multimappers). Furthermore, scientists working on the function and evolution of repetitive elements, particularly on transposable elements (TEs), have often expressed their concerns about most studies disregarding more than half

of the human genome [8–10], and proposed several strategies to alleviate the problem [9, 11].

While various computational strategies have been proposed to mitigate the challenges posed by multimappers, to the best of our knowledge, no established NGS data processing pipeline offers an entirely satisfactory solution. In general, most strategies make prior assumptions about the distribution of the reads, and calculate the probability of a multimapper mapping to each of its possible target loci using a specific statistical model. Most strategies assume that multimappers and uniquely mapping reads ("unimappers") are similarly distributed across the genome/transcriptome, and that loci/transcript segments with high unimapper coverage (e.g., [6, 7, 12–14]) or enriched for reads relative to, for example, what could be expected under a random distribution (e.g., [15]), are the most likely source of those multimappers. Some tools also incorporate information on the likelihood of sequencing errors and variations (e.g., [16]). Nevertheless, recent studies have shown that multimappers are concentrated into a few regions of the genome with especially poor unimapper coverage [17], and therefore their distribution does not match the distribution of unimappers. Consistent with this observation, to estimate the likelihood of a multimapper's origin some tools rely solely on attributes such as the mapping quality of the reads (e.g., [17, 18]) or the sequence similarity between the potential loci of origin of the reads (e.g., [19]), as well as on the proportion of multimappers shared between the potential loci/transcript segments of origin (e.g., [19]). Moreover, substantial effort has been invested into developing strategies that make minimal or no prior assumptions about the data. These strategies acknowledge multimappers by distributing them equally among all loci (e.g., [20, 21]) or randomly selecting one of their mappings (e.g., filtering for the secondary alignment flag using samtools-view [22]), and have been used as an ultimate solution for resolving ties when the mapping quality scores among multimapper's mappings are equally good (e.g., Bowtie2 with "-k" option [23]). A comprehensive review of available tools is out of scope of this work, and can be found in the literature (e.g., [9, 11]).

Typical repetitive elements in many genomes, including the human genome, include TEs, tandem repeats, and satellite and microsatellite DNA. But also members of certain gene families, such as the globin gene family, homeobox genes and the olfactory receptors, exhibit strong sequence similarity [24–26]. Consistently, it has been noted that the expression of highly repetitive members of the ubiquitin family [7] and HLA class II beta chain paralogues, specifically, *HLA-DRB5* [18], can be underestimated by the practice of discarding multimappers. Unfortunately, despite the evident issue, standard transcriptomic pipelines, including the ones introduced by the ENCODE Project Consortium [27], disregard multimappers by default [28]. Not surprisingly, 87% (27 out of 31) of the articles recently published in the high-impacted journals Nature, Nature Genetics, Science, and Cell that report on the findings of ChIP-seq or RNA-seq data analyses do not acknowledge multimappers, while the remaining ones only partially recognize them, for example, by considering at most 10 of the mapping loci or requiring a minimum mapping quality score (Additional file 1: Suppl. Table 1).

With the present study, we aim to draw attention to biases in the functional interpretation of NGS data that result from disregarding multimappers. We demonstrate the problem by comparing the strategy used by standard NGS pipelines (e.g., ENCODE Project Consortium [27]), which simply filter out multimappers, to simple "multimapper-aware" approaches [9]. Our contribution is not to provide a definitive solution for the problem, but rather, to demonstrate its potential functional-level implications. Specifically, we analysed 9 ChIP-seq and 16 RNA-seq datasets for a small but diverse group of human and mouse cell types and experimental conditions (i.e., targeted protein or histone modification for ChIP-seq data, different treatments and replicates for RNA-seq data). In conclusion, we urge for the implementation of strategies accounting for multimappers in NGS pipelines.

Methods

Literature search

A PubMed search was carried out to identify articles accounting for multimappers. The following terms were used for the query: (ChIP-seq[tiab] OR RNA-seq[tiab]) OR (ChIP-seq[MeSH Terms] OR RNA-seq[MeSH Terms]) AND ("Nature"[Journal] OR "Nat Genet."[Journal] OR "Cell"[Journal] OR "Science"[Journal]). A filter for publication date was applied for the period from 2022/8/01 to 2023/8/31.

If not specified, we assumed that the tools had been run with default parameters (Additional file 1: Suppl. Table 1).

Datasets

We selected four human and mouse datasets from the ENCODE Project data repository [27] for single-end ChIP-seq and pair-end RNA-seq with read (or read pair) length ranging from 50 to 101 bp (Additional file 1: Suppl. Table 2).

Repeat annotation

Repeat annotation was obtained from the RepeatMasker track of the UCSC Genome Browser [29]. Immediately adjacent or overlapping annotations for TEs with the

same "name" ("repName" in the RepeatMasker track) were merged. We further refer to all TEs with the same name as a TE "group".

Quality control and read mapping

Quality of raw ChIP-seq and RNA-seq samples was assessed using FASTQC v0.11.9 [30]. Reads were trimmed for adapters with Cutadapt 2.8 [31] and filtered with Trimmomatic v0.39 [32]. Bwa mem v0.7.17 [33] and BBMap v39.01 [34] were used to map reads against the human (GRCh38/hg38) and mouse (GRCm38/mm10) genome assemblies for ChIP-seq; STAR v2.7.10a [35] was used for RNA-seq. Gene annotations (GRCh38.p13 and GRCm38.p4) were obtained from GENCODE [36]. Duplicated reads were filtered out using PICARD v2.24.0 [37]. Reads mapping to non-chromosomal scaffolds and mitochondrial chromosome were excluded from the analysis of ChIP-seq samples. Only reads mapped in a proper pair were considered for RNA-seq data analysis; they were retrieved with SAMtools v1.10 [22]. The parameters used for each tool are listed in Additional file 1: Suppl. Table 3.

TE group age

The oldest clade in which the TEs from a given group can be assumed to have been active was retrieved from Dfam ("Clades" column, [38]).

TE group coverage

Bedmap v2.4.37 [39] was used to identify overlaps between the coordinates of read mappings and annotated TEs. Reads that mapped only once in the genome were considered "unimappers"; reads that mapped more than once were considered "multimappers". Read coverage was computed for each TE group as:

$$C_K = \sum_{k \in K} \sum_{r \in Q} \sum_{r_i \in M_r} \frac{I_k(r_i)}{|M_r|} \left(\frac{l_{k_{r_i}}}{L_r} \right)$$

where K is the set of all copies of a TE group, Q is the set of all reads in the library, M_r is the set of all loci to which read r (of length L_r) mapped and $|M_r|$ is the size of that set, and $l_{k_{r_i}}$ is the number of nucleotides of the ith mapping of read r_i, overlapping with TE copy k . For each mapping r_i of r

$$I_k(r_i) = \begin{cases} 1, \ if \ r_i \text{ overlaps with } k \\ 0, \ otherwise \end{cases}.$$

Gene expression quantification

Multimappers were defined as read pairs ("fragments") for which at least one read of the pair mapped more than once in the genome.

Standard gene expression quantification was performed with HTSeq-count (v2.0.2, [40]) using default parameters ("--nonunique none"), i.e., not accounting for multimappers. The expression value of a gene g was defined as H_g/L_g, where H_g is the count for gene g assigned by HTSeq-count, and L_g is the gene length as defined by its start and end coordinates in the R Ensembl BioMart database v2.54.0 [41].

To account for multimappers, we used a "multimapper-aware" strategy that counted fragments in genes based on the list of genes ("set S") overlapping with the fragment mappings generated by HTSeq-count [42]. Specifically, gene counts were computed for each gene g as:

$$C_g = \sum_{f \in Q} \sum_{f_i \in M_f} \frac{I_g(f_i)}{|M_f|}$$

where Q is the set of all fragments in the library, M_f is the set of all mappings in the transcriptome for fragment f and $|M_f|$ is the size of that set, and for each mapping f_i of f

$$I_g(f_i) = \begin{cases} 1, \text{ if } f_i \text{ overlaps with } g \\ 0, \ otherwise \end{cases}.$$

Note that if f_i overlaps not only with g but also with at least another gene, then $I_g(f_i) = 0$. This is the default behaviour of HTSeq-count (Additional file 3: Additional Material).

A gene g was considered "expressed" if $C_g > 0$. The multimapper-aware expression value of gene g was defined as C_g/L_g.

Genes were considered under-quantified by HTSeq-count if $\frac{C_g/L_g}{H_g/L_g} > 2$, where H_g is the count for gene g assigned by HTSeq-count.

Computations were repeated with simulated libraries constructed by trimming the 3' end of the read pairs to 25, 50 or 75 bp with Cutadapt 2.8 [31].

Functional analysis

The 50, 100 or 200 protein-coding genes with the highest expression values were subjected to functional analysis using the "compareCluster()" function of the R clusterProfiler package (v.4.6.0, [43]). Gene type was retrieved from R Ensembl BioMart database v2.54.0 [41].

Gene set enrichment analysis (GSEA)

GSEA [44] was conducted on the fold-changes between the normalised counts for all protein-coding genes computed with HTSeq-count and those computed with our multimapper-aware strategy using the "GSEA()" function of the clusterProfiler with 10,000 permutations and the C7: Immunologic Signatures" collection from the Human

Molecular Signatures Database (MSigDB) (v.2023.2, [45]). Gene counts were normalised with the trimmed mean of M-values (TMM) method, by calculating scaling factors using the "calcNormFactors()" and "cpm()" functions of the edgeR package (v.3.40.2, [46]) with default parameters.

Differential expression analysis

Differential expression analysis was performed for the mouse RNA-seq dataset (Additional file 1: Suppl. Table 2) using the DESeq2 (v.1.38.3, [47]) R package. Specifically, the "DESeq()" function was used with default options to compare gene expression across all time points (1 h, 2 h, 4 h, 6 h) following lipopolysaccharide treatment, relative to the untreated control group (0 h). Gene counts from the multimapper-aware strategy were rounded to the nearest integer. Genes with a false discovery rate (FDR) lower than 0.05 and a $\log_2$ fold-change lower than -1 (down-regulated) or greater than 1 (up-regulated) were considered differentially expressed. The "lfcShrink()" function was used for fold-change shrinkage, using the "apeglm" method.

Results

Inspection of exemplary ChIP-seq ENCODE [27] libraries (Datasets 1 and 2; Additional file 1: Suppl. Table 2) revealed that multimappers constitute a substantial proportion (9–51%) of all reads mapped to the human genome, although the exact numbers vary greatly depending on the mapping tool –Bwa mem [33] (26–32%) reported twice or more the number of multimappers than BBMap [34] (9–16%) – and the immunoprecipitated protein (22–51%) (Additional file 2: Suppl. Fig. 1). Counterintuitively, when adhering to the current working standards and guidelines for ChIP-seq, we observed that the read length had only a relatively modest influence on the proportion of multimappers. Specifically, extending the read length from 50 to 100 bp resulted in a 17% reduction when utilising BWA mem and a 40% reduction with BBMap (Fig. 1A). As expected, a large fraction (43–80%) of multimappers mapped to regions annotated as TEs. Motivated by this fact and by the enormous expansion of repetitive TE sequences in mammalian genomes –they comprise ~ 46% of the human genome–, we used TEs to explore the impact of multimappers on an epigenetic analysis based on ChIP-seq data. TE individual copies in the human genome vary widely in length, from 10 (e.g., members of L2a) to 153,104 bp (nested LTR12B), but span a median of 231 bp, mostly reflecting the relatively recent expansion of elements from the SINE Alu family (median of 294 bp, Fig. 1B). Thus, although not all TEs give rise to multimappers, the large fraction of multimappers derived from TEs can be explained by the fact that TE copies are not fully covered by conventional NGS reads. In the datasets included in this study, at least 70% of the reads mapping to 8–16% (up to 181 out of 1,160) of TE groups are multimappers (Methods). Specifically, when considering every possible mapping, multimappers tended to be associated with evolutionary *young* TEs, such as AluYa5, L1HS and SVAs, while unimappers were associated with *old* TEs (P-value $< 2.2 \times 10^{-16}$, Chi-squared test; Fig. 1C; Additional file 2: Suppl. Figs. 2–4). And although with some deviations in TE group coverage (e.g., 3–55% for HERV-Fc1_LTR2), we made similar observations when considering only a random mapping for each multimapper (Additional file 1: Suppl. Table 4). This is natural, since relatively young TEs have not had enough time to accumulate variations in their sequences, but has far-reaching consequences: using standard ChIP-seq pipelines will specifically underrepresent recently active TEs, hampering their study. Nevertheless, TE activity is known to be associated with diverse human diseases [48], and hence, rectifying this issue promptly is imperative.

ChIP-seq is not the only NGS technology concerned by the current prevailing approach to handling multimappers. Standard RNA-seq bioinformatic pipelines use tools such as HTSeq-count [40] and STAR [35] for quantifying the reads mapping to annotated genes, and these tools also deliberately disregard multimappers. Although multimappers are not as abundant in RNA-seq data as they are in ChIP-seq data, they are not negligible. Using human and mouse RNA-seq dendritic cell libraries to illustrate the problem, we found that ~ 10% (Fig. 2A) and ~ 5% of the fragments mapped to the human and mouse genomes, respectively, were multimappers (Additional file 2: Suppl. Fig. 5). Similarly to what we observed for ChIP-seq, the read length had a relatively modest influence on the proportion of multimappers. More precisely, for paired-end RNA-seq, increasing the read length from 50 to 100 bp resulted in a 28% reduction. Moreover, analysis using HTSeq-count and STAR geneCounts with default parameters revealed quantification differences for about 6% (777 out of 13,437) of the human and 4% (468 out of 12,561) of the mouse genes expressed in these cells compared to a simple multimapper-aware strategy (Methods). Specifically, these genes were underquantified by HTSeq-count and STAR geneCounts (Fig. 2B; Additional file 1: Suppl. Tables 5 and 6; Additional file 2: Suppl. Fig. 6), and most notably, they were not just random genes, but actually related to functions intrinsic to the biology of the samples under investigation, such as MHC class I and II immune responses and peptide antigen binding (Fig. 2C; Additional file 2: Suppl. Figure 7). GSEA analysis comparing the gene expression values from HTSeq-counts to those obtained with the

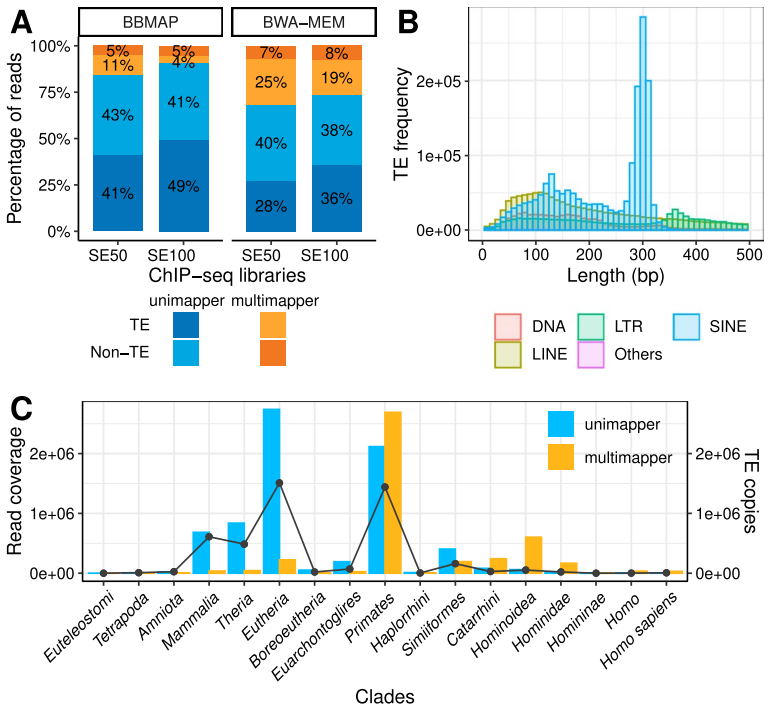

Fig. 1 Discarding multimappers leads to epigenetic mischaracterization of young TEs. **A** Percentage of uni- and multimapper reads mapping to portions of the human genome annotated as TEs and not annotated as TEs (non-TE) for dataset 1 containing two ChIP-seq libraries generated by the ENCODE consortium using single-end 50 bp ("SE50") and 100 bp ("SE100") reads. Mapping was performed with two different mapping tools: with BBMap and Bwa mem. TEs are the major source of ChIP-seq multimappers in the human genome. **B** Length distribution of TE individual copies. Only TEs shorter than 500 bp are shown. Note that ~ 74% (856 out of 1,160) of the TEs in the human genome are longer than 500 bp, spanning up to 153,104 bp. The bin width is 10 bp. TEs were classified as DNA, LTR (long terminal repeat), SINE (short interspersed nuclear element), LINE (long interspersed nuclear element) and Others (e.g., rolling-circle (RC), unknown classification). Standard NGS reads are too short to fully cover most TE copies, explaining why TEs often give rise to multimappers. **C** Read coverage (bar plot on the left y-axis; see Methods) per clade for the SE100 ChIP-seq library (dataset 1) for uni- and multimappers. Reads were mapped using Bwa mem. TEs found in multiple clades (e.g., L1HS and L1P1) were assigned to the "younger" clade (*Homo* and *Hominoidea*, respectively). Only 30 out of 1,160 different TEs in the human genome (~ 3%) have not been annotated to any clade and were not represented. The number of TE copies for each clade (line plot on the right y-axis) shows that the majority of TEs are *Primates* and *Eutherian*-specific. Evolutionary young TEs are prone to be underrepresented when excluding multimappers from ChIP-seq analysis

multimapper-aware strategy further supported the association with perturbations of the immune system (Additional file 2: Suppl. Fig. 8). Perhaps more critically, these quantification biases can also impact the identification of differentially expressed genes. Indeed, we observed a substantial number of genes (9–81) that were exclusively differentially expressed when quantified with HTSeq-counts compared to the multimapper-aware strategy, or vice versa. Additionally, these genes were enriched for antigen presentation molecular functions (Additional file 2: Suppl. Figures 9 and 10). Naturally, RNA-seq dendritic cell libraries are no exception. Thus, for a collection of RNA-seq libraries of human lung carcinoma treated with three different drugs (i.e., dexamethasone, hydrocortisone or mapracorat), we found multimappers represented 6–9% of the fragments mapped to the genome, and 6–7% of the expressed genes were under-quantified when discarding multimappers (Additional file 2: Suppl. Fig. 11). In line with our findings in dendritic cells, accounting for multimappers resulted in differences in functional analysis, although in this case, the discrepancies were smaller (Additional file 1: Suppl. Table 7, Additional file 2: Suppl. Fig. 12 and 13). In essence, contingent upon the characteristics of the gene families expressed in the sample of interest, disregarding multimappers during the analysis of RNA-seq data may severely hinder the identification of critically relevant biological functions and processes.

Discussion

For over a decade, scientists have grappled with the challenge of unambiguously assigning a substantial fraction of NGS short-reads to their original genomic loci. Most standard NGS pipelines filter out reads whose origin is ambiguous. Over time, numerous computational strategies have been proposed to acknowledge multimappers.

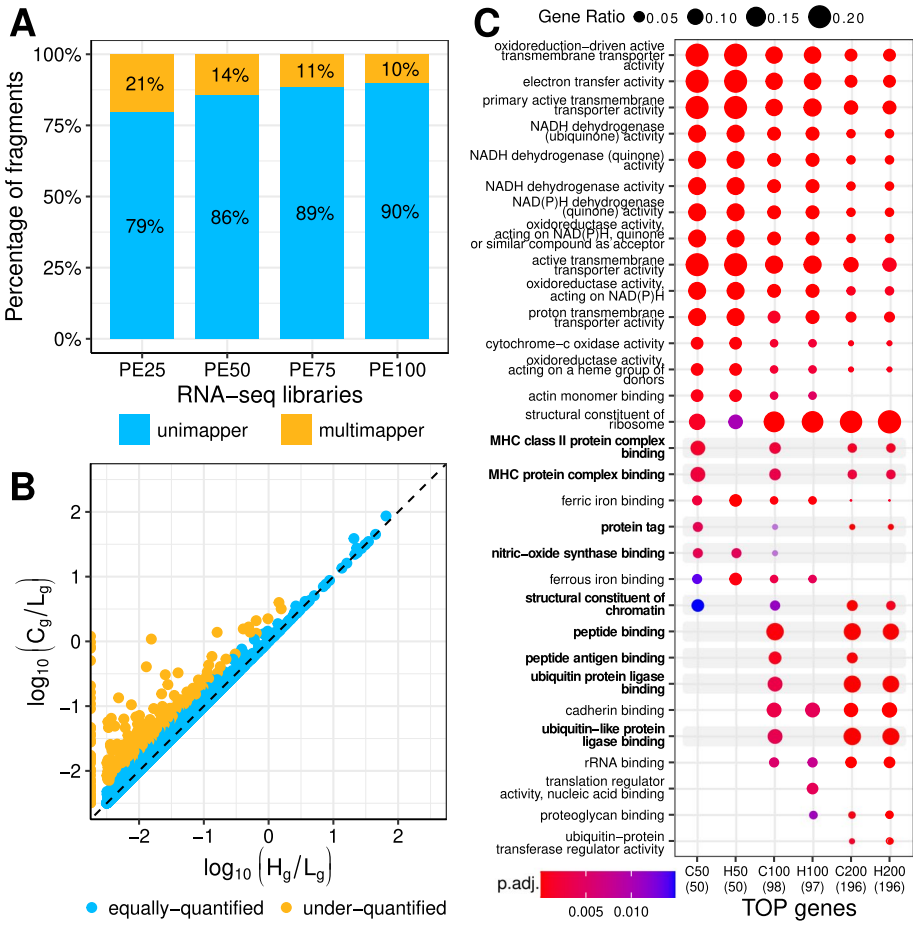

Fig. 2 Discarding multimappers leads to functional mischaracterization of repetitive gene families. **A** Percentage of uni- and multimapper fragments mapping to human genome for a RNA-seq library of dataset 3 generated by the ENCODE consortium using pair-end 100 bp ("PE100") and thereof simulated libraries with read pairs of length 25, 50 and 75 bp ("PE25", "PE50" and "PE75", respectively). Within the read lengths assessed, the difference in the proportion of multimappers was modest (10–21%). **B** Scatter plot showing gene expression values computed with HTSeq-count using default parameters ("–nonunique none"; x-axis) and by our "multimapper-aware" strategy (y-axis) for PE100. Each dot represents a protein-coding gene and is coloured differently depending on whether it is considered (approximately) equally-quantified or under-quantified by HTSeq-count (see Methods). The dashed line indicates identical gene expression values. About 6% (777 out of 13,437) expressed genes are under-quantified when discarding multimappers. **C** Gene ontology (GO) enrichment analysis of the 50, 100, and 200 protein-coding genes with the highest expression values in PE100 as computed by HTSeq-count ("H50", "H100", and "H200", respectively) or our "multimapper-aware" strategy ("C50", "C100", and "C200", respectively). GO enrichment analysis was performed for the "molecular function" category and using the "org.Hs.eg.db" annotation for the human genome. The q-value threshold was set to 0.01. Dot size represents the ratio between the number of genes in the given GO term (y-axis) and the number of genes annotated in each category (shown in brackets, below the label of each gene set on the x-axis). Dot colour indicates the *P*-value adjusted by Benjamini-Hochberg (BH, "p.adj."). Neglecting multimappers leads to the underrepresentation of genes associated with specific GO terms (indicated in bold and with grey shading)

All of them make prior assumptions, mostly about the distribution of reads in the NGS data (e.g., [6, 7, 12–14]). These assumptions have been naturally accepted as valid and only recently, it has been questioned whether they are valid [17]. To date, no gold-standard NGS pipeline exists that completely resolves the problem. It is up to the researchers to decide which assumptions are reasonable, and ultimately which biases are acceptable.

In this study, we investigated the implications of how multimappers are processed for the functional analysis of NGS data using two "multimapper-aware" approaches. One of the approaches accounts for multimappers by dividing the number of reads assigned to a locus/transcript segment by the total number of loci/transcript segments to which the reads map. The second approach randomly selects a mapping for each multimapper from the set of all multimapper's mappings. The main advantage of these strategies is that they rely on a parsimonious set of assumptions, which makes them simple, intuitive and widely applicable. This alignment with Occam's

Razor reduces uncertainty and promotes generalizability. Meanwhile, the vast majority of strategies proposed in the literature allocate multimappers to a locus/transcript segment according to the distribution of unimappers or based on read mapping quality scores. While these strategies may seem more appealing and well-suited at first glance, they are flawed. First and foremost, the underlying assumptions are not necessarily true [17]. But even if multimappers could be assumed to be distributed as unimappers, and hence, the unimapper coverage served as a good indicator for inferring the loci of origin of multimappers, the low fraction of unimappers mapping to genomic elements like the members of the ubiquitin B family –only 3% of the reads mapping to these genomic elements are unimappers [7]– would make this inference unreliable. Similarly, the probability that a read is incorrectly mapped alone is insufficient for determining the locus of origin of a multimapper, as it can be easily influenced by other factors, such as sequencing errors and sequence variants. Naturally, despite their advantages, straightforward solutions to the problem such as the two "multimapper-aware" strategies we employed also have their limitations, as they may not comprehensively capture the entirety of the complexity of the problem. In particular, the approaches we adopted are suitable for calculating read coverage at the level of TE group/gene families, but do not allow to quantify how much a TE copy/transcript is expressed in comparison to other TE copies/transcripts of the same TE group/gene family [11].

Using exemplary random ChIP-seq and RNA-seq datasets, we showed that discarding multimappers can lead to biases in functional genomic/transcriptomic analyses. As might be expected, the magnitude of the biases varies with the dataset and the impact is more pronounced when shorter reads are used. Prevailing NGS platforms like Illumina produce massive quantities of highly accurate sequencing reads, but these reads are relatively short. Generally, we noted that the proportion of multimappers is determined by the read length. However, consistent with previous observations [49, 50] we found that the use of longer or paired-end Illumina reads does not result in substantial differences. Moreover, our findings suggest that the fraction of multimappers mapped to regions annotated as TEs and members of repetitive gene families also depends on the mapping tool (previously observed by [51]), targeted proteins or histone modifications, and treatment. Furthermore, also the identity of the TEs and genes most affected by the way multimappers are handled depends on the aforementioned factors, suggesting that the biases stemming from the practice of discarding multimappers may vary in severity, contingent upon the underlying biological context. Finally, it is important to note that the datasets illustrating the issue were chosen randomly. While an examination of a broader range of datasets may be warranted to unveil more nuanced trends, there is no compelling evidence to undermine the robustness of our general conclusion—that neglecting multimappers introduces biases in the functional analysis of NGS data.

Interpreting the functional significance of expressed (or differentially expressed) genes is often a primary goal in RNA-seq analysis. When genes exhibit sufficiently high sequence similarity, the practice of discarding multimappers is likely to affect the quantification of paralogous gene families, genes with internally repeated domains, and multiple isoforms of the same gene. Notable examples of such groups of genes include HLA class I (e.g., *HLA-B*, *HLA-E*) and class II (e.g., *HLA-DRA*, *HLA-DPA1*), polyubiquitin genes (e.g., *UBB*, *UBC*), chromatin (e.g., *MRNIP*) and cytoskeleton (e.g., *TUBB, TUBB2B*) components, and the recently discovered *BOLA2B* genes. It is worth mentioning that biases in gene quantification impact differentially expression analysis as well, potentially leading to both false positives and false negatives. These effects are likely to be exacerbated if poorly expressed genes are filtered out before testing for differential expression, a common practice.. Since in this study we have only focused on mRNA, we anticipate that many other types of RNA (e.g., miRNAs, rRNAs, tRNAs) present in multiple copies might be underestimated when discarding multimappers. Importantly, functional analysis may not always reveal underrepresented functions, as we found for dendritic cell libraries, in which functions related to adaptive immunology were well associated with multimappers, but can be minor, as for the analysed lung cancer libraries. Researchers had previously identified isolated instances of genes exhibiting varying quantification results depending on how multimappers were handled. Here, we demonstrate that these effects extend beyond individual genes and manifest at the functional level.

Despite its intuitive nature, the problem posed by multimappers and their impact on functional NGS analysis are routinely disregarded by standard bioinformatics pipelines. This oversight results in the neglect of clusters of repetitive genomic elements with highly similar members. We therefore believe that addressing the problem outlined in this study may entail applying one or more multimapper-aware strategies and contrasting their results with those of strategies that do not account for multimappers. Furthermore, emerging NGS technologies such as PacBio and Oxford Nanopore enable the acquisition of ultra-long reads, having already reached the impressive mark of more than 2 Mbp [52], and thus, hold the potential to substantially reduce the number of ambiguously mapping reads. However, they are currently

limited by their higher cost and lower accuracy when compared to Illumina NGS. An alternative to achieve the desired outcome could be combining these two technologies. Ultimately, it becomes imperative that new computational guidelines acknowledging multimappers are established and disseminated by major projects.

Conclusions

Our research shows that neglecting multimappers during NGS data processing can have a substantial impact on biological inferences drawn from genomic and transcriptomic data. To the best of our knowledge no other article has explored the functional-level implications of this practice embedded in the ENCODE guidelines. Notably, we showed that the issue extends beyond specific scientific communities, such as those dedicated to the study of TEs. Indeed, our findings emphasise that even a seemingly routine task such as performing a gene ontology (GO) enrichment analysis on any given RNA-seq dataset can be susceptible to biases. And consequences are far-reaching: candidates identified for further functional assays may be considerably suboptimal.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-024-10344-9.

Additional file 1. Supplementary Tables 1–7.

Additional file 2. Supplementary Figs. 1–10.

Additional file 3. Additional Material and Additional Figs. 1–20.

Acknowledgements
Not applicable.

Authors' contributions
MAP: Methodology, Software, Visualization, Investigation, Formal analysis, Writing-Original draft preparation. SW: Software, Investigation. LT: Conceptualization, Methodology, Supervision, Writing-Reviewing and Editing. All coauthors have read and approved the final version of the manuscript.

Funding
Open access funding provided by Graz University of Technology. This research was funded in whole, or in part, by the Austrian Science Fund (FWF) [P33437] awarded to L.T. For the purpose of open access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

Availability of data and materials
ChIP-seq and RNA-seq datasets were acquired, processed, and analysed from datasets described in Additional file 1: Suppl. Table 2. Code generated and used during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 1 February 2024 Accepted: 24 April 2024
Published online: 08 May 2024

References

1. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods. 2007;4:651–7.
2. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10:57–63.
3. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science. 2008;320:1344–9.
4. Transcription Factor ChIP-seq Data Standards and Processing Pipeline. https://www.encodeproject.org/chip-seq/transcription_factor/. Accessed 1 Feb 2024.
5. Sequencing Read Length. https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/read-length.html. Accessed 1 Feb 2024.
6. Chung D, Kuan PF, Li B, Sanalkumar R, Liang K, Bresnick EH, et al. Discovering transcription factor binding sites in highly repetitive regions of genomes with multi-read analysis of ChIP-Seq data. PLoS Comput Biol. 2011;7:e1002111.
7. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008;5:621–8.
8. Lanciano S, Cristofari G. Measuring and interpreting transposable element expression. Nat Rev Genet. 2020;21:721–36.
9. O'Neill K, Brocks D, Hammell MG. Mobile genomics: tools and techniques for tackling transposons. Philos Trans R Soc Lond B Biol Sci. 2020;375:20190345.
10 Teissandier A, Servant N, Barillot E, Bourc'his D. Tools and best practices for retrotransposon analysis using high-throughput sequencing data. Mob DNA. 2019;10:52.
11. Deschamps-Francoeur G, Simoneau J, Scott MS. Handling multi-mapped reads in RNA-seq. Comput Struct Biotechnol J. 2020;18:1569–76.
12. Faulkner GJ, Forrest ARR, Chalk AM, Schroder K, Hayashizaki Y, Carninci P, et al. A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE. Genomics. 2008;91:281–8.
13. Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN. RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics. 2010;26:493–500.
14. Liu Y, Ma Y, Salsman E, Manthey FA, Elias EM, Li X, et al. An enrichment method for mapping ambiguous reads to the reference genome for NGS analysis. J Bioinform Comput Biol. 2019;17:1940012.
15. Newkirk D, Biesinger J, Chon A, Yokomori K, Xie X. AREM: aligning short reads from ChIP-sequencing by expectation maximization. J Comput Biol. 2011;18:1495–505.
16. Ji Y, Xu Y, Zhang Q, Tsui K-W, Yuan Y, Norris C Jr, et al. BM-map: Bayesian mapping of multireads for next-generation sequencing data. Biometrics. 2011;67:1215–24.
17. Shah RN, Ruthenburg AJ. Sequence deeper without sequencing more: Bayesian resolution of ambiguously mapped reads. PLoS Comput Biol. 2021;17:e1008926.
18. Consiglio A, Mencar C, Grillo G, Marzano F, Caratozzolo MF, Liuni S. A fuzzy method for RNA-Seq differential expression analysis in presence of multireads. BMC Bioinformatics. 2016;17(Suppl 12):345.
19. McDermaid A, Chen X, Zhang Y, Wang C, Gu S, Xie J, et al. A new machine learning-based framework for mapping uncertainty analysis in RNA-Seq read alignment and gene expression estimation. Front Genet. 2018;9:313.

20. Criscione SW, Zhang Y, Thompson W, Sedivy JM, Neretti N. Transcriptional landscape of repetitive elements in normal and cancer human cells. BMC Genomics. 2014;15:583.
21. Almeida da Paz M, Taher L. T3E: a tool for characterising the epigenetic profile of transposable elements using ChIP-seq data. Mob DNA. 2022;13:29.
22. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9.
23. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.
24. Efstratiadis A, Posakony JW, Maniatis T, Lawn RM, O'Connell C, Spritz RA, et al. The structure and evolution of the human beta-globin gene family. Cell. 1980;21:653–68.
25. Holland PWH, Booth HAF, Bruford EA. Classification and nomenclature of all human homeobox genes. BMC Biol. 2007;5:47.
26. Olender T, Lancet D, Nebert DW. Update on the olfactory receptor (OR) gene superfamily. Hum Genomics. 2008;3:87–97.
27. Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, et al. The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res. 2018;46:D794–801.
28. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. 2012;22:1813–31.
29. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome Res. 2002;12:996–1006.
30. Babraham Bioinformatics - FastQC a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed 1 Feb 2024.
31. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 2011;17:10.
32. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.
33. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.
34. Bushnell B. BBMap: A fast, accurate, splice-aware aligner. 2014.
35. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.
36. Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 2019;47:D766–73.
37. Picard. https://broadinstitute.github.io/picard/. Accessed 1 Feb 2024.
38. Storer J, Hubley R, Rosen J, Wheeler TJ, Smit AF. The Dfam community resource of transposable element families, sequence models, and genome annotations. Mob DNA. 2021;12:2.
39. Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK, et al. BEDOPS: high-performance genomic feature operations. Bioinformatics. 2012;28:1919–20.
40. Anders S, Pyl PT, Huber W. HTSeq–a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31:166–9.
41. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat Protoc. 2009;4:1184–91.
42. Counting reads in features with htseq-count — HTSeq 0.11.1 documentation. https://htseq.readthedocs.io/en/release_0.11.1/count.html. Accessed 1 Feb 2024.
43. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innovation (Camb). 2021;2:100141.
44. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102:15545–50.
45. Castanza AS, Recla JM, Eby D, Thorvaldsdóttir H, Bult CJ, Mesirov JP. Extending support for mouse data in the Molecular Signatures Database (MSigDB). Nat Methods. 2023;20:1619–20.
46. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–40.
47. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:1–21.
48. Hancks DC, Kazazian HH Jr. Active human retrotransposons: variation and disease. Curr Opin Genet Dev. 2012;22:191–203.
49. Chhangawala S, Rudy G, Mason CE, Rosenfeld JA. The impact of read length on quantification of differentially expressed genes and splice junction detection. Genome Biol. 2015;16:131.
50. Li W, Freudenberg J. Mappability and read length. Front Genet. 2014;5:381.
51. Oliva A, Tobler R, Cooper A, Llamas B, Souilmi Y. Systematic benchmark of ancient DNA read mapping. Brief Bioinform. 2021;22:bbab076.
52. Payne A, Holmes N, Rakyan V, Loose M. BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files. Bioinformatics. 2019;35:2193–8.

Publisher's Note