Published online 13 July 2022

NAR Genomics and Bioinformatics, 2022, Vol. 4, No. 3 **1**
https://doi.org/10.1093/nargab/lqac052

Quantification of mutant–allele expression at isoform level in cancer from RNA-seq data

Wenjiang Deng [1], Tian Mou[2], Yudi Pawitan[1,*] and Trung Nghia Vu [1,*]

[1]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden and [2]School of Biomedical Engineering, Shenzhen University, Shenzhen, China

Received May 10, 2021; Revised June 26, 2022; Editorial Decision July 01, 2022; Accepted July 04, 2022

ABSTRACT

Even though the role of DNA mutations in cancer is well recognized, current quantification of the RNA expression, performed either at gene or isoform level, typically ignores the mutation status. Standard methods for estimating allele-specific expression (ASE) consider gene-level expression, but the functional impact of a mutation is best assessed at isoform level. Hence our goal is to quantify the mutant–allele expression at isoform level. We have developed and implemented a method, named MAX, for quantifying mutant–allele expression given a list of mutations. For a gene of interest, a mutant reference is constructed by incorporating all possible mutant versions of the wild-type isoforms in the transcriptome annotation. The mutant reference is then used for the RNA-seq reads mapping, which in principle works similarly for any quantification tool. We apply an alternating EM algorithm to the read-count data from the mapping step. In a simulation study, MAX performs well against standard isoform-quantification methods. Also, MAX achieves higher accuracy than conventional gene-based ASE methods such as ASEP. An analysis of a real dataset of acute myeloid leukemia reveals a subgroup of NPM1-mutated patients responding well to a kinase inhibitor. Our findings indicate that quantification of mutant–allele expression at isoform level is feasible and has potential added values for assessing the functional impact of DNA mutations in cancers.

INTRODUCTION

The role of DNA mutations in the initiation and progression of cancers is well recognized. In the era of individualized medicine, the mutation profile in a specific patient is used as biomarker for prognosis and prediction of response to therapy (1,2). Most mutation sites are heterogeneous; in fact, this characteristic is one of the filters used in mutation callers such as Mutect (3). So, we can expect that both wild-type and mutant alleles are expressed. However, existing methods for the estimation of RNA expression from RNA-seq data (4–8), either at gene or isoform level, typically ignores the mutation status. Biologically, the impact of a mutation is likely to be mediated by the expression level of the mutant allele, it is informative to quantify the mutant–allele expression separately from the wild-type alleles. Hence, our goal is to develop, given a list of mutations, a method to estimate the mutant–allele expression based on the RNA-seq data.

Traditional gene-expression microarrays do not include multiple alleles due to DNA variants in a gene. RNA-seq data potentially contains information of the allelic heterogeneity. Analysis of allele-specific expression (ASE) has been commonly done for normal tissues, assuming there are polymorphic sites—usually single nucleotide polymorphisms (SNPs)—within the gene (9–11). Current algorithms to estimate ASE are generally gene-based and mostly based on individual samples (12,13). Moreover it is typically DNA-based, even when analysing RNA-seq data. For example, a recent study (9) employs the read counts-covering SNVs based on mapping to a genome reference rather than to a transcriptome reference. These two mappings in general produce conceptually distinct datasets, because only the latter contains all the known alternative transcripts. The quantification of ASE is used to assess allelic imbalance, which might be associated with phenotypic diversity including diseases among individuals. ASE is also related to the so-called expression quantitative-trait locus (eQTL), although in general eQTLs do not have to reside within a gene (14).

During its maturation process, for approximately 50% of human genes, the mRNA of a gene is alternatively spliced to produce potentially distinct transcripts (15,16). To be clear, the term ‘transcript’ naturally refers to a biological entity, while ‘isoform’ refers to a logical entity representing the RNA sequence of a transcript. However, when there is no danger of confusion, we use the terms interchangeably.

Alternative splicing is an important cellular mechanism to generate transcriptomic and phenotypic diversity. The

*To whom correspondence should be addressed. Email: yudi.pawitan@ki.se
Correspondence may also be addressed to Trung Nghia Vu. Email: trungnghia.vu@ki.se

functional impact of a mutation in a gene is best assessed at isoform level, since the codon-level translation of the RNA to amino-acids is only meaningful at isoform level. A specific mutation inside a gene may not even appear in some isoforms of the gene. We can see this, for instance, in NPM1 gene, which we describe in detail later. It has eight wild-type (WT) isoforms, but even though there are 14 distinct mutations in the gene (collected from 135 NPM1-mutated patients of the BeatAML cohort), two WT isoforms do not have any mutant versions. More generally, the mutation load and expression level of a mutant isoform might be different across the different mutant isoforms. Therefore, estimation of the mutant–allele expression at isoform level is important for investigating the mutation functional impact. A recent study implements the ideas of splice junctions and splice graphs to quantify isoform expression, which seems promising when the transcript reference is uncertain or incomplete (17).

Here, we describe a method for the quantification of Mutant–Allele eXpression—called MAX—based on RNA-seq data. To avoid extra complexities in the procedure, we assume that a list of mutations is available, either from the same or from an external set of samples. One can also use cancer mutations available in public databases such as COSMIC (18). When mutation data are available for the same set of samples, a more flexible analysis is possible; see the description of MAX2 below. Briefly, given a set of known wild-type isoforms in the transcriptome annotation and the list of mutations, we construct a new transcriptome annotation that contains the list of all candidate mutant isoforms. In principle, we could then use any previous tool available for isoform-level quantification, such as Sailfish (4), Kallisto (6) or Salmon (5). We found, however, that the estimation is challenging because there are substantial sequence similarities between the wild-type and mutant isoforms, and also amongst the mutant isoforms themselves. Our previously developed method XAEM (7) is more suitable for this purpose, particularly in identifying and merging isoforms that have close sequence similarities.

When compared using simulated data, MAX performs well against the procedures based on Salmon (5) and RSEM (8), and achieves higher accuracy than conventional gene-based ASE methods such as ASEP (11) . We apply MAX to analyse a real RNA-seq dataset ($n = 461$) of acute myeloid leukemia (AML) from the BeatAML project (19), including three of the most common mutations in AML: NPM1, FLT3 and TP53. The analysis reveals a subgroup of NPM1-mutated patients with low NPM1 mutant-expression that has a better drug response than those with high NPM1 mutant-expression. In summary, we have shown that quantification of the mutant–allele expression from RNA-seq data is feasible and has potential added values for assessing the functional impact of DNA mutations in cancer.

MATERIALS AND METHODS

Overview of MAX

Figure 1 shows the overview of the mutant–allele quantification using MAX. We start with the gene model from the standard transcription reference (panel A); we use the gene NAA20 for illustration. From a list of mutations in the exonic region, we construct a collection of mutant isoforms (panel B) to be added to the standard reference. This extended reference will be called 'WT+mutant Reference'. Once the reference is constructed, we map the RNA-seq data from multiple samples to produce read-count data y (panel C). The initial design matrix X for the mutated gene is then constructed following the method described in section below. The WT and mutant isoform expression is the parameter β in the Poisson model $y \sim X\beta$ and estimated using the alternating expectation-maximization (AEM) algorithm (D). The output is the read count for each wild-type and mutant isoform. For down-stream analyses that require comparisons across samples, a counts-per-million (CPM) value is calculated to normalize the counts relative to the library size in each sample. The AEM algorithm is described in more detail in (7), and the implementation of MAX is publicly available at https://github.com/wenjiangdeng/MAX/.

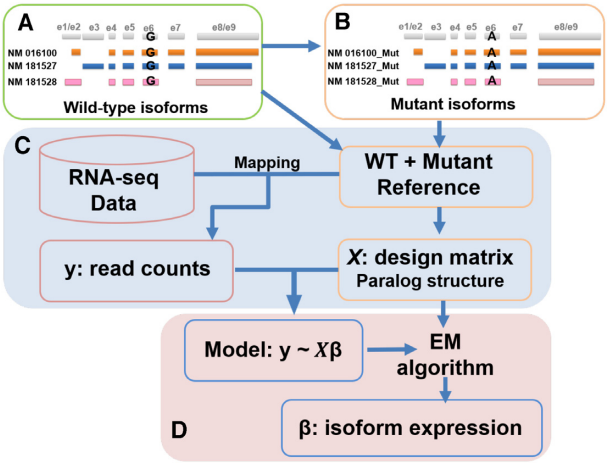

Figure 1. Overview of the mutant–allele expression quantification using MAX. (**A**) The collection of wild-type isoforms from a gene (NAA20 is used for illustration), and (**B**) the group of mutant isoforms as the basis of WT+mutant reference with the G/A mutation in exon e6. (**C**) Mapping of the RNA-seq data to the WT+mutant reference and the construction of initial design matrix X. (**D**) The quantification of isoform expression β and update of X using the AEM algorithm.

Isoform quantification model

The isoform quantification model has been described previously (7). Briefly, after mapping, the RNA-seq reads that map to a gene are summarized into a read-count vector y, which is assumed Poisson with mean μ. The most commonly used model is

$$\mu = X\beta, \tag{1}$$

where β is the vector of isoform expression values to be estimated. X is a design matrix which integrates multiple attributes such as isoform length, non-uniformity effects and RNA-seq biases. Mathematically, the elements of X translate the isoform-level expression into expected read-counts. X also summarizes the exon-sharing between isoforms (see the example below). In general we use an alternating EM algorithm, where in step (i) given y and an initial X, the parameter β is estimated using the EM algorithm, and in step

Table 1. The WT design matrix X for the gene NAA20

EqClass	NM_016100	NM_181527	NM_181528
001	0	0	0.40
010	0	0.26	0
100	0.04	0	0
101	0.10	0	0.10
110	0.48	0.41	0
111	0.38	0.33	0.50

(ii) given y and β, X is updated using the EM algorithm, and iterate back to (i) until convergence.

To construct the initial X matrix, we follow a computational scheme as described previously (7). Briefly, the R package Polyester (20) is used to simulate the RNA-seq data. The setting of the read length, mean and standard deviation of fragment length are obtained from real data. For each isoform, the simulated data have a read depth of 20×, which guarantees a high coverage of each nucleotide in the isoform. The RNA-seq reads are mapped to the reference transcriptome using the mapping tool Rapmap (21). Table 1 shows an example of the WT design matrix X for the *N*-acetyl transferase 5 (NAA20) gene, which contains three isoforms.

A total of six equivalence classes (eqClasses) are shared between the three wild-type isoforms. The entries in X indicate the proportions of reads distributed by each isoform to the different equivalent-classes. Each equivalence class (eqClass) is identified by a binary pattern that tells us which isoforms contribute to the eqClass. In RNA-seq mapped data processing, an eqClass represents all the reads that map to the same set of isoforms that define the eqClass. In some cases, an eqClass does represent a biological exon, but in general an eqClass represents only the sequence similarities across isoforms as implied by the transcriptome reference. We can see this by comparing the exon map in Figure 1A with Table 1. Analytical evaluation of X is possible, but extremely complicated. Therefore, in MAX we build it computationally from simulated mapped reads. In practice, we use Rapmap for a fast read alignment, which outputs an eqClass table (22). As shown in Supplementary Table S1, this table contains detailed information for each eqClass.

Construction of mutant X

To construct the initial X for a mutant gene, we need to consider the mutant versions of each wild-type isoform. If there are M WT isoforms and N distinct mutations in a gene, the potential number of mutant isoforms is up to $M \times 2^N$. The maximum occurs if each isoform contains all the mutation positions, and all possible present-absent combinations of mutations exist. Since the number of mutations is potentially large, for example the FLT3 gene in the BeatAML cohort contains 120 distinct mutations, it is not practical and sometimes not even feasible to code and quantify each possible mutant isoform separately. Statistically, the large X matrix will also lead to indeterminate solution in model (1). To avoid this problem, we need a more flexible way to integrate the mutations.

Specifically, we consider estimating only the sum of all the mutant isoforms associated with one WT isoform. Crucially, this implies that joint mutations do not need to be coded explicitly. For example, if two mutations occur in one isoform, there is no need to code the mutant isoform that contains both mutations. It is sufficient to code one mutant isoform for each mutation. The reason is that mismatches in mapping are allowed, where if the singly-mutated versions exist in the reference then the jointly-mutated version will be differentiated from the WT version, and will instead be mapped to the closest one of the singly-mutated versions.

So we first generate one mutant isoform for each mutation in the isoform. This process produces up to $M \times N$ mutant isoform sequences, which are substantially less than the $M \times 2^N$ potential mutant isoforms mentioned above. The sequences are then added into the WT reference and used as the reference for RNA-seq data mapping. In the processing of the mapping output, we then combine the mutant isoforms associated with one WT isoform into one mutant isoform. We illustrate the case where there are only two mutant versions—isoform_Mut1 and isoform_Mut2—for a WT version isoform_WT. In Step 1, for every eqClass that contains isoform_Mut1 and isoform_Mut2, we relabel isoform_Mut1 and isoform_Mut2 as isoform_Mut. In Step 2, we compare the isoform patterns from all eqClasses, and merge eqClasses with the same binary pattern by adding the corresponding read counts. The new eqClass table is then used to construct the mutant X matrix. MAX also inherits the procedure to merge paralogs from XAEM (7) method. However, to avoid merging wild-type isoforms with mutant isoforms, it applies the procedure only for the wild-type isoforms, then the mutant version follows the paralog structure of the wild-type.

In the NAA20 gene example, there are three WT isoforms: NM_016100, NM_181527 and NM_181528. We simulate a G-A point mutation at the position of Chr20:20026790. The mutation is carried by all three isoforms, which means that the mutant NAA20 will have six isoforms in total: the original three WT isoforms and the mutant isoforms NM_016100_Mut, NM_181527_Mut and NM_181528_Mut. For comparison, the wild-type and mutant X matrices are given in Tables 1 and 2. The number of eqClasses in the mutant X increases from 6 to 16, a large increase in complexity due to just a single mutation. The X matrix here is well conditioned, so it is possible to estimate the expression of all isoforms and their mutant versions. Note that there are no eqClasses that correspond to the mutant NM_016100 uniquely. This means it is not necessary to have reads that cover both the mutation and the isoform-defining part uniquely.

MAX2: extension of MAX for heterogeneous samples

Empirically, a mutated gene has one or a few variants in a given sample, and they may vary from sample to sample, potentially creating heterogeneity between samples. Sample heterogeneity violates our model, where the same X matrix is assumed across the samples. The single X matrix estimated from the pool of all samples could differ from the appropriate X for an individual sample. Statistically, this induces bias in the estimation, so sample heterogeneity will reduce the accuracy of a quantification method. To solve this

Table 2. The initial design matrix X for the mutant NAA20 gene, which contains six isoforms and 16 eqClasses

	NM_016100		NM_181527		NM_181528	
EqClass	WT	Mut	WT	Mut	WT	Mut
000001	0	0	0	0	0	0.16
000010	0	0	0	0	0.16	0
000011	0	0	0	0	0.23	0.22
000100	0	0	0	0.11	0	0
001000	0	0	0.11	0	0	0
001100	0	0	0.16	0.16	0	0
010001	0	0.07	0	0	0	0.08
010100	0	0.12	0	0.11	0	0
010101	0	0.01	0	0.01	0	0.04
100010	0.07	0	0	0	0.08	0
101000	0.12	0	0.11	0	0	0
101010	0.01	0	0	0	0.04	0
110000	0.05	0.04	0	0	0	0
110011	0.03	0.03	0	0	0.03	0.03
111100	0.36	0.36	0.31	0.32	0	0
111111	0.36	0.36	0.31	0.31	0.46	0.46

issue, we develop MAX2, an extension of MAX for heterogeneous samples.

We assume that the mutation data are available, for instance, in a separate exome sequencing. Since each DNA mutation corresponds to its own list of mutant isoforms, we use the observed DNA mutations as input in MAX2. Briefly, MAX2 starts by clustering the samples based on their common mutation(s). Then for each cluster we derive an initial mutant X that correspond to the relevant mutation(s) that identify the group. Finally, the AEM algorithm is applied to each cluster to obtain the isoform abundance. If the number of individuals in a cluster is deemed too few, such as less than five, we do not update the X matrix, so the AEM algorithm reduces to the standard EM.

Real RNA-seq data

To investigate the mutation-specific expression, we study a cohort of AML patients from the BeatAML project, which provides a comprehensive resource including omics data, clinical records and drug response data (19). A total of 461 RNA-seq samples are included for isoform quantification. In the original study of BeatAML, the whole-exome sequencing (WES) data were produced as input of Mutect (3) and Varscan2 (23) to detect the single-nucleotide variations and indels. We collect mutations that are predicted to have functional impact for further analysis.

Previous studies have proposed a two-hit model for the tumorigenesis of AML, involving two classes of mutations (24). According to the model, class-I mutations provide proliferative and cancer-cell survival advantages, while class-II mutations impair the processes of cell differentiation and apoptosis. The most common class-I mutated genes include FLT3 (FMS-like tyrosine kinase 3) and TP53, while NPM1 is the most common in class-II (25). The FLT3 gene is a crucial component which involves haematopoiesis such as proliferation and differentiation. Mutations in FLT3 have been strongly associated with high blast counts, increased risk of relapse and unfavorable prognosis (26). The most common type of FLT3 mutation is an internal tandem duplication (ITD), where the length of duplication ranges from 40 to 400 base pairs (bp). The FLT3-ITD occurs in about 27% of AML patients, which is significantly related to the poor overall survival (27).

TP53 is the key player in the apoptosis pathway and is one of the most commonly mutated genes in cancer (28). TP53 mutation in AML patients is associated with poor survival (29). NPM1 mutations represent another common group of genetic abnormalities in AML patients. The mutations usually involve exon 12 in NPM1, which occur in 8% of pediatric AML and 30% of adult AML cases (30). The most frequent mutation in NPM1 is a 4 bp insertion at chr5:171410539. In this study, we focus on analysing mutant–allele expression of the FLT3, NPM1 and TP53 genes. Further information about the mutations in these genes from the BeatAML data is given in Supplementary File II.

Simulated RNA-seq data

The simulation of RNA-seq data has been commonly used to evaluate the performance of quantification methods (5,6). To mimic the real RNA-seq dataset, we derive the expression values from a human cancer cell-line HCT116 (31), which is processed using the Sailfish method (4). The average library size per sample is nine million, or ∼420 read-pair counts per isoform. We simulate both non-mutated and mutated RNA-seq samples for a comprehensive assessment. In the non-mutated samples, we only assign read counts to the wild-type isoforms, which means that the mutant isoforms are not expressed. This is aiming to evaluate the false positive rate in each method. As described above, in this simulation, we focus on FLT3, NPM1 and TP53. According to the BeatAML data, the FLT3 gene contains 120 distinct mutations, the NPM1 14 mutations and the TP53 49 mutations.

To simulate realistic mutated samples, we use the observed DNA-mutations in BeatAML patients, so each sample has one or a few mutations in the gene of interest, and the mutation profile varies across samples. For FLT3, there are 179 patients carrying 120 unique mutations, which result in 202 mutation events. Thus, each patient has an average of 1.13 mutations. For NPM1, a total of 135 mutation events occur in 135 patients, which indicate that each patient carries only one mutation on average. For TP53, there are 67 mutation events in 54 patients, so that each patient has an average of 1.24 mutation events. For each sample, we generate only mutant isoforms relevant to the mutation(s) of that sample. We then assign equal read counts to a wild-type isoform and its mutant isoform. If there are multiple mutant forms of the wild-type isoform (due to multiple mutations), each mutant isoform is expressed equally to its wild-type expression. The software Polyester is used to simulate paired-end sequencing reads (20). We set the read length at 100 bp, the average fragment length at 250 bp and the standard deviation at 25.

Method comparisons

The alignment-free methods, such as Sailfish (4), Salmon (5) and Kallisto (6), are widely used to quantify the isoform expression. These methods have highly similar performance (7,32), so we implement Salmon as the main com-

Table 3. Comparison of median APEs in 100 non-mutated and 100 mutated samples for FLT3, NPM1 and TP53

	FLT3		NPM1		TP53	
	WT	Mut	WT	Mut	WT	Mut
Non-mutated samples						
MAX	0.05	0	0.06	0	0.20	0
Salmon	0.06	0.04	0.07	0	0.39	0
RSEM	0.06	0	0.08	0	0.57	0
Mutated samples						
MAX	0.16	0.14	0.06	0.07	0.40	0.37
MAX2	0.13	0.10	0.07	0.07	0.13	0.12
Salmon	0.11	0.10	0.08	0.07	0.38	0.38
Salmon_PTR	0.10	0.09	0.07	0.07	0.44	0.40
RSEM	0.13	0.11	0.08	0.10	0.61	0.58
RSEM_PTR	0.09	0.10	0.08	0.10	0.56	0.54

parison to MAX. In principle, once a transcriptome reference annotation is established, one could use Salmon to estimate isoform-level expression. Our comparisons are fair as far as they are based on exactly the same WT and WT + mutant transcriptome references that we use for MAX. We also implement RSEM in combination with Bowtie2 as an alignment-based method for comparison (8). The gene model hg38 is used as the genome and transcriptome reference. The scripts and commands to run each method are uploaded in MAX Github repository, the link is https://github.com/WenjiangDeng/MAX.

Given that the mutation profiles of simulated samples are available, as a fair comparison to MAX2, we have constructed the personalized transcriptome reference (PTR) and re-run Salmon and RSEM for isoform quantification. The methods are called Salmon_PTR and RSEM_PTR in Table 3.

To measure the accuracy of each method, we calculate the absolute proportion error (APE) defined by

$$\text{APE} = |E - T|/(T + 1). \quad (2)$$

where E is the estimated read count and T is the true read count (with 1 added to avoid division by zero). We summarize the median APE from each isoform across all simulated samples as the final metric. In computing the APE for MAX/MAX2, paralogs are treated like other isoforms. Paralogs are a unique feature of MAX/MAX2, but neither Salmon nor RSEM are aware of or warn about the paralog problem. So, while they are not 'apples to apples', the APE comparisons of the default output from each method indicate the real performance of the existing methods when paralogs are necessary, as in the case of TP53 gene. However, we subsequently also compare the results when we include the paralog information from MAX into the output of Salmon and RSEM.

When using the WT + mutant reference, both the wild-type and mutant isoforms are estimated. This means that we can use the known non-mutated samples, which should produce no mutant–allele expression, to assess the false positive rate of a method. In this case, we use the wild-type expression T to assess the mutant–allele expression estimate E.

RESULTS

Mutant expression of the FLT3 gene in BeatAML patients

We start by illustrating the quantification of the mutant expression of FLT3. The wild-type FLT3 contains 24 exons, organized into two isoforms: NM_004119 and NR_130706. In the BeatAML data, we observed 120 unique mutations in the exonic regions. In standard quantification, the mutation status is ignored and the RNA-seq reads are mapped to the WT transcriptome reference. In MAX we construct a WT+mutant reference, by incorporating 240 possible mutant isoforms. Therefore, in addition to the expression of WT isoforms, we also have the expression of mutant isoforms. As described above, the expression values of the mutant isoforms are combined according to their WT versions.

Figure 2 compares the isoform abundance using MAX and the standard WT reference versus the WT + mutant reference for 122 FLT3-mutated samples. Each isoform for each sample will have three expression values: (i) the standard expression estimate from the WT reference, (ii) the WT expression estimates using WT+mutant reference and (iii) the mutant expression using WT + mutant reference. Note that when using the WT reference, we can only get the standard WT expressions, while the mutant–allele expression is not available. In contrast, using the WT+mutant reference, we can quantify both WT and mutant allele expressions. As expected, the WT-allele expression in the WT + mutant reference (grey circles) is smaller than the standard expression estimated using the WT reference. However, the sum of the WT and mutant–allele expression of each isoform, estimated based on mapping to the WT + mutant reference (black circles), match the expression estimated from the WT reference. The results indicate that, for mutated genes, the standard quantification using the WT reference indeed includes the mutant–allele expression.

Simulation results

To investigate the performance of MAX in quantifying the mutation-specific expression, we simulate the RNA-seq data of 100 non-mutated samples and another 100 mutated samples. From the BeatAML data the FLT3 gene has 120 unique mutations, NPM1 has 14 mutations, and TP53 has 49 distinct mutations. The numbers of WT isoforms for FLT3, NPM1 and TP53 are 2, 8 and 15, respectively; and a total of 240, 84 and 522 mutant isoforms are incorporated in the WT+mutant reference for these genes. Because of substantial sequence similarities, MAX merges the 15 WT isoforms of the TP53 gene into four paralogs. The mutant versions are merged in the same way. The procedure to merge paralogs is inherited from XAEM (7) method. In the 100 non-mutated samples, we only assign read-counts to the wild-type isoforms, while in the mutated samples, an equal number of read counts are assigned to both the wild-type and mutant isoforms. We then implement MAX, RSEM and Salmon to quantify the isoform expression. The median APE for each isoform is calculated to evaluate the accuracy. Since the number of isoforms is large, we summarize the median APE across all isoforms for each gene as the final APE in Table 3.

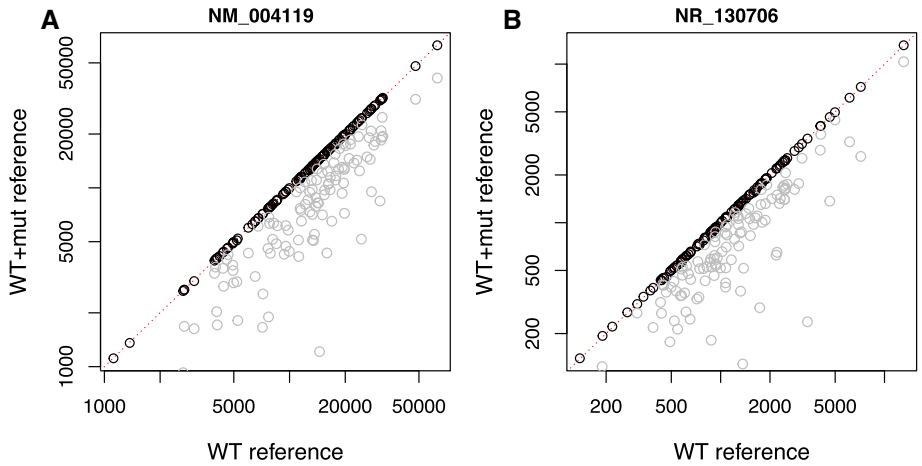

Figure 2. Comparison of the isoform quantification of the FLT3 gene in 122 mutated samples in the BeatAML real dataset, using the WT reference vs WT + mutant reference. The grey circles represent the expression of the WT isoform based on the WT (x-axis) versus the expression of WT isoform based on the WT + mutant references (y-axis). The black circles are the expression of the WT isoform based on the WT reference (x-axis) versus the sum of WT and mutant–allele expressions based on the WT + mutant reference (y-axis).

Figure 3 shows the estimated and true expression of the mutant isoforms of FLT3, the circles represent the 100 simulated samples. It can be seen that the circles are distributed near the diagonal line, which indicates that MAX, RSEM and Salmon have accurate and similar performances. The exact value of accuracy is summarized in Table 3. The corresponding plots for NPM1 and TP53 are presented in Supplementary Figures S1 and S2, which demonstrate similar results.

Table 3 shows the comparison of APEs between MAX, Salmon and RSEM, combining the isoforms of the three genes. Supplementary Tables S2– S5 show the results for each constituent isoform. For the WT isoforms in non-mutated samples, MAX and Salmon generally perform a little bit better than RSEM. For the FLT3 gene, the median APEs of MAX for NM_004119 and NR_130706 are 0.04 and 0.06, respectively, compared with 0.05 and 0.06 for Salmon, and 0.05 and 0.08 for RSEM (Supplementary Table S2). When summarized across isoforms in Table 3, the corresponding median APEs are 0.05, 0.06 and 0.06. However, Salmon has some false positives for mutant FLT3 isoforms, giving a median APE 0.04. For NPM1, the median APEs of MAX, Salmon and RSEM are 0.06, 0.07 and 0.08, respectively. For the TP53 gene, the quantification results are generally worse than for FLT3 and NPM1; the APEs for MAX, Salmon and RSEM increase to 0.20, 0.39 and 0.57, respectively. It seems reasonable however that as the number of constituent isoforms increases, the estimation problem becomes harder.

Isoform quantification in mutated samples is potentially noisier, because the mutant isoforms are set to be expressed so there is more chance for them to influence the overall performance. Indeed, as seen in Table 3, RSEM's performance becomes worse for both WT and mutant isoforms, with median APEs larger than MAX and Salmon. In contrast, MAX and Salmon have a stable performance in mutated samples. For example, the median APEs of MAX2 in WT and mutant FLT3 are 0.13 and 0.1, respectively; for Salmon the values are 0.11 and 0.1, respectively. As expected, MAX2 improves on MAX for the FLT3 and TP53 genes, which have a rich mutation landscapes across samples. In contrast, there is no significant improvement for the NPM1 gene, as it is dominated by a single mutation: 4-bp TCTG insertion. Overall, the estimates from alignment-free methods, i.e. MAX and Salmon, are highly accurate for both WT and mutant isoforms in non-mutated and mutated samples. RSEM performs somewhat worse than the other two methods. Salmon is better than MAX for FLT3 in the mutated samples, and comparable to MAX2 for NPM1; otherwise, MAX2 is more accurate than Salmon.

We now turn to the results for RSEM and Salmon based on the personalized transcriptome reference (RSEM_PTR and Salmon_PTR). For the FLT3 gene, which has heterogeneous mutations across individuals, the median APEs of RSEM decrease from 0.13 to 0.09 for the WT alleles, and from 0.11 to 0.10 for the mutated alleles. The performance of Salmon improves slightly when using the personalized reference. However, for the NPM1 gene, the median APEs between the two runs were practically the same. Intriguingly, for the TP53 gene, RSEM's accuracy improves from 0.61 to 0.56 and from 0.58 to 0.54 for WT and Mut alleles, respectively; but the accuracy of Salmon becomes a little bit worse, where the APEs increase from 0.38 to 0.44 and from 0.38 to 0.40 for WT and Mut alleles, respectively. These findings suggest that there can be some improvements in APEs but it is not guaranteed. The improvement depends on the gene and the heterogeneity of the mutation profiles.

We have also checked the scenario where 100% read counts are assigned to the mutant isoforms. The simulation setting is the same as for Table 3 (100 mutated samples with mutation profiles taken from the BeatAML data). The results in the Supplementary Table S6 show that MAX is slightly better than Salmon, and these two methods are better than RSEM. We further investigate whether the paralog merging leads to more accurate estimates than the level of individual transcripts of other methods. We apply the same paralog merging of the TP53 isoforms for both RSEM and Salmon, then calculate the new median APEs for the mu-

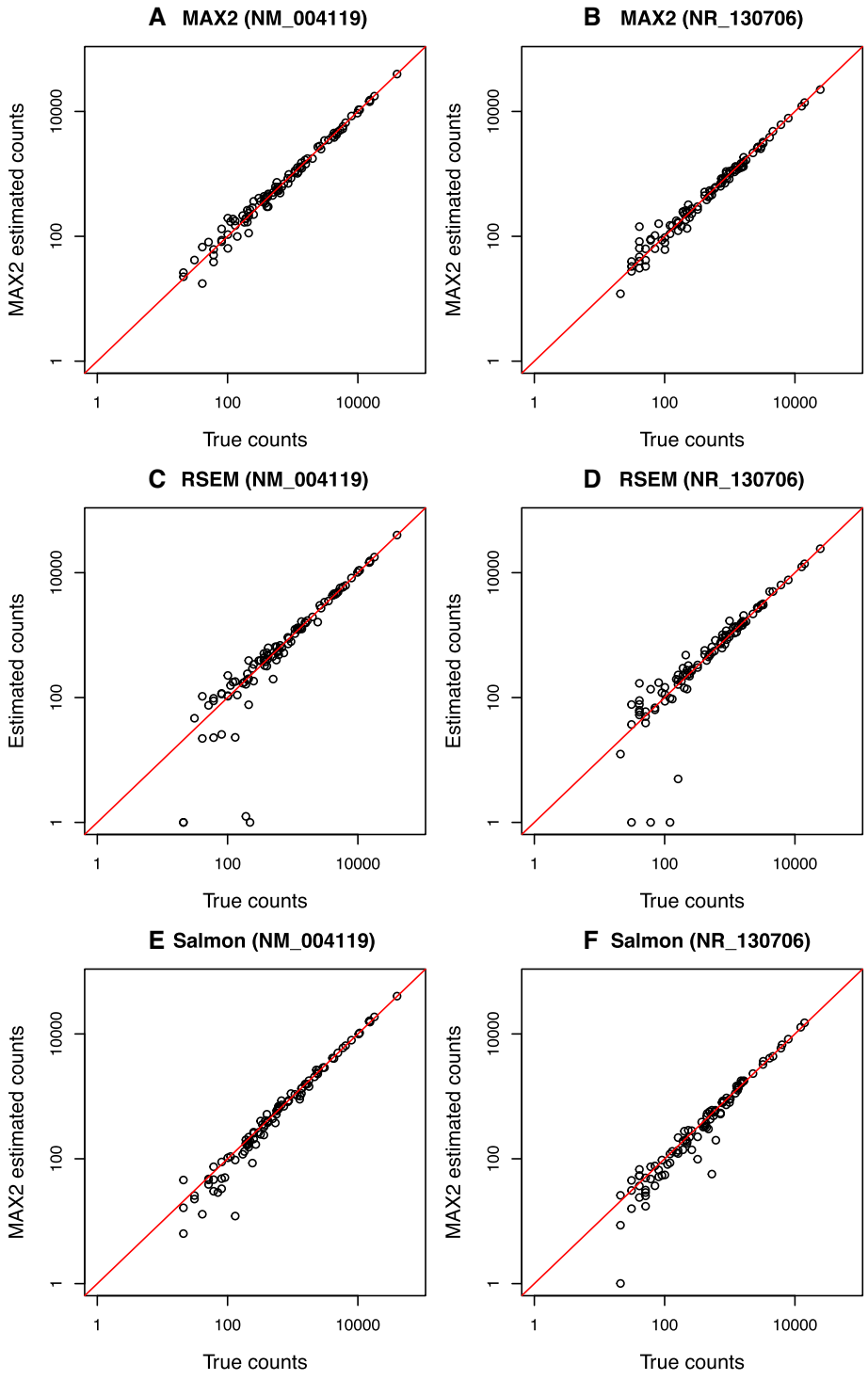

Figure 3. Comparison between true and estimated values for the mutant isoforms in FLT3 gene. A total of 100 simulated FLT3-mutated samples are included in this comparison. Each circle represents one sample. (**A**) The estimated values of NM_004119 from MAX2 and (**B**) estimations of NR_130706 from MAX2. (**C**) The estimated values of NM_004119 and (**D**) estimations of NR_130706 from RSEM. (**E**) and (**F**) are estimates from Salmon. The x-axis and y-axis are in log scale.

Table 4. The median APE of ASEP method of FLT3 from the 100 simulated RNA-seq samples

		NM_004119		NR_130706	
Sample group	Num. of samples	WT	Mut	WT	Mut
one SNV	13	0.08	0.12	0.07	0.11
>one mutation	29	1.37	0.66	1.25	0.68
one ITD mutation	58	0.85	0.86	0.89	0.85
Total	100	0.89	0.89	0.80	0.79

Table 5. The comparison of median APEs of ASEP, MAX and MAX2 using the 100 simulated data carrying only one SNV mutation

	NM_004119		NR_130706	
	WT	Mut	WT	Mut
ASEP	0.29	0.27	0.25	102
MAX	0.24	0.25	0.22	39
MAX2	0.12	0.08	0.13	0.05

tated samples. The Supplementary Table S7 shows that using the paralog merging both RSEM and Salmon achieve lower median errors. Since TP53 has a complex mutation profile, in this case MAX2 is more suitable than MAX and it performs better than the other methods.

Comparison with a conventional ASE method ASEP

We apply a conventional gene-based ASE tool named ASEP (11), which estimates allele-specific expression across a population of samples, to analyze the 100 mutated samples. To get isoform-level estimates, it is assumed that the mutant:WT ratio estimated at gene level applies to the isoform level. The assumption is satisfied in the simulation setting above. We first implement STAR (33), Varscan2 (23) and Pindel (34) to produce the input files for ASEP, which are read counts supporting the REF and ALT allele. We note that Pindel was designed for DNA-seq rather than RNA-seq data, so its use here might not be as intended by its authors. We then use ASEP to estimate the ASE at gene level, hence get the mutant:WT allelic ratio. The XAEM method is applied to quantify isoform-level expression using the wild-type transcript reference (7). We then apply the mutant:WT ratio from ASEP to XAEM-based estimates to get the allele-specific expression at the isoform level.

Here, we take FLT3 for illustration, since it only has two isoforms, so it is straightforward to compute and compare the mutant and WT isoforms. We use the same 100 mutated samples in Table 3 and the APE is used as the performance metric. The results are summarized in Table 4. The overall performance of ASEP for the total 100 samples is very poor, with APEs around 0.8–0.9, compared to APEs around 0.10–0.13 for MAX2, Salmon or RSEM.

To understand the details, we divide the 100 samples into three groups according to their mutation profiles. There are 13 samples with only one SNV. This group is the most natural case for ASEP. The median APEs for WT and mutant NM_004119 are 0.08 and 0.12, respectively; for NR_130706 the APEs are 0.07 and 0.11. For the WT isoforms these are even smaller than the overall APEs of the other methods. This indicates that gene-based methods such as ASEP can indeed work well in this highly specific scenario (one SNV and equal mutant:WT ratio across isoforms).

However, when it comes to the 29 samples with more than one mutation (e.g. two SNVs, or one SNV and one ITD mutation), the APEs increase greatly to ∼1.3 and ∼0.67 for the WT and mutant isoforms. Thus substantially worse than the other methods. In fact, the predominant genomic alteration of FLT3 in AML patients is the ITD mutation, which involves tandem duplications of varying length across individuals. As shown in Table 4, the median APEs are ∼0.85 in the ITD group (n=58), which indicates that ASEP has a really poor performance for this mutation. Overall, these results demonstrate that the conventional ASE methods are not applicable as a general method for quantifying mutation-specific expression at isoform level, even when the mutant:WT ratio is equal across isoforms.

As ASEP works well only in the case of one SNV mutation, we shall continue under this setting. There are 17 unique SNVs among the real BeatAML patients. We then sample the 17 SNVs with replacement to generate 100 SNVs and simulate the corresponding 100 RNA-seq samples, where each sample carries only one SNV mutation. To produce a challenging dataset, in each sample, we only assign read counts to mutant NM_004119, but mutant NR_130706 is not expressed, so that the mutant:WT ratios of NM_004119 and NR_130706 are 1:1 and 0:1, respectively. We then run ASEP, MAX and MAX2 to quantify the expression. MAX2 is included because here the samples have heterogeneous mutation profiles. ASEP-based estimates are computed as described above. The APEs are calculated and summarized in Table 5.

The median APEs of ASEP for WT and mutant NM_004119 are 0.29 and 0.27, respectively. MAX has slightly lower APEs, which are about 0.25. MAX2 achieves a substantially higher accuracy for WT and mutant NM_004119, the APEs are 0.12 and 0.08, respectively. For WT NR_130706, the APEs across three methods are similar with those of NM_004119, where MAX2 has the smallest APE. The greatest challenge is for the mutant NM_004119, which is not expressed. ASEP has a high APE at 102, which indicates substantial false positive rate, which is expected since it presumes equal mutant:WT ratios across isoforms. The APE of MAX is 39, which is smaller than ASEP but still shows many false positives. In contrast, MAX2 is highly accurate and achieves a small APE at 0.05. This high accuracy shows the ability and advantage of MAX2 in dealing with heterogeneous samples. Taken together, this analysis demonstrates that an isoform-based method, under a general scenario of unequal mutant:WT expression ratio across isoforms, can have substantially better performance than conventional gene-based ASE tools.

Real data analysis

We have estimated the WT and mutant isoform expression of FLT3, NPM1 and TP53 in the BeatAML data. The number of mutated samples are 122, 82 and 36, respectively, and all the non-mutated samples are included in the quantification for validation purposes. The expression profile of the isoforms of NPM1 is given in Figure 4A. This gene has eight WT and six mutant isoforms; the mutants are put next

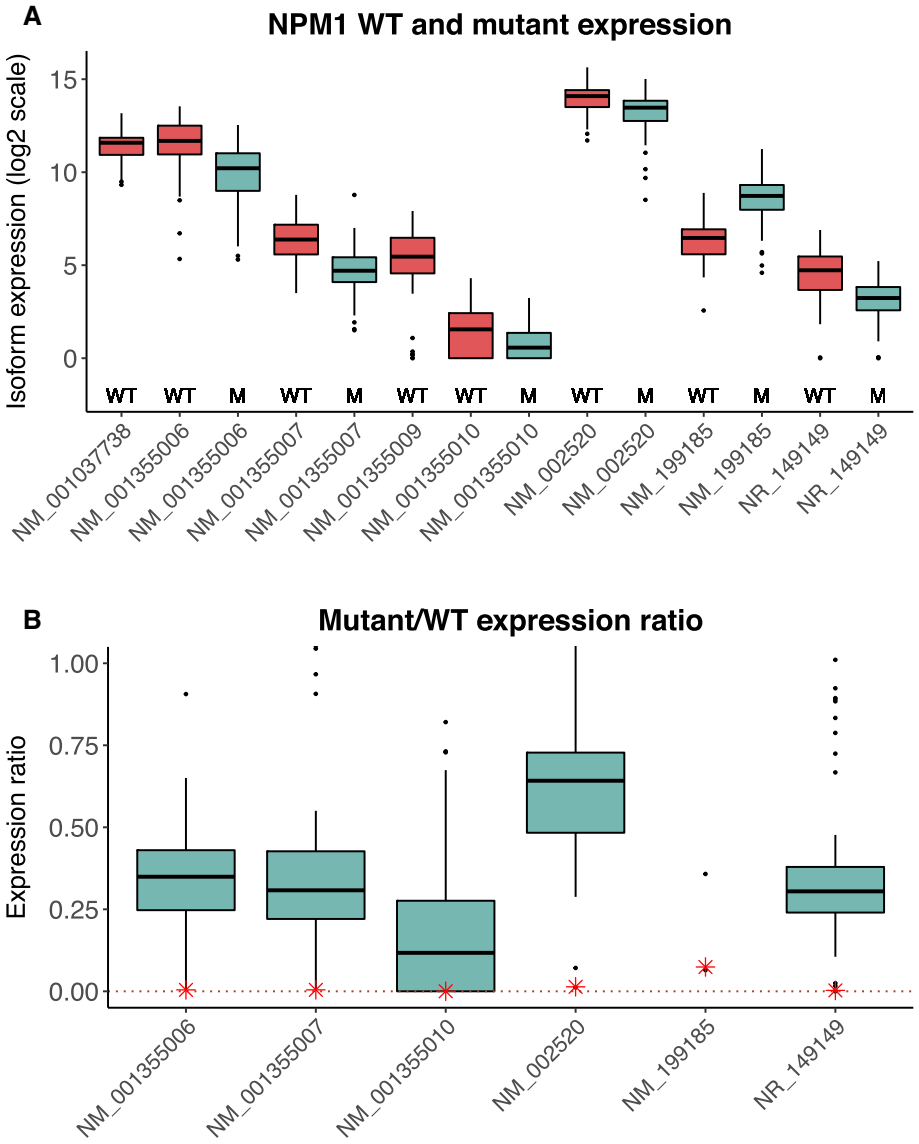

Figure 4. (**A**) Boxplots of WT and mutant expression of the NPM1 gene in 82 NPM1-mutated samples from the BeatAML data. The WT and mutant status are given at the bottom of the panel. (**B**) Boxplots of the mutant/WT expression ratio. The isoforms are in the same order as in (A). The red stars at the bottom of the panel are the corresponding median ratios from the 379 non-mutated samples. (Note: the median ratio for NM_199185 is 5.21, hence the boxplot for this isoform falls above the plotting region.)

to the WT version. Isoforms number 1 (NM_001037738) and number 6 (NM_001355009) do not have mutant versions.

Figure 4B shows the mutant/WT expression ratios for the six mutant isoforms. All mutant isoforms, except for NM_199185, do not reach the same expression level as their WT versions. This could be partly due to the clonality, where only some fraction of the cells carry the mutations. The red stars at the bottom of panel (B) are the corresponding median ratios from the 379 non-mutated samples. This shows that there is little false positive mutant-expression estimated in the non-mutated samples. Similar figures for the FLT3 and TP53 are shown in Supplementary Figures S3 and S4. We also observe little false-positive mutant expression in the non-mutated samples of these genes.

We next assess the biological significance of the mutant–allele expression. Even though mutations in the NPM1 and FLT3 genes are well-known driver mutations in AML, as shown in the survival curves of Supplementary Figure S5, the NPM1 and FLT3 mutation status alone are not associated with survival (logrank P-value = 0.83 and 0.10, respectively). We note that these patients were normally treated, so we are not looking at their natural history. It is possible that differential treatment effects in the mutated and non-mutated groups attenuated any underlying/treatment-naive differences. The clinical significance of these mutations is, however, visible in their interaction: the group with the best survival is found among the ITD-negative NPM1-positive (logrank P-value= 0.0009 versus ITD-positive NPM1-positive). One possible explanation is that the ITD-negative NPM1-positive group respond well to their treatment. Or,

alternatively, the ITD-positive NPM1-positive patients are poor responders. So, we ask the question whether there is a subgroup of the ITD-positive NPM1-positive patients that are good responders to some experimental treatments.

We first define two subgroups within the NPM1-positive patients according to their level of mutant–allele expression. Since the mutant isoforms vary substantially both in their absolute expression level and the relative expression to the WT version, we choose only the mutant isoform of the dominant isoform NM_002520_Mut for further analysis. High mutant expression is defined as the ratio of mutant/WT expression greater than its median value.

We then compare the high versus low mutant-expression subgroups in terms of their drug response; statistical comparisons are performed using the standard *t*-test. We collect a total of 32 263 experiments of these AML patients from the *ex-vivo* drug screening of 122 drugs in the BeatAML project. For each experiment, the drug sensitivity is measured in terms of its IC50 and AUC (area under the curve) metric. A small value of either metric indicates an effective drug for that specific tumor, or similarly the tumor is responsive to the drug. Vice versa, a high value indicates an ineffective drug for the tumor, or a drug-resistant tumor.

Figure 5A and B shows the volcano plots of the 122 drugs. We now consider only drugs that are significant in both the IC50 and AUC values. This analysis identifies VX-745 and panobinostat as potentially effective drugs in the high mutant-expression subgroup. VX-745 is a p38α mitogen-activated protein (MAP) kinase inhibitor. Panobinostat is an inhibitor of histone deacetylase, which regulates gene transcription, cell-cycle progression, and apoptosis.

Figure 5C and D shows the boxplots of the IC50 and AUC as response to VX-745 in the six subgroups defined by ITD status and NPM1 mutant expression level. The latter is only defined within the NPM1-positive patients. Among the ITD-negative patients (three left-most boxplots), both mutant subgroups have better drug response than the NPM1-negative patients (P-value= 0.04 for low mutant-expression versus NPM1-negative group for the IC50 metric and 0.0004 for the AUC; there was no significant difference between the low- versus high-mutant groups for both metrics).

However, within the ITD-positive group (three right-most boxplots), a good drug response is achieved only by the low mutant-expression subgroup (P-value = 0.001 versus the high mutant-expression subgroup for the IC50, and P-value = 0.03 for the AUC). Supplementary Figure S6 shows that the total NPM1 gene expression does not carry the same information (P-value = 0.39 for the low- versus high-mutant subgroups.) In summary, based on the mutant–allele expression information and in conjunction with ITD-status, we have identified a subgroup of NPM1-positive with a potentially good response to VX-745.

As shown in Supplementary Figure S7, panobinostat seems to kill most cells at low dose, so there is not much scope for individualized therapy. However, there is also some evidence that high mutant-expresssion of NPM1 in the ITD+ group is associated with higher drug resistance (P-value=0.05 versus the low mutant-expression group).

DISCUSSION

One reason why MAX performs well against other methods is likely due to the handling of sequence similarities. Mutant alleles have strong sequence similarities with their wild-type versions as well as with each other. Statistically, the model is of the form $\mu = X\beta$, where the design matrix X captures the exon sharing between isoforms. The sequence similarities will result in an X with a poor condition number or nearly singular. This will generally produce large variability in the estimates. MAX can identify and avoid/reduce this problem, since X is available explicitly. This facilitates analysis of the problem and suggests solutions by constructing paralogs or merging of mutant isoforms. In contrast, standard methods such as Salmon, which attempt to jointly estimate the expression of all isoforms in the whole transcriptome and do not have any explicitly defined X matrix, lack a natural way to perform anything similar.

To some extent, mutant–allele expression is similar to allelic-specific expression (ASE), but as far as we know currently used methods to estimate ASE are gene-based. We have previously described that isoform level is a more meaningful context to assess the impact of mutations. Moreover, ASE is typically assessed in germ-line tissues, where allelic imbalance is often the main interest as it might lead to phenotypic variation. However, mutant–allele expression in cancer may have a functional impact that has nothing to do with allelic imbalance. Interestingly, mutant allele expression from RNA-seq is currently not commonly measured nor reported.

The analyses of a large collection of non-mutated samples in the BeatAML data indicate very little false positive mutant expression. We have used this fact as a validation of MAX in real data. However, it also means that in principle we can use MAX for gene-centric mutation detection from RNA-seq data, based on an external list of mutations detected by exome sequencing on independent samples or from some public databases of known mutations. However, this approach will presume that the mutated gene(s) are expressed, as obviously we cannot detect mutant expression if it is unexpressed.

Our work was partly motivated by the challenge of discovering an effective therapy in AML, one of the most common hematological malignancies, accounting for approximately 80% in acute leukemia patients (35). The pathogenesis of AML is associated with the abnormally proliferated and differentiated myeloid stem cells, which are theoretically driven by somatic mutations. We have used a large cohort of AML patients from the BeatAML project to illustrate our method and have revealed an association between mutant–allele expression and drug response in a subgroup of NPM1-mutated patients. The current advances in treatment have improved the outcome of young patients significantly. However, the prognosis of the elderly patients, which account for the majority of new cases, remains poor and challenging. The information on the expression pattern of mutation alleles using MAX could provide novel insights for individualized treatment of AML patients.

Our study has some strengths and weaknesses. The main strength is that MAX is based on a flexible tool that was

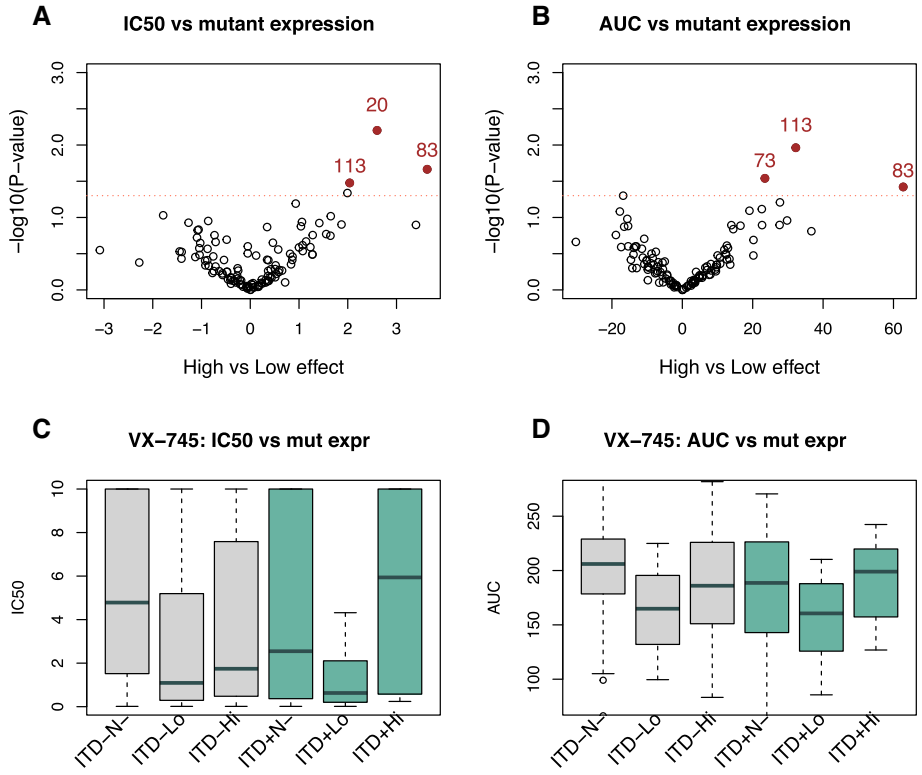

Figure 5. (A and B) Volcano plots of the association between drug response and mutant expression in the BeatAML data. The drug response is measured by IC50 (**A**) and AUC (**B**); 122 drugs are included in the analysis. The x-axis is the difference in drug response between the high vs low mutant subgroups, and the y-axis is the corresponding $-\log_{10}P$-value. The concordant drugs 83 and 113 are panobinostat and VX-745, respectively. (**C**) Boxplots of IC50s of the six groups defined by ITD status and NPM1 mutant expression. The sample sizes for the six groups are 196, 12, 20, 47, 15, 10, respectively. (**D**) Same as (C) using the AUC values for drug response.

previously shown to perform well against other isoform-quantification methods. We have assessed the performance of MAX in a realistic simulation study and applied it to a large cohort of 461 AML samples. In the real data analysis, it shows promising added value in identifying a subgroup of patients that respond well to specific therapy. But our study also has a number of shortcomings. The discovery of the interesting subgroup is based on small number of patients ($n = 32$ patients with ITD-positive and NPM1-positive), so more work is needed to validate the result. The simulation study focuses on single genes such as TP53 and NPM1; this may violate some distributional assumptions of isoform expression. It may also reduce the spurious mapping of reads from elsewhere in the transcriptome. We have however tried to make the simulation setting as realistic as possible. Firstly, note that MAX, RSEM and Salmon are all accurate for non-mutated samples, so it could be more practical and reasonable to just focus on certain mutated genes. All mutation patterns and wild-type expression levels are as observed in the real BeatAML data. In our previous method (specifically during the construction of transcription clusters in XAEM (7)), we can check for each gene whether there are reads that spuriously map from elsewhere. For the three genes we study here there is no issue of spuriously mapped reads (if there were, the X matrices for these genes will include other genes).

One of MAX's current shortcomings is that the mutations must be in the exonic regions. If a driver mutation is in the intronic or in the intergenic regions, then it is not clear how to define mutant–allele expression in a meaningful way. The concept of expression quantitative-locus (eQTL) could be extended to these mutations. MAX assumes that the samples are relatively homogeneous so that a single design matrix X is appropriate. This is violated, for example, in the analysis of TP53 gene, which contains many mutations, generating potentially heterogeneous samples. To deal with this problem, we have developed an extension called MAX2, which uses only the simplest clustering based on mutation profile. A more sophisticated clustering might improve the method further. These issues are worthy of future investigation.

CONCLUSION

We have developed and implemented a method named MAX for isoform-level quantification of mutant–allele expression. MAX provides biologically more meaningful information than the standard quantification of mutated genes, which is included in the wild-type allele expression. We have shown in the simulation study that MAX performs well against an implementation based on a standard isoform-quantification method. In the analysis of real dataset of AML patients, we reveal a subgroup of NPM1-mutated patients with a good drug response. Overall, MAX represents a promising and informative tool for RNA-seq data analysis.

DATA AVAILABILITY

The MAX pipeline and source codes can be found at https://github.com/wenjiangdeng/MAX. We also upload the original dataset and scripts for RNA-seq data simulation and analysis in github repository, the folder is Scripts_and_Files. The BeatAML dataset used in this study is available from the BeatAML project (19).

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

We gratefully acknowledge the computational resources provided by the Swedish National Infrastructure for Computing (SNIC) in Uppsala, which is partially funded by the Swedish Research Council through grant agreement no. 2018-05973. We are also grateful to the investigators and the patients who had contributed to the BeatAML study.
Author contributions: W.J.D., Y.P. and T.N.V. contributed the conception, method development, software implementation, data analysis and report writing. T.M. contributed to the data analysis, method development and report writing. All authors have read and approved the manuscript, method tutorial and supplementary information

FUNDING

Swedish Cancer Fonden, the Swedish Research Council (to V.R. in part); Swedish Foundation for Strategic Research (to S.S.F.); China Scholarship Council [201600160085].
Conflict of interest statement. The authors declare that they have no competing interests.

REFERENCES

1. Devarakonda,S., Rotolo,F., Tsao,M.-S., Lanc,I., Brambilla,E., Masood,A., Olaussen,K.A., Fulton,R., Sakashita,S., McLeer-Florin,A. *et al.* (2018) Tumor mutation burden as a biomarker in resected non–small-cell lung cancer. *J Clin. Oncol.*, **36**, 2995.
2. Suo,C., Deng,W., Vu,T.N., Li,M., Shi,L. and Pawitan,Y. (2018) Accumulation of potential driver genes with genomic alterations predicts survival of high-risk neuroblastoma patients. *Biol. Direct*, **13**, 14.
3. Cibulskis,K., Lawrence,M.S., Carter,S.L., Sivachenko,A., Jaffe,D., Sougnez,C., Gabriel,S., Meyerson,M., Lander,E.S. and Getz,G. (2013) Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.*, **31**, 213–219.
4. Patro,R., Mount,S.M. and Kingsford,C. (2014) Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat. Biotechnol.*, **32**, 462–464.
5. Patro,R., Duggal,G., Love,M.I., Irizarry,R.A. and Kingsford,C. (2017) Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods*, **14**, 417–419.
6. Bray,N.L., Pimentel,H., Melsted,P. and Pachter,L. (2016) Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, **34**, 525–527.
7. Deng,W., Mou,T., Kalari,K.R., Niu,N., Wang,L., Pawitan,Y. and Vu,T.N. (2020) Alternating EM algorithm for a bilinear model in isoform quantification from RNA-seq data. *Bioinformatics*, **36**, 805–812.
8. Li,B. and Dewey,C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323.
9. Mayba,O., Gilbert,H.N., Liu,J., Haverty,P.M., Jhunjhunwala,S., Jiang,Z., Watanabe,C. and Zhang,Z. (2014) MBASED: allele-specific expression detection in cancer tissues and cell lines. *Genome Biol.*, **15**, 405.
10. Grant,A.D., Vail,P., Padi,M., Witkiewicz,A.K. and Knudsen,E.S. (2019) Interrogating Mutant Allele Expression via Customized Reference Genomes to Define Influential Cancer Mutations. *Sci. Rep.-UK*, **9**, 12766.
11. Fan,J., Hu,J., Xue,C., Zhang,H., Susztak,K., Reilly,M.P., Xiao,R. and Li,M. (2020) ASEP: gene-based detection of allele-specific expression across individuals in a population by RNA sequencing. *PLoS Genet.*, **16**, e1008786.
12. Harvey,C.T., Moyerbrailean,G.A., Davis,G.O., Wen,X., Luca,F. and Pique-Regi,R. (2015) QuASAR: quantitative allele-specific analysis of reads. *Bioinformatics*, **31**, 1235–1242.
13. Raghupathy,N., Choi,K., Vincent,M.J., Beane,G.L., Sheppard,K.S., Munger,S.C., Korstanje,R., Pardo-Manual de Villena,F. and Churchill,G.A. (2018) Hierarchical analysis of RNA-seq reads improves the accuracy of allele-specific expression. *Bioinformatics*, **34**, 2177–2184.
14. Khansefid,M., Pryce,J.E., Bolormaa,S., Chen,Y., Millen,C.A., Chamberlain,A.J., Vander Jagt,C.J. and Goddard,M.E. (2018) Comparing allele specific expression and local expression quantitative trait loci and the influence of gene expression on complex trait variation in cattle. *BMC Genomics*, **19**, 793.
15. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
16. Bhuiyan,S.A., Ly,S., Phan,M., Huntington,B., Hogan,E., Liu,C.C., Liu,J. and Pavlidis,P. (2018) Systematic evaluation of isoform function in literature reports of alternative splicing. *BMC Genomics*, **19**, 637.
17. Ma,C., Zheng,H. and Kingsford,C. (2021) Exact transcript quantification over splice graphs. *Algorithms Mol. Biol.*, **16**, 5.
18. Forbes,S.A., Beare,D., Boutselakis,H., Bamford,S., Bindal,N., Tate,J., Cole,C.G., Ward,S., Dawson,E., Ponting,L. *et al.* (2017) COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.*, **45**, D777–D783.
19. Tyner,J.W., Tognon,C.E., Bottomly,D., Wilmot,B., Kurtz,S.E., Savage,S.L., Long,N., Schultz,A.R., Traer,E., Abel,M. *et al.* (2018) Functional genomic landscape of acute myeloid leukaemia. *Nature*, **562**, 526–531.
20. Frazee,A.C., Jaffe,A.E., Langmead,B. and Leek,J.T. (2015) Polyester: simulating RNA-seq datasets with differential transcript expression. *Bioinformatics*, **31**, 2778–2784.
21. Srivastava,A., Sarkar,H., Gupta,N. and Patro,R. (2016) RapMap: a rapid, sensitive and accurate tool for mapping RNA-seq reads to transcriptomes. *Bioinformatics*, **32**, i192–i200.
22. Ntranos,V., Kamath,G.M., Zhang,J.M., Pachter,L. and David,N.T. (2016) Fast and accurate single-cell RNA-seq analysis by clustering of transcript-compatibility counts. *Genome Biol.*, **17**, 112.
23. Koboldt,D.C., Zhang,Q., Larson,D.E., Shen,D., McLellan,M.D., Lin,L., Miller,C.A., Mardis,E.R., Ding,L. and Wilson,R.K. (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.*, **22**, 568–576.
24. Conway,E., Prideaux,S. and Chevassut,T. (2014) The epigenetic landscape of acute myeloid leukemia. *Adv. Hematol.*, **2014**, 103175.
25. Renneville,A., Roumier,C., Biggio,V., Nibourel,O., Boissel,N., Fenaux,P. and Preudhomme,C. (2008) Cooperating gene mutations in acute myeloid leukemia: a review of the literature. *leukemia*, **22**, 915–931.
26. Daver,N., Schlenk,R.F., Russell,N.H. and Levis,M.J. (2019) Targeting FLT3 mutations in AML: review of current knowledge and evidence. *Leukemia*, **33**, 299–312.
27. Sakaguchi,M., Yamaguchi,H., Najima,Y., Usuki,K., Ueki,T., Oh,I., Mori,S., Kawata,E., Uoshima,N., Kobayashi,Y. *et al.* (2018) Prognostic impact of low allelic ratio FLT3-ITD and NPM1 mutation in acute myeloid leukemia. *Blood Adv.*, **2**, 2744–2754.
28. Schon,K. and Tischkowitz,M. (2018) Clinical implications of germline mutations in breast cancer: TP53. *Breast Cancer Res. Treat.*, **167**, 417–423.
29. Ciurea,S.O., Chilkulwar,A., Saliba,R.M., Chen,J., Rondon,G., Patel,K.P., Khogeer,H., Shah,A.R., Randolph,B.V., Perez,J.M.R. *et al.* (2018) Prognostic factors influencing survival after allogeneic

transplantation for AML/MDS patients with TP53 mutations. *Blood*, **131**, 2989–2992.
30. Thiede,C., Koch,S., Creutzig,E., Steudel,C., Illmer,T., Schaich,M., Ehninger,G. and (DSIL),D.S.L. (2006) Prevalence and prognostic impact of NPM1 mutations in 1485 adult patients with acute myeloid leukemia (AML). *Blood*, **107**, 4011–4020.
31. Wu,A.R., Neff,N.F., Kalisky,T., Dalerba,P., Treutlein,B., Rothenberg,M.E., Mburu,F.M., Mantalas,G.L., Sim,S., Clarke,M.F. *et al.* (2014) Quantitative assessment of single-cell RNA-sequencing methods. *Nat. Methods*, **11**, 41.
32. Zhang,C., Zhang,B., Lin,L.-L. and Zhao,S. (2017) Evaluation and comparison of computational tools for RNA-seq isoform quantification. *BMC Genomics*, **18**, 583.
33. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
34. Ye,K., Schulz,M.H., Long,Q., Apweiler,R. and Ning,Z. (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, **25**, 2865–2871.
35. Kouchkovsky,D. (2016) Acute myeloid leukemia: a comprehensive review and 2016 update. *Blood Cancer J.*, **6**, e441.