MINTIE: identifying novel structural and splice variants in transcriptomes using RNA-seq data

Marek Cmero[1,2,3], Breon Schmidt[1,2,4], Ian J. Majewski[5,6], Paul G. Ekert[1,2,7,8], Alicia Oshlack[1,2,4]*, Nadia M. Davidson[1,2,4]*

1. Peter MacCallum Cancer Centre, Melbourne, Vic, Australia.
2. Murdoch Children's Research Institute, Parkville, Australia.
3. Sir Peter MacCallum Department of Oncology, The University of Melbourne, Parkville, Australia.
4. School of BioSciences, University of Melbourne, Parkville, Australia.
5. Walter and Eliza Hall Institute, Parkville, Australia.
6. Faculty of Medicine, Dentistry and Health Sciences, University of Melbourne, Parkville, Australia.
7. Children's Cancer Institute, UNSW, Sydney, Australia.
8. Department of Paediatrics, University of Melbourne, Parkville, Australia.

* Joint corresponding authors, supervised this project equally

Abstract

Structural DNA variants can modify gene function by altering transcript sequences, and have been shown to be drivers in both cancer and rare diseases. Although there are now many methods to detect structural variants from Whole Genome Sequencing (WGS), RNA-sequencing (RNA-seq) remains under-utilised as a technology for the detection of gene altering structural variants. Calling fusion genes from RNA-seq data is well established, but other transcriptional variants such as fusions with novel sequence, tandem duplications, large insertions and deletions, and novel splicing are difficult to detect using existing approaches.

To identify all types of variants in transcriptomes, we developed MINTIE, an integrated pipeline for RNA-seq data. We take a reference free approach, which combines de novo assembly of transcripts with differential expression analysis, to identify up-regulated novel variants in a case sample.

We validated MINTIE on simulated and real data sets and compared it with seven other approaches for finding novel transcriptional variants. We found MINTIE was able to detect all defined variant classes at high rates (>70%) while no other method was able to achieve this.

We applied MINTIE to RNA-seq data from a cohort of acute lymphoblastic leukemia patient samples and identified several novel clinically relevant variants, including an unpartnered recurrent fusion involving the tumour suppressor gene RB1, and a tandem duplication in IKZF1, an ALL-associated gene linked to poor outcome. We further demonstrate the ability for MINTIE to identify transcriptional changes in rare disease samples associated with variants of unknown significance. We posit that MINTIE will be able to identify new disease variants across a range of cancers and other disease types.

Introduction

Rearrangements of the genome can disrupt or modify gene function and have been implicated as the causal event in disease. In cancer, somatic genomic rearrangements are common and can alter the genomic landscape to drive oncogenesis and cancer progression[1–3]. Whole genome sequencing (WGS) has successfully been used to detect structural variants (SVs) and profile their frequency in cancers[4] and rare diseases[5,6]. However, clinically relevant variants can occur alongside benign events, making prioritisation of important events difficult. This is especially true in cancers with genomic instability.

Transcriptome profiling has been used to interpret the functional impact of genomic variants through alterations in gene expression, transcript sequence, or both[7]. Some rearrangements that alter gene structure may be more effectively detected from RNA sequencing (RNA-seq). In particular, numerous computational approaches have been developed to reliably call fusion genes[8,9]. Similarly, a number of methods now exist to detect novel gene splicing[10–12] and the utilisation of RNA-seq to detect and interpret splicing variants has been shown to improve diagnostic yields in rare Mendielian disorders[13,14].

While there are many methods to detect fusion genes and novel splice variants (NSVs) from RNA-seq, transcribed structural variants (TSVs) involving a single gene, such as deletions,

inversions, internal tandem duplications (ITDs) and partial tandem duplications (PTDs) are difficult to detect, and only a handful of tools have been developed[15–17]. However, these variant types can be clinically important; for example IKZF1 deletions in Acute Lymphoblastic Leukemia[18], FLT3 internal tandem duplications (ITDs) and MLL partial tandem duplications (PTDs) in Acute Myeloid Leukemia[15–17,19]. The prevalence of TSVs in disease is not yet known, but small SVs (<100kbp) that could give rise to TSVs are frequently seen in WGS[4,20]. In addition, fusion-finding tools are likely to miss non-canonical fusions, which we define as a fusion transcript that includes non-reference sequence at the fusion boundary, or a fusion product created between a gene and an intergenic region.

Fusion finding tools typically use discordant or split read detection approaches, where reads are first mapped to either a reference genome[21,22] or transcriptome[23,24], then discordant reads are detected and a set of filtering steps are performed. While mapping information from genome alignments can be used to detect non-canonical fusions and other TSVs, most fusion finding tools do not consider these variants. However, some exceptions exist; Arriba (https://github.com/suhrig/arriba/) is a recent method that utilises STAR's alignments to detect fusions, including non-canonical fusions, as well as PTDs. SQUID[16], which is a method based on the idea of rearranging the reference genome using concordant and discordant reads, can detect non-canonical fusions and some other TSVs. This approach has also been extended to handle heterogeneous (i.e. multi-allelic) contexts[25]. These approaches rely heavily on the reference genome and generally have limitations in detecting small variants (<100bp) such as INDELs and ITDs.

An alternative to detecting variants from read alignment is to use de novo assembly. This approach attempts to reconstruct a sample's transcriptome, then aligns the assembly back to the reference genome to identify variants. Assembly is able to reconstruct transcripts containing variants of all types and sizes, which avoids potential biases caused by alignment to the reference genome. KisSplice[26] is an example of a reference free method focused on splice variants. It calls splice variants based on bubble-detection within the assembly De Bruijn graph. De novo assembly approaches typically generate a large number of transcripts, many of which will be misassemblies, leading to a potentially large false positive rate. Additionally, these methods are generally slower to run due to the computational requirements of de novo assembly. One strategy to reduce false positives and improve speed is to only consider a

specific list of target genes, a strategy employed by TAP[15]. Another approach, used by DE-kupl[27], is to perform differential expression of k-mers against a set of control samples. One limitation of the DE-kupl approach is that it compares two groups, each with multiple samples, and is therefore not appropriate for detecting variants in a single sample.

Existing methods are biased towards detecting certain classes of variants. Here we propose an alternative approach, called MINTIE, which can detect any kind of anomalous sequence insertion/deletion (≥7bp by default) or splicing (flanked by ≥20bp by default) in any gene. MINTIE is an RNA-seq pipeline that combines the advantages of full de novo assembly with differential expression to identify unique variants in a case sample versus a set of controls.

We tested MINTIE on a simulation of 1,500 variants including fusions, TSVs and NSVs and showed that it is highly sensitive across many different variant types. We compared the performance to seven other methods and found that MINTIE was the only one able to consistently detect all classes of variants we simulated at high recall rates (>70%). In addition, we ran MINTIE on RNA-seq samples from AML and normal blood, where we confirmed its sensitivity in detecting FLT3-ITDs, KMT2A-PTDs and fusions on real data. Normal (non-cancer) samples were used to assess the background rate of variants. In these samples, MINTIE reported a median of 155 genes per sample containing transcriptional variants, demonstrating a low background rate.

We used MINTIE to discover several interesting alterations in known driver or tumour suppressor genes in a cohort of paediatric acute lymphoblastic leukemia (ALLs) from the Royal Children's Hospital, Melbourne. We found a recurrent non-canonical fusion involving RB1 with downstream intergenic regions of the genome. We also discovered novel splicing of ETV6, and PTDs of IKZF1 and PAX5. Finally, we demonstrate that MINTIE is able to detect novel splicing caused by splice site mutations in the context of rare diseases.

Results

Algorithm overview

MINTIE is an approach for detecting novel transcribed variants that utilises de novo assembled transcripts, which are then prioritised for novelty using differential expression. The MINTIE pipeline (Figure 1) is divided into four main steps: transcriptome assembly of the case sample, pseudo-alignment of the case and controls to the assembled and reference transcripts, differential expression to identify upregulated novel features, and annotation of novel transcripts. One case sample and at least two controls are required as input. The pipeline is always run in a 'single case versus N controls' fashion (although multiple cases versus the same control set can be run in parallel). Conceptually, the idea of MINTIE is to perform whole-transcriptome de novo assembly, remove features consistent with reference transcripts, then perform differential expression on non-reference features versus a set of controls. Significantly over-expressed assembled transcripts are aligned to the genome and variants identified. The novelty of this approach is that we use differential expression testing to select assembled transcripts that contain highly expressed novel sequence. Therefore, we completely avoid using alignment to a reference genome to define novel sequence. Instead, we identify novel sequence by testing gene expression on equivalence classes that are unique to the assembly.

In more detail, once a de novo assembly of a case sample is obtained, by default using SOAPdenovo-Trans[28], the assembled transcripts are merged with a standard transcriptome reference (we use CHESS v2.2[29] by default), and a Salmon[30] index is created for this file. Salmon is then run on the case and N controls (where controls are ideally samples of the same tissue type). Salmon equivalence class (EC) counts are matched across all samples. ECs represent the set of transcripts that a given read is equally compatible with, and have been used as a basis for differential expression[31], differential isoform usage[32], as well as fusion detection[33]. MINTIE retains EC counts that are compatible with de novo assembled transcripts only (no reference transcripts), and all other EC counts are discarded. This step identifies novel ECs where there is no compatible match to an existing reference transcript, as ECs containing reference transcripts are assumed to be consistent with the reference, and therefore uninteresting. The expression of these novel transcript ECs are then compared by performing

differential expression testing between 1 case and N controls with edgeR[34]. This step is required to remove common but unannotated transcripts, as well as to remove erroneous assemblies where there is no discernable expression above background. Significant ECs (false discovery rate, FDR<0.05 and log 2 fold change, logFC>5 by default) are then retained, after which the transcripts corresponding to these ECs are extracted and aligned to the genome. MINTIE then performs annotation, filtering, and estimates variant allele frequency for each given variant. See Methods for more detail.

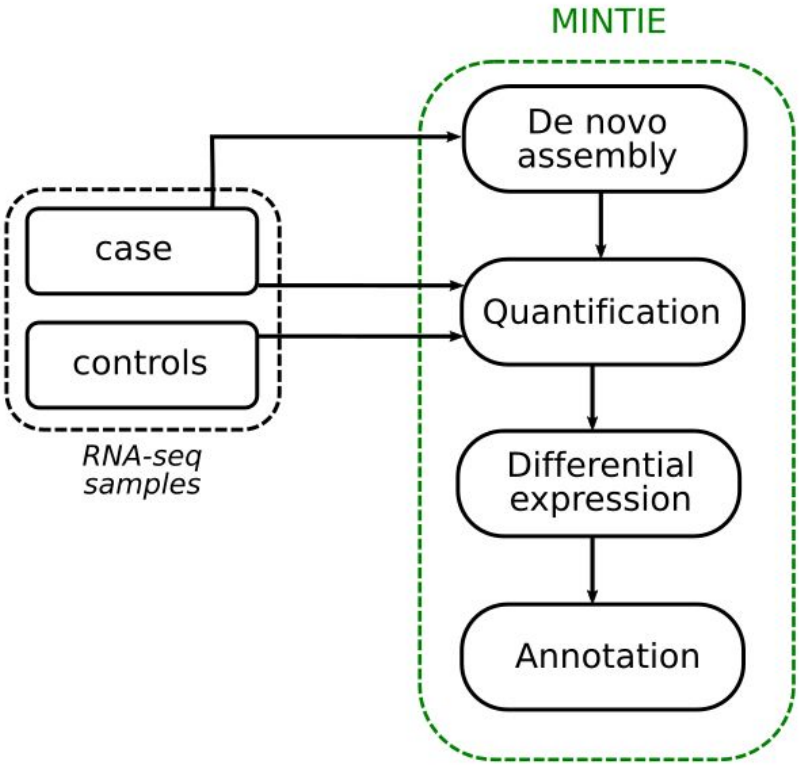


Figure 1 | *The MINTIE pipeline uses one case sample and N controls and consists of four steps: i) de novo assembly is performed on the case sample; ii) assembled* transcripts*, along with all reference transcripts, are quantified using pseudo-alignment of reads for both case and controls; iii) differential expression is performed between the case versus controls for equivalence classes that are identified as originating from assembled transcripts only, which contain novel sequence (i.e. we do not consider quantifications consistent with the reference); and iv) annotation is performed on assembled transcripts that are overexpressed in the case sample.*

MINTIE detects more types of variants than other tools

In order to test the utility of MINTIE across a wide range of variant types we simulated 1,500 variants by inserting, deleting, combining or duplicating sequences from 100 randomly selected hg38 UCSC RefSeq[35] transcripts. We then generated reads from this modified reference using ART-Illumina[36] v2.5.8. Fifteen different classes of variant were simulated (Figure 2): 500 fusions (100 each of canonical fusions, unpartnered fusions, and fusions with extended exons, novel exons and insertions at their fusion boundaries), 500 transcribed structural variants (TSVs) (100 each of deletions, insertions, ITDs, PTDs and inversions), 500 novel splice variants (NSVs) (100 each of extended, novel and truncated exons, retained introns and unannotated splice variants) and 100 unmodified background genes (see Methods). Variants were simulated at 50x coverage in a heterozygous fashion (i.e. both modified and unmodified transcripts were simulated for the variant sample, resulting in 50x variant coverage and 50x wild-type coverage). We also generated a control sample by simulating reads over wild-type (unmodified) transcripts at 100x for all transcripts used in the variant sample.

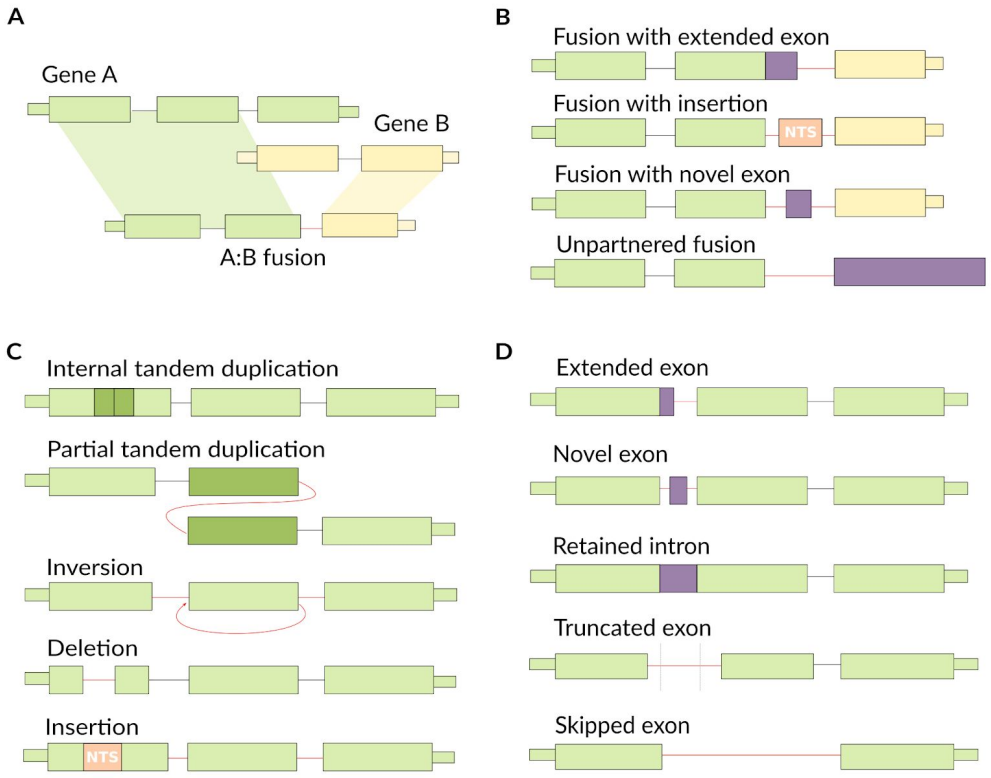


Figure 2 | *Variant types detected by MINTIE. (NTS = non-templated sequence.) a. Canonical fusions. Typically defined as a transcriptional product of two genes, joined at exon-exon boundaries. b. Non-canonical fusions. Fusions that join exon-boundaries to non-gene regions, as well as fusions without a second gene partner. c. Transcribed structural variants (TSVs). May include internal tandem duplications (ITDs), partial tandem duplications (PTDs), inversions, deletions and insertions. d. Novel* splice variants *(NSVs). Includes extended exons, novel exons, retained introns, truncated exons and skipped exons.*

We ran MINTIE and seven other variant detection methods (TAP[15], Barnacle[17], SQUID[16], JAFFA[23], Arriba (https://github.com/suhrig/arriba), KissSplice[11] and StringTie[12] on the reads generated from the simulated variants. MINTIE was run with control samples generated from the same transcripts without the variants (see methods). MINTIE detected 99.2% of fusions, 91.8% of TSVs and 90.6% of NSVs. MINTIE had lower recall on retained introns (72%) and deletions (88%). Manual inspection revealed that missed cases were mostly due to mis-assembly or insufficient read coverage. Mis-assembly could potentially be improved with another assembler. Insufficient read coverage particularly affected some deletion variants that were near the end or start of the variant transcript and thus less likely to have adequate read support. MINTIE showed similar sensitivity to the best performing tools for fusion detection (Arriba and SQUID) and novel splicing (StringTie and KissSplice). Over the complete data set MINTIE had significantly higher sensitivity than any other tool for the broad range of possible TSVs (Figure 3). Given that only 100 unmodified (wild-type) transcripts were simulated, we expected minimal false positives to be identified. No false positive calls were identified for SQUID and Arriba. JAFFA reported six false positive fusion genes (across five calls, all at low confidence). MINTIE, TAP, StringTie, Barnacle and KisSplice called 9, 10, 13 and 21 false positives respectively, while KisSplice had the highest number at 27. MINTIE, StringTie, TAP and Barnacle reported calls in background genes (those simulated without any variants) with 1, 1, 2 and 3 hits respectively. All other reported false positives were outside of simulated gene regions. Manual inspection of MINTIE's false positives revealed that these calls were largely due to poor transcript alignment (in background genes), or alignment to highly similar gene sequences (for example, to a pseudogene of the variant gene).

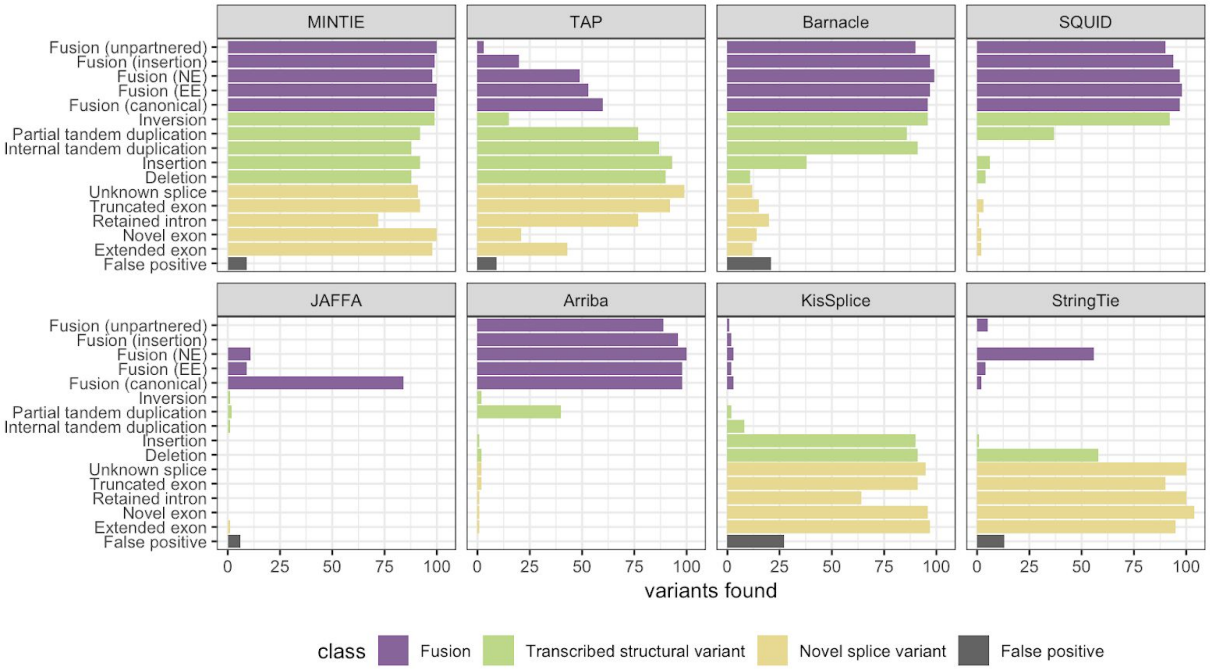


Figure 3 *| Recall and false positives for 1,500 simulated variants for tools listed. Fusions were considered detected if either (or both) fusion partner(s) were present.*

MINTIE identifies known fusions and ITDs in leucegene samples

In order to demonstrate the sensitivity of MINTIE in real samples we applied MINTIE to a set of 78 AML samples with a variety of known fusions, internal tandem duplications and partial tandem duplications (Supplementary Table 3). This included the NUP98-NSD1 fusions that were validated from Lavelle et al.[33] (N=7). We also tested MINTIE on samples containing CBFB-MYH11 fusions, RUNX1-RUNX1T1 fusions and FLT3-ITDs known to be in the core binding factor (CBF) AML data[27], as well as a cohort containing KMT2A-PTDs[37], identified by Audemard et al.[38] (Table 1). In all these analyses we used 13 normal leucegene samples as controls (5 granulocytes, 5 monocytes and 3 total white blood cells). In the AML data sets, a median of 508 variant genes was reported per sample (range 266-2365). All NUP98-NSD1 fusions were detected in the seven-patient cohort, as were the 20 RUN11-RUNX1T1 fusions in the CBF cohort. 24 of 26 CBFB-MYH11 fusions; inspection of the data revealed that the missed fusions were filtered out due to low log fold-change values of their respective ECCs. Additionally, all FLT3-ITDs were detected in the CBF and NUP98-NSD1 cohorts (default parameters needed to be adjusted for patient sample *11H022* in order to detect the small 3bp FLT3-ITD). MINTIE was able to find 16 of 24 KMT2A-PTDs in the samples identified by Audemard et al.[38] (Supplementary Table 1). The variants that were missed had low coverage (<17 reads). Of the 8 missed variants, 6 were due to insufficient assembly and 2 were filtered out by the CPM filter due to low counts in their respective ECs).

Table 1 | *Summary of true positive variants detected by MINTIE in Leucegene AML samples. *The 3bp FLT3-ITD in patient 11H022 could be detected by changing MINTIE's minimum gap size parameter to 3.*

Cohort	Gene(s)	Type	Detected	Total samples
Core-binding factor	CBFB-MYH11	Fusion	24	26
	RUNX1-RUNX1T1	Fusion	20	20
	FLT3	ITD	2*	3
NUP98-NSD1	NUP98-NSD1	Fusion	7	7
	FLT3	ITD	4	4
KMT2A-PTD	KMT2A	PTD	16	24

MINTIE detects a low number of background variants

We applied MINTIE to a set of 23 non-cancer healthy adult sorted blood cell samples obtained from Leucegene[39]. We expected these samples to have low numbers of transcriptional variants compared to cancer samples, and thus provide an estimate to the background rate of detected variants found by MINTIE. This set was composed of several cell types: T cells (N=5), B cells (N=5), granulocytes (N=5), monocytes (N=5) and total white blood cells (N=3). Each sample was run with MINTIE using all other samples of the same cell type as controls.

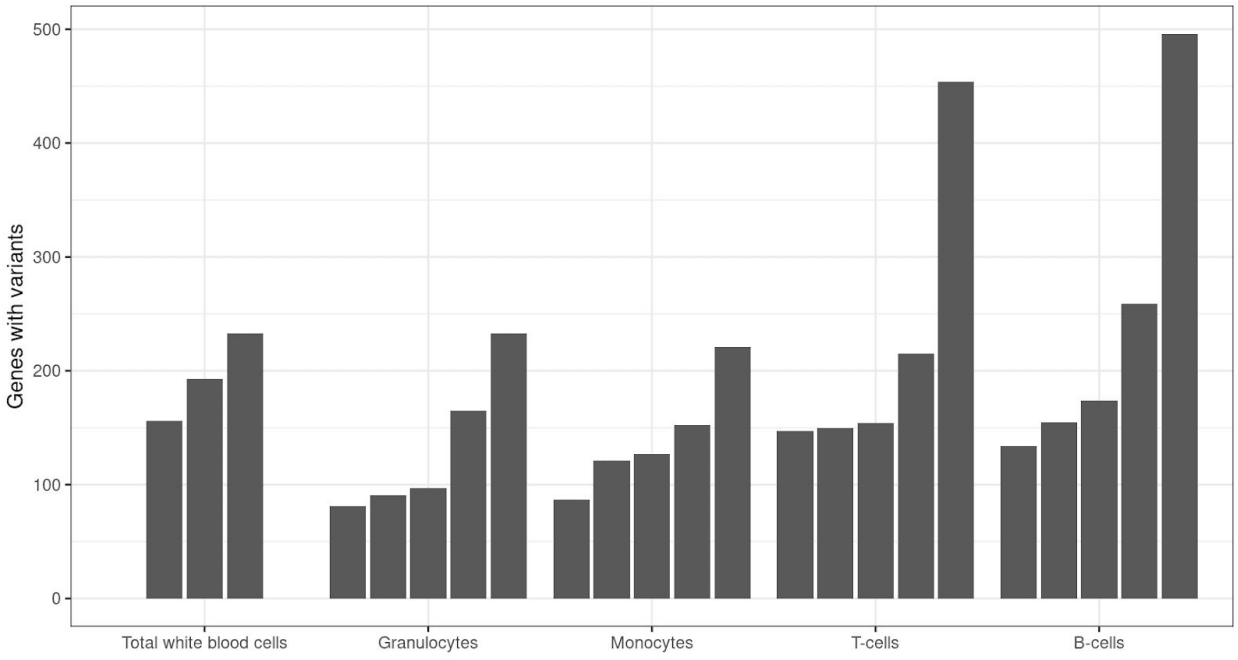


Figure 4 | *The number of genes with variants found per sample across the 23 non-cancer cell types from Leucegene samples. Each bar represents a different sample within the cell type. Samples are ordered by the total number of genes with detected variants.*

Figure 4 shows the total number of genes containing transcriptional variants identified across the non-cancer samples. In general, MINTIE had a low background variant rate, with a median of 155 affected genes across samples (range 81-496). We identified 20,203 expressed genes across the non-cancer samples (>1 CPM in at least one sample), thus the range translates to 0.4-2.6% of genes identified as containing variants. 81.5% of the 2,688 identified variant genes were protein coding (the rate in the expressed genes was 72%). 18.6% of variants were

classified as splicing variants and may represent true variation in the samples. The number of variants called was highly variable between samples, even within samples of the same type. We found only a weak correlation between library size and number of transcriptional variants detected (spearman correlation coefficient = 0.1408). A large fraction of variants (44%) were of unknown type and may be due to poor alignment of the de novo assembled transcripts to the genome. We note that one B-cell sample contained the largest number of variant calls (1,119 variants across 174 genes). Upon manual inspection, we noted that 793 of these variants were clustered in a 1MB region at the end of the long arm of chromosome 14 containing immunoglobulin genes, likely indicating that this sample had high expression of immunoglobulins, which are commonly rearranged.

Well-matched controls reduce the number of background variants

We found that using correctly matched controls was important for filtering out the normal range of transcriptional diversity in a given sample. We compared the number of variant genes observed in normal total white blood cells (TWBCs) when using different controls: the same cell type (TWBCs), as well as granulocytes, monocytes, T-cells and B-cells. Using monocytes as controls resulted in the fewest number of variant genes found across all three samples (N=517), followed by TWBCs (N=580) and granulocytes (N=655) (Supplementary Figure 3). TWBCs against themselves did not yield the lowest variant count; this is likely due to the lower number of controls available (2 for TWBCs vs. 5 for monocytes), and the high degree of gene expression similarity of monocytes to TWBCs (Supplementary Figure 1 for a PCA plot). T-cells and B-cells cluster together in gene expression, distinct from TWBCs, and using them as controls identified the most variant genes at 1126 and 1153 respectively. These results confirm that comparing against controls of similar expression profiles combined with enough control samples, resulted in fewer total number of variants found.

We also ran the cohort of 24 leucegene AML samples containing KMT2A-PTDs against 3 different sets of controls: i) 13 normal (non-cancer) controls (5 granulocytes, 5 monocytes and 3 total white blood cells), ii) a reduced set of 3 normal controls containing one sample from each cell type, and iii) a set of 13 AML samples from a different cohort. MINTIE was able to detect 16/24 variants across all three control groups (Supplementary Table 1). Differences were observed in total variants found (Supplementary Figure 2), with a median of 508 variant genes (range 266-2365) found per sample for the 13 normal controls compared to 794 variant genes

(range 501-2562) for the reduced set of 3 non-cancer control samples and a median of 213.5 variant genes (range 129-2093) found when using the 13 AML controls. If we assume most variants are background events, unrelated to cancer, this suggests that using samples with well matched transcriptomes as controls reduces the number of detected background variants. We also found some evidence that using more controls reduced the number of variants.

MINTIE identifies novel transcriptomic variants in B-ALL

In order to test MINTIE's ability to discover previously undetected events we applied MINTIE to a set of 91 samples across a cohort of 87 paediatric B-ALL patients from the Royal Children's Hospital, Melbourne Australia[40] (Supplementary Table 2). As the control group we used a subset of 34 samples from the cohort with already known driver fusions, identified by molecular test, and validated in the RNA-seq with fusion calling (Supplementary Note 1). A median of 50 variant genes was found per sample (range 20-507). We filtered on variants found in 379 recurrently altered genes (not considering immunoglobulin genes) in over 2,500 paediatric cancer samples reported in two landscape papers[41,42], and found 315 variants across 134 genes. The top recurrent genes were ETV6 (25 variants across 10 samples), IKZF2 (10 variants across 2 samples) and PAX5 (9 variants across 8 samples) (Supplementary Figure 5).

One sample contained an in-frame, single-exon tandem duplication in IKZF1. In-frame deletions of IKZF1 are recurrent in paediatric ALL and associated with poor prognosis[43]. They are believed to act in a dominant negative manner and we hypothesise that this might also be the case for the IKZF1-PTD we detected. We also found one sample with an in-frame tandem duplication of 4 exons in PAX5. A subtype of B-ALL is defined by PAX5 alterations, with a range of diverse variant types across patients (fusions, amplifications and mutations)[44].

In addition we found three samples with novel splicing of ETV6 (Figure 5B). In one sample, Case 1, we observed a unique alternative splicing event between exons 2-5, resulting in lowered expression of intervening exons, indicating a potential deletion. Karyotyping of the sample indicated a complex rearrangement of ETV6, but no fusion involving ETV6 was found.

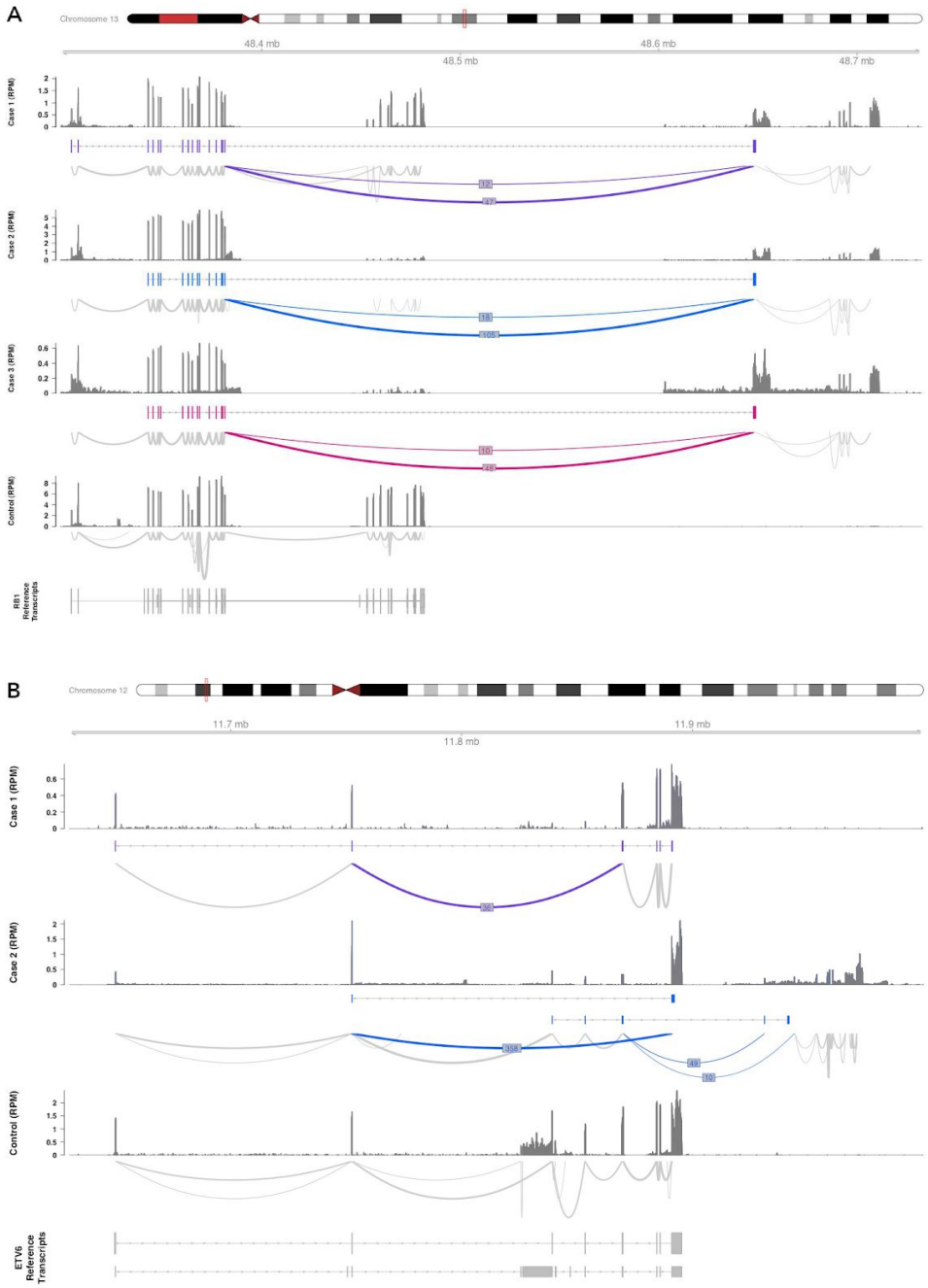


Figure 5: *Examples of novel variants detected in a cohort of paediatric B-ALL patient samples. (A) Three cases were found to have an RB1 unpartnered fusions and (B) two cases with altered splicing of EVT6. Shown for each case in top to bottom order is the read coverage, novel assembled transcript and splicing across the variant. Novel splice junctions are coloured with their corresponding number of supporting reads. For comparison, a representative control sample is shown (bottom) with corresponding expressed reference transcripts.*

The same exon skipping event was observed in a single TCGA Lung Squamous Cell Carcinoma sample, found with SeqOthello[45]. In the other two B-ALL samples, Case 2, which were from diagnosis and relapse of the same patient, we observed two novel ETV6 isoforms: splicing between exons 2-8 and splicing between exon 5 and two downstream novel intergenic exons. Expression of exons 6 and 7 was absent.

A B-ALL subtype classifier based on gene expression was run on the four samples harbouring ETV6 splicing, IKZF1-PTD and PAX5-PTD variants. All samples were found to have a strong probability of belonging to their respective class, ETV6-RUNX1-like, IKZF1 N159Y and PAX5alt (data not shown), suggesting these events may be drivers.

Interestingly, MINTIE detected an unpartnered recurrent fusion involving the tumour suppressor gene RB1 from the end of exon 17 to an intergenic region approximately 165kb downstream (Figure 5A) in three samples. In all three samples, two splice variants of the fusion were seen, which involved different novel downstream exons (Supplementary Table 2). Read coverage dropout suggested a 188kb genomic deletion involving the 3' end of RB1 in one sample, however, this was not clear in the two other samples. By querying TCGA for this event using SeqOthello[45], one of the fusion splice variants involving RB1's exon 17 and the same downstream intergenic region was detected in two other samples, a breast invasive carcinoma and an esophageal carcinoma. We hypothesise that these fusions result in a loss of function of RB1.

MINTIE identifies novel splicing and transcribed structural variants in rare disease

To demonstrate the utility of MINTIE in rare disease, we analysed RNA sequencing data from patients with rare muscle disorders from Cummings et al.[14]. In the study, diagnosis was made by identifying altered transcripts using RNA-seq in pathogenic genes with an identified variant of unknown significance. From this data set there were 10 patients that had RNA-seq available with 13 splicing variants that could potentially be detected by our method. We used 10 muscle RNA-seq samples from GTEx[46] as controls, and ran MINTIE transcriptome wide using RNA-seq alone without knowledge of which genes had DNA mutations. MINTIE identified the pathogenic found the correct splice variants for 8/13 events (Table 4).

Table 4 | *List of variants detectable by MINTIE from Cummings et al.[14] rare muscle disease data.*

Patient	Gene	Junction 1	Junction 2	DNA variant	RNA variant	Detected Variant
E2	NEB	chr2:151687733-151690727		SNP	skipped exon	Y
	NEB	chr2:151687733-151688175		SNP	extended exon	Y
	NEB	chr2:151687733-151688007		SNP	extended exon	Y
C9	NEB	chr2:151531896-151533330		SNP	extended exon	Y
N25	NEB	chr2:151498352-151498491		SNP	novel exon	N (failed novel junction criteria)
N31	COL6A1	chr21:45989778-45989894	chr21:45989965-45990258	SNP	novel exon	Y
N32	COL6A1	chr21:45989778-45989894	chr21:45989965-45990258	SNP	novel exon	Y
C11	RYR1	chr19:38467720-38468966		SNP	truncated exon	Y
C1	POMGNT1	chr1:46194651-46195811		SNP	skipped exon	N (low CPM)
	POMGNT1	chr1:46189357-46189458		SNP	retained intron	Y
E4	TTN	chr2:178580609-178581906	chr2:151498552-151499298	INDEL	skipped exon	N (not assembled)
N22	TTN	chr2:178777317-178777460		SNP	truncated exon	N (low CPM)
C3	DMD	chrX:31729748-31819975		inversion-deletion	skipped exon	N (EC not significant)

A novel NEB exon in patient N25 was missed due to an overlapping exon from RIF1, as MINTIE will only call novel exons in regions where no annotated exon is present. Two other variants were filtered out due to low expression (exon skipping in POMGNT1 and exon truncation in TTN), while an exon skipping event in TTN which was only supported by 2 junction reads, failed to be assembled. A transcript harbouring the DMD skipped exon in patient C3 was not found to be significant in the DE step as a sizable number of reads supporting the variant transcript were found in the controls. Manual inspection revealed that the assembled transcript for this DMD case contained two variants, one being the target variant (the skipped exon), and a secondary variant (a homozygous SNP). The secondary variant was not in the reference transcriptome but was found in 9/10 controls. Cases such as this are rare, but represent one potential weakness

of an EC-based method, whereby multiple novel variants assembled into the same transcript cannot be resolved.

Patient samples C2 and C4 were identified with large-scale inversions in the DMD gene using DNA sequencing that manifested as lower exonic coverage across regions of the gene. Cummings et al. did not report an exon skipping RNA-seq feature, and thus we could not check specific boundaries in the RNA-seq. However MINTIE did find novel transcriptomic variants in DMD in both of these samples. Running BLAT[47] on these soft-clip sequences revealed that they aligned on the opposite strand to the gene sequence, supporting an inversion variant. This is an example where MINTIE can provide insight into DNA-level structural variants solely from the RNA-seq.

Computational Performance

Computer resources are often seen as a limitation to the application of de novo assembly in variant detection. In MINTIE, we used SOAPdenovo-trans as the default assembler due to its computational efficiency compared to other assemblers. MINTIE also leverages pseudo-alignment to minimise compute times. To benchmark performance, we selected three samples (containing the least, median and most variants called) from the Leucegene KMT2A-PTD cohort and ran MINTIE against the set of 13 controls previously described. Tests were run on a machine with 32 cores and 190GB of available memory with MINTIE's bpipe pipeline concurrency limited to 32. The variant numbers also corresponded with the number of read pairs, ranging from 67 million to 113 million. Compute times were between 3 hours and 24 minutes and 7 hours and 4 minutes, where the run time scaled with the number of reads. Maximum memory usage did not appear to be significantly correlated with the read number, as all samples were between roughly 90-100GB max memory usage. The first three steps, dedupe, trim and assemble steps account for approximately over half the run time, and most of the memory usage. For instance, in sample 10H007, these stages took 4 hours and 4 minutes, and used at maximum 102.64GB of memory. The rest of the pipeline finished in 3 hours and used at maximum 16.95GB of memory. These benchmarking results demonstrate that MINTIE's utility is not hindered by computational considerations, and a large number of samples can be processed on a medium sized stand-alone server in a reasonable time frame.

Table 2 | *Compute times for three selected samples from the Leucegene KMT2A-PTD cohort, including the sample with the least variants called, the median, and the most. Reads were 100bp and samples were run against 13 AML controls with 8 cores on a machine with 190GB of available memory and 32 cores.*

Sample	Variants	Read pairs	Time	Max mem (GB)
08H012	128 (min)	67,789,336	3h24m	101.47
07H155	214 (median)	98,946,255	6h42m	89.85
10H007	2093 (max)	113,017,564	7h04m	102.64

Discussion

Here we present the MINTIE pipeline, a catch-all caller for transcribed variants ≥ 7bp in size. MINTIE frames variant calling as a problem of detecting sample specific expression of transcripts with novel sequence, without the use of a reference genome. MINTIE achieves this by performing a *de novo* assembly to reconstruct all transcripts expressed in a sample. It then leverages the unique approach of using equivalence class counts to reliably detect differential expression of unannotated sequence compared to a set of controls. Once novel transcripts are identified, MINTIE performs a number of steps to filter out transcripts found in the reference annotation and finally annotates the remaining variants into subtypes.

We demonstrated MINTIE's ability to find a diverse range of fusions, structural and splice transcriptomic variants by simulating 1500 variants, where MINTIE achieved >93% detection recall. We further benchmarked fusion callers, transcribed structural variant callers and splice assemblers and found that MINTIE detected the full range of simulated variant types, while the other methods could not. We further validated the ability of MINTIE to detect fusions and duplications in a set of validated AML cancer samples. Even for fusion finding, where many reliable methods exist, MINTIE's sensitivity was comparable. Thus, we propose that MINTIE could be used as a single tool to detect any large alterations, replacing the need to run multiple tools.

Using the Leucegene data, we showed that well matched controls are important for reducing the number of false discoveries. For many cases, other disease samples with similar expression profiles can be used as controls. In this instance we advise using samples with known driver variants, so that the unique variants to be detected are unlikely to be present in the controls. Identical variants that are present in the set of controls will limit detection sensitivity in the case sample.

MINTIE has the potential to discover a wide range of important events that have thus far evaded detection due to the limitation of conventional analysis approaches. We demonstrated this by running MINTIE on a cohort of 87 B-ALL samples, where we found interesting events in disease relevant genes including 3 samples with the same RB1 unpartnered fusion, IKZF1 and PAX5 PTDs, and 2 samples with altered splicing of ETV6.

It is important to consider some of the limitations of MINTIE. The methodology has been optimised to detect a large number of variant types and thus some specialised approaches geared at detecting fewer variant types outperform MINTIE in accuracy and runtime. Additionally, MINTIE requires sufficient supporting reads for accurate de novo assembly and to obtain significance when testing differential up-regulation. A further consideration is that MINTIE is, to some extent, reliant on appropriate control samples. Finally, MINTIE is a method based on RNA sequencing, and can only identify structural rearrangements that affect transcribed regions. Similarly, lowly expressed genes or variants that knock out the expression of a gene will be difficult to detect.

MINTIE utilises de novo transcriptome assembly to reconstruct transcript sequences. While this has the advantage of allowing complex variants to be detected, the run-time and RAM requirements are generally higher than many of the alignment based approaches. However, long read transcriptome sequencing technology is readily being adopted and MINTIE could be easily extended to use polished long read data in hybrid with short reads, avoiding the disadvantages of assembly.

Here we have presented a novel, fast, transcriptome-wide approach for detecting rare transcribed variants, and have demonstrated the use of the tool through simulations, real cancer and non-cancer samples and have demonstrated the detection of a number of alterations, some

of which are clinically relevant. Although we have used RNA-seq from cancer to validate MINTIE, it can be applied to diseases other than cancer, where SVs or altered splicing are suspected. We propose that MINTIE will be used to identify novel drivers and loss-of-function variants in cancer and in rare diseases. Unlike existing methods, which are targeted at specific types of rearrangements or splicing, MINTIE is a catch-all approach. We have tested MINTIE on several types of variants, but there are potentially many other types which MINTIE could detect. For example, retrotransposon insertions into genes or complex events involving multiple rearrangements, which could be generated by chromoplexy or chromothripsis. Little is known about the frequency or characteristics of transcribed variants, but we now have a method to uncover and study them.

Methods

Pipeline detail

De novo assembly: RNA-seq reads of the case sample are de-duplicated using fastuniq[48] v1.1 and trimmed using Trimmomatic[49] v0.39 and then de novo assembled using SOAPdenovo-trans[28] v1.03 with kmers of 49 and 79. (Optionally, Trinity[50] or rna-SPAdes[51] may be used for assembly, or the user may provide their own assembly.) The different kmer assemblies are merged and the sequences are deduplicated.

Quantification: all assembled transcripts are merged with the CHESS[29] v2.2 transcriptome reference sequence, which is then indexed by Salmon[30] v0.14. Salmon is then run on this index for the case and control samples in single-end mode (otherwise short assembled variants may not be correctly counted). Sequence bias is corrected (--seqBias) and equivalence class counts (ECCs) are extracted (--dumpEq).

Differential expression: equivalence classes (ECs) are matched between all samples by membership (i.e. an EC composed of transcripts A, B and C in sample 1 is given an identifier, and will be considered the same EC as one composed of the same transcripts in sample 2). Library sizes are calculated prior to any filtering. ECs containing *only* assembled transcripts (i.e.

the EC does not contain any transcripts from the reference) are kept and, after light filtering (CPM > 0.1 in the case sample), the case sample is compared to the controls using edgeR[34] v3.26.5. Dispersions are estimated using the classic (non-GLM) model[52], and transcripts are fitted using edgeR's GLM quasi-likelihood fit function[53] with the robust flag set to true[54]. edgeR's GLM likelihood-ratio test[55] is used to perform differential expression. All transcripts associated with significant ECs (FDR < 0.05 and logFC > 5) are retained.

Annotation: all significant transcripts are aligned using GMAP[56] (2019-05-12) to the hg38 p9 genome with all alternative transcripts removed. All aligned transcripts are extracted and compared to the CHESS[29] v2.2 reference GTF to identify fusions, TSVs and novel splicing. We require the aligned transcript to contain at least 30 base-pairs of aligned sequence and 30% of the transcript maps to somewhere on the genome. Transcripts not overlapping any known reference exons are discarded. Broadly, we then set the criteria for variants where: i) the transcript contains an insertion or deletion of at least 7 base-pairs, ii) the transcript has a soft-clip or hard-clip (based on the GMAP's alignment) that is least 20 base-pairs in length, iii) the transcript contains a splice junction not seen in the reference transcriptome, and iv) the transcript contains a block of at least 20 base-pairs that is not normally transcribed in the reference genome. Passing category (i), (ii), (iii) and/or (iv) will retain the transcript as potentially harbouring a candidate variant. We discard any variants matching category (iv) where no splicing occurs (i.e. the transcript is aligned as an unspliced block); even when considering retained intron variants, we require assembly of a transcript where some splicing occurs elsewhere in the transcript.

Annotated variants are further refined by matching novel blocks with novel splice junctions and identifying valid donor/acceptor sites for novel junctions (defined as observing the sequences AG/GT or AC/CT before the splice junction). (The motif checking functionality is used by default, but may be disabled by the user.) As extended/novel exons should create novel exon boundaries, we consolidate these variants and discard any extended/novel exons not supported by a corresponding novel junction. Novel and extended exons must be accompanied by at least one novel junction at either (or both) ends of the block and have a valid donor or acceptor motif. Alignment gaps (of at least 7bp) within single exons are checked to capture potential novel intron, and exon ends are checked for truncation by at least 20bp. A single transcript may have

multiple annotated variants, but only those matching specified filters are retained. The variant sizes listed above for filtering are defaults only and may be adjusted in MINTIE.

Selecting appropriate control samples

Control samples ideally have the same transcriptional profile as the case samples but without the variants of interest. Controls can be from normal samples from the same tissue type as the case sample if they still have a similar transcriptional program as the case. As RNA sequencing data from the given tissue may be difficult to obtain, other samples from the same cancer type (and same tissue type) can be used as controls. However, using other cancer samples may be a better choice with the caveat that recurrent events might be missed. At least 2 control samples must be used, however we recommend using >10.

Simulations

We simulated 1500 variants using code available in the MINTIE source code repository (https://github.com/Oshlack/MINTIE/tree/master/simu/run_simu.py). Simulations were generated by extracting sequence from the transcripts listed in the hg38 UCSC RefSeq reference, and simulating reads from the resulting sequence. 100 variants from 15 variant types were generated (five fusion types: canonical, extended exon, novel exon, with insertion and unpartnered, five TSV types: insertions, deletions, ITDs, PTDs and inversions, and five novel splice variants: extended exons, novel exons, truncated exons, skipped exons and retained introns. Only transcripts from genes that did not overlap any other genes were used in the simulation. Additionally, each transcript had to have at least 3 exons to be considered as a simulation transcript.

All fusions were simulated by selecting the first two and the last two exons from two random transcripts from different genes, and inserting the intervening sequence. Canonical fusions contained no intervening sequence, while fusions with extended exons inserted 30-200bp of intronic sequence from the end of the second exon of the first transcript. Similarly, fusions with novel exons contained intronic sequence 30-200bp downstream with a size of 30-200bp. Non-canonical fusions with insertions were generated by inserting 7-50bp of randomly-generated sequence between the two fusion transcripts.

Small TSVs were generated by inserting, duplicating or deleting sequence within randomly selected exons from randomly selected transcripts. These small variant types were between 7 and 50 base-pairs and had to reside at least 10bp within the exon. Inversions and partial-tandem duplications were generated by selecting 1-3 random exons within a transcript and either inverting or duplicating their sequence in tandem. Lastly, splice variants were generated by extending or placing novel exons downstream of a randomly selected exon. To ensure that novel or extended exons did not overlap exons from other transcripts (or downstream exons of the same transcript), each candidate exon was checked for these potential overlaps (which would otherwise result in obfuscation of the variant, or the wrong variant type being created). Novel junction variants were created by selecting a random pair of exons and checking whether an existing junction existed between them (creating a transcript with this junction if not). Two randomly-selected neighbouring exons were both truncated at their facing ends (end and start respectively) by 30-200bp. Retained introns included a random intronic sequence from a given transcript that was >30bp. The presence of correct splicing motifs was not considered for the simulation.

In addition to each variant gene, the sequence to the unaltered wild-type gene was added to the simulated case sample's reference. An additional 100 unaltered background genes were also added to the case sample. A control sample reference was also generated, which included the unaltered wildtype sequence only for all simulated transcripts. ART-illumina[36] v2.5.8 was run on the corresponding references with 100bp paired-end reads with a fragment size of 300 and coverage of 50 (transcripts should thus have an effective coverage of 100, given the bi-allelic reference containing variant and wildtype transcripts).

The simulated variant fastq files were run through the MINTIE pipeline using Trinity as the assembler. Motif checking was disabled and test mode was enabled. The only difference between test mode and standard mode is that test mode uses an exact test and dispersion is set to 0.1, due to the low number of genes being tested and lack of replicates. All other parameters were left at default. Seven other tools were run on the resulting simulation output: TAP[15] (commit 8940e45), Barnacle[17] v1.0.4, SQUID v1.5[16], JAFFA[23] v1.07, Arriba (https://github.com/suhrig/arriba/) v1.1.0, KisSplice[26] v2.4.0 and StringTie[12] v1.3.6. TAP was run using the tap2.py pipeline with kmer lengths of 49 and 79, and `--max_diff_splice` set to 4 to bypass motif checking. Other parameters were kept at default. The transcript reference was

generated with PAVfinder's extract_transcript_sequence.py script, using the UCSC RefGene sequence reference as input. Barnacle was run with its default hg19 reference. SQUID was run using STAR's alignments. JAFFA was run using its direct mode and default reference. All fasta output from KisSplice was merged and aligned to the genome using GMAP. References used in generating simulations were used whenever possible. Default parameters were used for each tool unless otherwise specified. Samtools v1.8, blat v36.1, STAR v2.5.3a and bwa v0.7.17 were used by all methods that required these tools.

Variants were counted as true positives for all tools run for the benchmarking analysis if a variant was reported in the same gene name as the simulated variant (for MINTIE, Arriba and TAP). Gene coordinates were used for all other tools where gene names were not reported (SQUID, StringTie and KissSplice) or where using specific references proved prohibitive (Barnacle). Only transcripts marked as novel were considered for StringTie. Sequence results from KisSplice were aligned with GMAP[56] (2019-05-12), and aligned coordinates were extracted in order to obtain variant positions. As long as a variant coordinate was found within the expected gene region, this was counted as a hit. Only one gene/coordinate match was needed (of a partnered fusion pair) to be called a true positive. As Barnacle was run using a hg19 reference, we used UCSC liftover (https://genome.ucsc.edu/cgi-bin/hgLiftOver) to convert the hg38 simulation coordinates to hg19 (6 TSVs/NSVs and 9 fusion gene coordinates were partially deleted in the liftover; the TSVs/NSVs could not be counted, while the fusion were only affected at one coordinate and could still be counted).

Running MINTIE on Leucegene samples

MINTIE was run on four Leucegene cohorts: the NUP98-NSD1 fusion cohort[57] , the CBF AML cohort[58], the KMT2A-PTD cohort[37] and the non-cancer peripheral blood (normals) data set[39]. We used the patient IDs identified by Audemard et al.[59] to identify the samples containing KMT2A PTDs, as these were not disclosed in the original paper. For all AML data sets, we used the normal data as controls (including all 24 samples). We also ran the KMT2A-PTD data set against a set of randomly selected AML samples from the CBF cohort (see Supplementary Table 3 for the control samples that were used). The normal samples were run individually against the remaining samples from that cell type. (The pooled peripheral CD34+ blood cells were not considered for this analysis.)

Salmon[30] (v0.14) was run separately on the leucegene normal adult peripheral blood samples in paired-end mode with default parameters and the --seqBias flag. Counts were summarised at the gene level using tximport[60] (v1.12.3) with the countsFromAbundance='lengthScaledTPM' parameter. The voom function from Limma[61] v3.42.0 was used to normalise libraries, and the base prcomp function (R v3.6.2) was used to perform principal component analysis.

Running MINTIE on B-ALL samples

B-ALL and T-ALL samples used in this study were obtained from the Royal Children's Hospital; data is described in Brown et al[40]. 7 samples were excluded due to significantly different sequencing characteristics (shorter read lengths and using an unstranded protocol). Patient samples were selected as controls if they had a positive molecular test and the driver fusions were identified in the RNA-seq data. If there were multiple samples obtained from the same patient, these were not considered as controls. This resulted in 34 controls (see Supplementary Note 1 for a list).

Running MINTIE on rare disease samples

MINTIE was run on each of the 12 available samples that had RNA-variants described in Cummings et al. against a set of 10 muscle sample controls from GTeX[46]. Due to the relatively short read size (76bp), we selected k-mer sizes of 69, 49 and 29 for SOAPdenovo-trans. Motif checking was turned off, due to the presence of SNP-induced splice sites described in the study. We manually inspected variants called and transcripts assembled in order to determine whether a given variant was called.

Code availability

- MINTIE software: https://github.com/Oshlack/MINTIE.
- Code for generating paper analyses and figures: https://github.com/Oshlack/MINTIE-paper-analysis (which can be viewed at https://oshlacklab.com/MINTIE-paper-analysis).

Data availability

- The simulation is available from 10.5281/zenodo.3825926, or can be reproduced using the run_simu.py script under https://github.com/Oshlack/MINTIE/tree/master/simu, using the same random seeds as the fullsimu_params.ini file.
- The Leucegene cohorts are available on the Sequence Read Archive: the NUP98-NSD1 fusion cohort[57] (GSE49642, GSE67039 and GSE52656), the CBF AML cohort[58] (GSE49642, GSE52656, GSE62190, GSE66917 and GSE67039), the KMT2A PTD cohort[37] (accessions GSE49642, GSE52656, GSE66917 and GSE67039) and the non-cancer peripheral blood (normals) data set[39] (accession GSE51984).
- The paediatric B-ALL samples are available from the European Genome-Phenome Archive (accession number EGAS00001004212).
- The rare disease and control data are available from dbGaP under accession IDs phs000655.v3.p1 (SRA IDs SRR5051543, SRR5038795, SRR5051356, SRR5020927, SRR5039275, SRR5034854, SRR5020918, SRR5033001, SRR5034830, SRR5034681, SRR5034787 and SRR5039335) and phs000424.v6.p1 (SRR809444, SRR810225, SRR811771, SRR813656 SRR813983, SRR809595, SRR810249, SRR812773, SRR813802 and SRR815020) respectively.

Acknowledgements

We would like to thank Jinze Liu and Xiaofei Zhang for querying TCGA with SeqOthello for us, and help from Christoffer Flensburg on interpreting variants and software suggestions. Computing resources for this work were provided by MCRI and The University of Melbourne Science IT. This work was funded by NHMRC project grant APP1140626.

Authors' contributions

MC: Formal Analysis, Methodology, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; BS: Visualization; IJM: Writing – Review & Editing, PGE: Writing – Review & Editing; AO: Conceptualization, Supervision, Methodology, Writing – Original Draft

Preparation, Writing – Review & Editing; NMD: Conceptualization, Supervision, Methodology, Software, Writing – Original Draft Preparation, Writing – Review & Editing.

References

1. Saito, M. *et al.* Development of Lung Adenocarcinomas with Exclusive Dependence on Oncogene Fusions. *Cancer Res.* **75**, 2264–2272 (2015).
2. Patch, A. *et al.* Whole-genome characterization of chemoresistant ovarian cancer. *Nature* 489–494 (2015) doi:10.1038/nature14410.
3. Grimwade, D. *et al.* Refinement of cytogenetic classification in AML Younger adult patients treated in UKMRC. *Blood* **116**, 354–366 (2010).
4. Li, Y. *et al.* Patterns of somatic structural variation in human cancer genomes. *Nature* **578**, 112–121 (2020).
5. Sanchis-Juan, A. *et al.* Complex structural variants in Mendelian disorders: identification and breakpoint resolution using short- and long-read genome sequencing. *Genome Med.* **10**, 95 (2018).
6. Holt, J. M. *et al.* Identification of Pathogenic Structural Variants in Rare Disease Patients through Genome Sequencing. *bioRxiv* 627661 (2019) doi:10.1101/627661.
7. Cancer, T. *et al.* Genomic basis for RNA alterations in cancer. *Nature* **578**, 129–136 (2020).
8. Haas, B. J. *et al.* Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. *Genome Biol.* **20**, 1–16 (2019).
9. Kumar, A. *et al.* Substantial interindividual and limited intraindividual genomic diversity among tumors from men with metastatic prostate cancer. *Nat. Med.* **22**, 1–13 (2016).
10. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–78 (2012).

11. Sacomoto, G. A. T. *et al.* KISSPLICE: de-novo calling alternative splicing events from RNA-seq data. *BMC Bioinformatics* **13**, 1–12 (2012).
12. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
13. Gonorazky, H. D. *et al.* Expanding the Boundaries of RNA Sequencing as a Diagnostic Tool for Rare Mendelian Disease. *Am. J. Hum. Genet.* **104**, 1007 (2019).
14. Cummings, B. B. *et al.* Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Sci. Transl. Med.* **9**, (2017).
15. Chiu, R., Nip, K. M., Chu, J. & Birol, I. TAP: a targeted clinical genomics pipeline for detecting transcript variants using RNA-seq data. *BMC Med. Genomics* **11**, 79 (2018).
16. Ma, C., Shao, M. & Kingsford, C. SQUID: Transcriptomic structural variation detection from RNA-seq. *Genome Biol.* **19**, 1–16 (2018).
17. Swanson, L. *et al.* Barnacle: detecting and characterizing tandem duplications and fusions in transcriptome assemblies. *BMC Genomics* **14**, 550 (2013).
18. Mullighan, C. G. *et al.* Deletion of *IKZF1* and Prognosis in Acute Lymphoblastic Leukemia. *N. Engl. J. Med.* **360**, 470–480 (2009).
19. Bolouri, H. *et al.* The molecular landscape of pediatric acute myeloid leukemia reveals recurrent structural alterations and age-specific mutational interactions. *Nat. Med.* (2017) doi:10.1101/125609.
20. An integrated map of structural variation in 2,504 human genomes | Nature. https://www.nature.com/articles/nature15394.
21. STAR-Fusion: Fast and Accurate Fusion Transcript Detection from RNA-Seq | bioRxiv. https://www.biorxiv.org/content/10.1101/120295v1.abstract.
22. Kim, D. & Salzberg, S. L. TopHat-Fusion: An algorithm for discovery of novel fusion

transcripts. *Genome Biol.* **12**, (2011).

23. Davidson, N. M., Majewski, I. J. & Oshlack, A. JAFFA: High sensitivity transcriptome-focused fusion gene detection. *Genome Med.* **7**, 43 (2015).
24. Melsted, P. *et al.* Fusion detection and quantification by pseudoalignment. *bioRxiv* 166322 (2017) doi:10.1101/166322.
25. Qiu, Y., Ma, C., Xie, H. & Kingsford, C. Detecting transcriptomic structural variants in heterogeneous contexts via the Multiple Compatible Arrangements Problem. *Algorithms Mol. Biol.* **15**, 9 (2020).
26. Sacomoto, G. A. T. *et al.* KISSPLICE: de-novo calling alternative splicing events from RNA-seq data. *BMC Bioinformatics* **13**, 1–12 (2012).
27. Philippe, N. *et al.* Exhaustive capture of biological variation in RNA-seq data through k-mer decomposition. *bioRxiv* 12–14 (2017) doi:10.1101/122937.
28. Xie, Y. *et al.* SOAPdenovo-Trans: De novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics* **30**, 1660–1666 (2014).
29. Pertea, M. *et al.* CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome Biol.* **19**, 332825 (2018).
30. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 021592 (2017).
31. Yi, L., Pimentel, H., Bray, N. L. & Pachter, L. Gene-level differential analysis at transcript-level resolution. *Doi.Org* 190199 (2017) doi:10.1101/190199.
32. Cmero, M., Davidson, N. M. & Oshlack, A. Using equivalence class counts for fast and accurate testing of differential transcript usage. *F1000Research* **8**, 1–30 (2019).
33. Vu, T. N. *et al.* A fast detection of fusion genes from paired-end RNA-seq data. *BMC*

Genomics 1–13 (2018) doi:10.1186/s12864-018-5156-1.

34. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinforma. Oxf. Engl.* **26**, 139–40 (2010).
35. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733-745 (2016).
36. Huang, W., Li, L., Myers, J. R. & Marth, G. T. ART: a next-generation sequencing read simulator. *Bioinformatics* **28**, 593–594 (2012).
37. Lavallée, V.-P. *et al.* The transcriptomic landscape and directed chemical interrogation of MLL-rearranged acute myeloid leukemias. *Nat. Genet.* **47**, 1030–1037 (2015).
38. Audemard, É. *et al.* Target variant detection in leukemia using unaligned RNA-Seq reads. *bioRxiv* 295808 (2018) doi:10.1101/295808.
39. Pabst, C. *et al.* GPR56 identifies primary human acute myeloid leukemia cells with high repopulating potential in vivo. *Blood* **127**, 2018–2027 (2016).
40. Brown, L. M. *et al.* The application of RNA sequencing for the diagnosis and genomic classification of pediatric acute lymphoblastic leukemia. **4**, 1–3 (2020).
41. Gröbner, S. N. *et al.* The landscape of genomic alterations across childhood cancers. *Nature* **555**, 321–327 (2018).
42. Ma, X. *et al.* Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours. *Nat. Publ. Group* (2018) doi:10.1038/nature25795.
43. Mullighan, C. G. *et al.* Deletion of IKZF1 and Prognosis in Acute Lymphoblastic Leukemia. *N. Engl. J. Med.* **360**, 470–480 (2009).
44. Gu, Z. *et al.* PAX5-driven subtypes of B-progenitor acute lymphoblastic leukemia. *Nat. Genet.* doi:10.1038/s41588-018-0315-5.

45. Yu, Y. *et al.* SeqOthello: Query over RNA-seq experiments at scale. *bioRxiv* 258772 (2018) doi:10.1101/258772.

46. Consortium, T. Gte. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).

47. BLAT—The BLAST-Like Alignment Tool. https://genome.cshlp.org/content/12/4/656.abstract.

48. Xu, H. *et al.* FastUniq: A Fast De Novo Duplicates Removal Tool for Paired Short Reads. *PLOS ONE* **7**, e52249 (2012).

49. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

50. Haas, B. J. *et al. De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity*. *Nat Protocols* vol. 8 (2014).

51. Bushmanova, E., Antipov, D., Lapidus, A. & Prjibelski, A. D. rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. *GigaScience* **8**, giz100 (2019).

52. Chen, Y., Lun, A. T. L. & Smyth, G. K. Differential Expression Analysis of Complex RNA-seq Experiments Using edgeR. in *Statistical Analysis of Next Generation Sequencing Data* (eds. Datta, S. & Nettleton, D.) 51–74 (Springer International Publishing, 2014). doi:10.1007/978-3-319-07212-8_3.

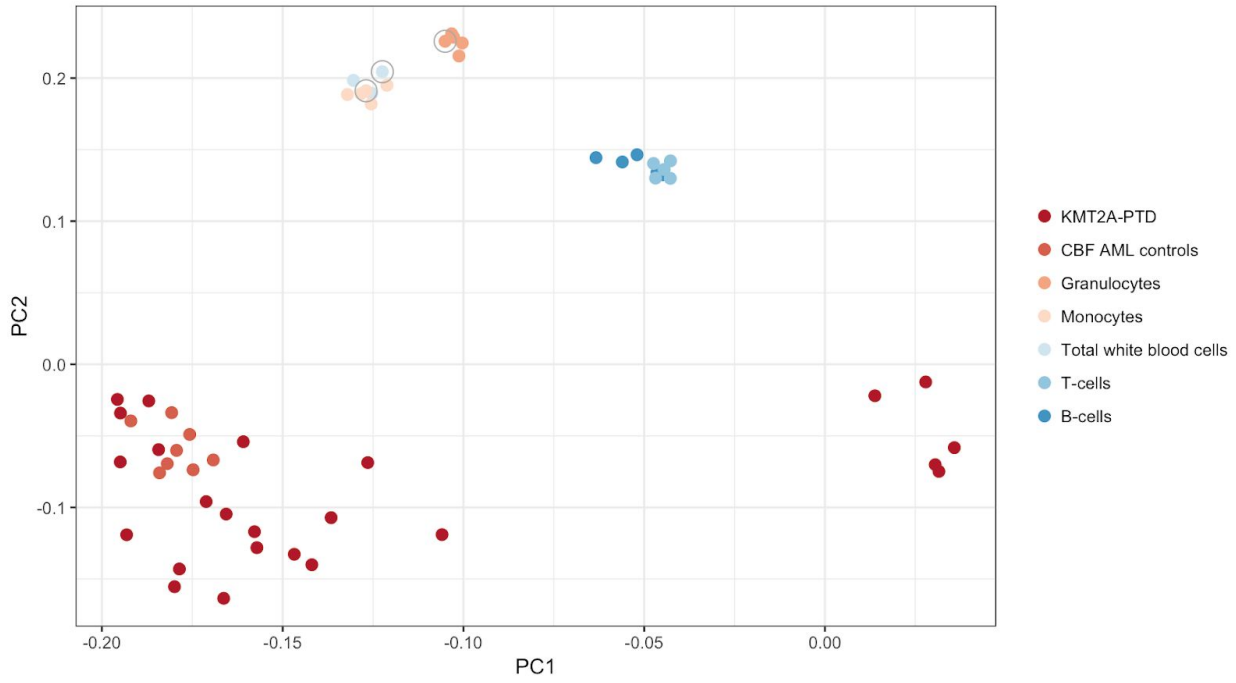
53. Lund, S. P., Nettleton, D., McCarthy, D. J. & Smyth, G. K. Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunken dispersion estimates. *Stat. Appl. Genet. Mol. Biol.* **11**, (2012).

54. Phipson, B., Lee, S., Majewski, I. J., Alexander, W. S. & Smyth, G. K. ROBUST HYPERPARAMETER ESTIMATION PROTECTS AGAINST HYPERVARIABLE GENES

AND IMPROVES POWER TO DETECT DIFFERENTIAL EXPRESSION. *Ann. Appl. Stat.* **10**, 946–963 (2016).

55. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* **40**, 4288–4297 (2012).
56. Wu, T. D., Reeder, J., Lawrence, M., Becker, G. & Brauer, M. J. GMAP and GSNAP for Genomic Sequence Alignment: Enhancements to Speed, Accuracy, and Functionality. in *Statistical Genomics: Methods and Protocols* (eds. Mathé, E. & Davis, S.) 283–334 (Springer, 2016). doi:10.1007/978-1-4939-3578-9_15.
57. Lavallée, V. P. *et al.* Identification of MYC mutations in acute myeloid leukemias with NUP98-NSD1 translocations. *Leukemia* **30**, 1621–1624 (2016).
58. Thompson, M. P., Waters, T. M., Kaplan, E. K., Mckillop, C. N. & Martin, M. G. RNA-sequencing analysis of core binding factor AML identifies recurrent ZBTB7A mutations and defines RUNX1-CBFA2T3 fusion signature. **128**, 872–875 (2016).
59. Audemard, É. *et al.* Target variant detection in leukemia using unaligned RNA-Seq reads. *bioRxiv* 295808 (2018) doi:10.1101/295808.
60. Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research* **4**, 1521 (2016).
61. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47–e47 (2015).

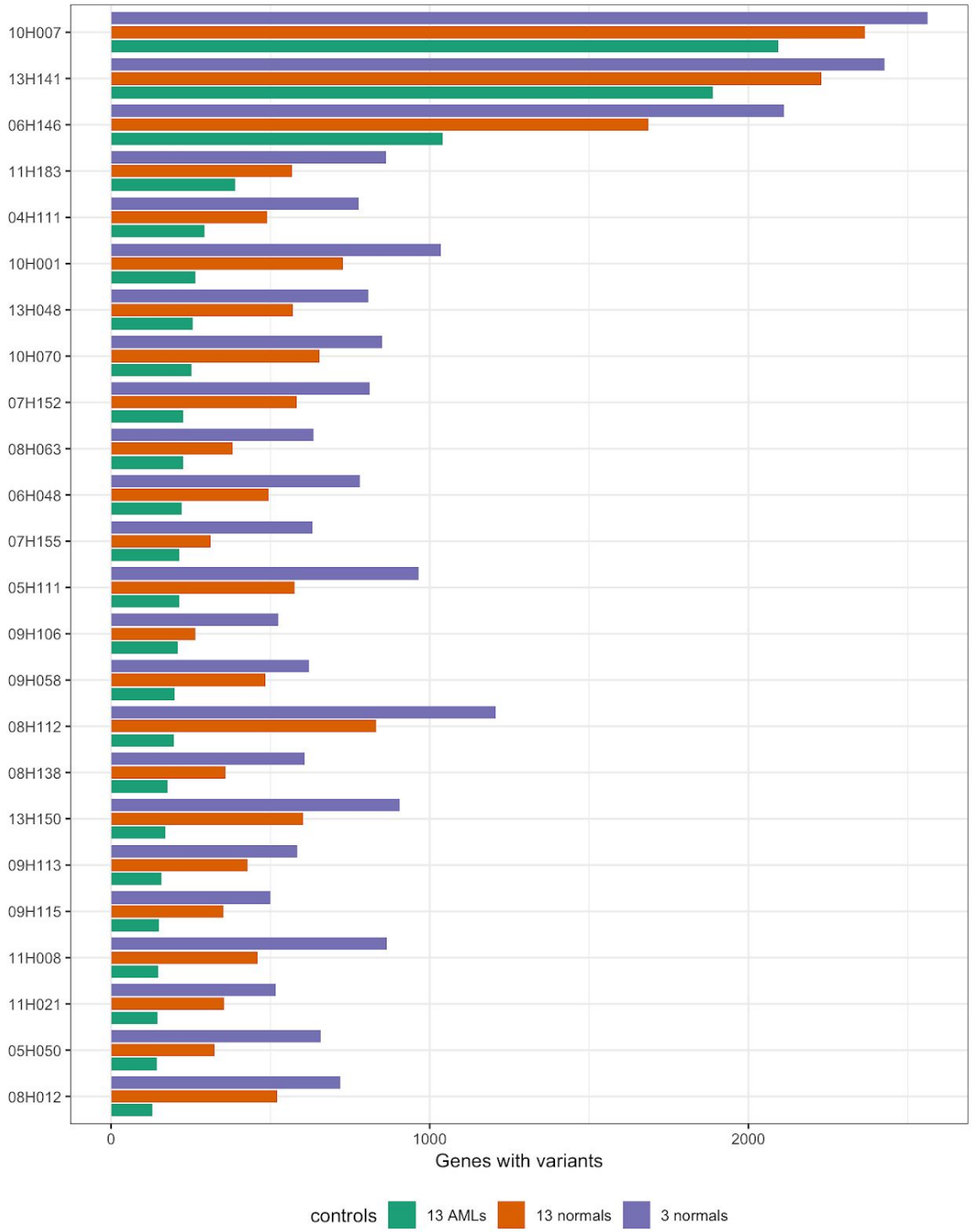
Supplementary Tables and Figures



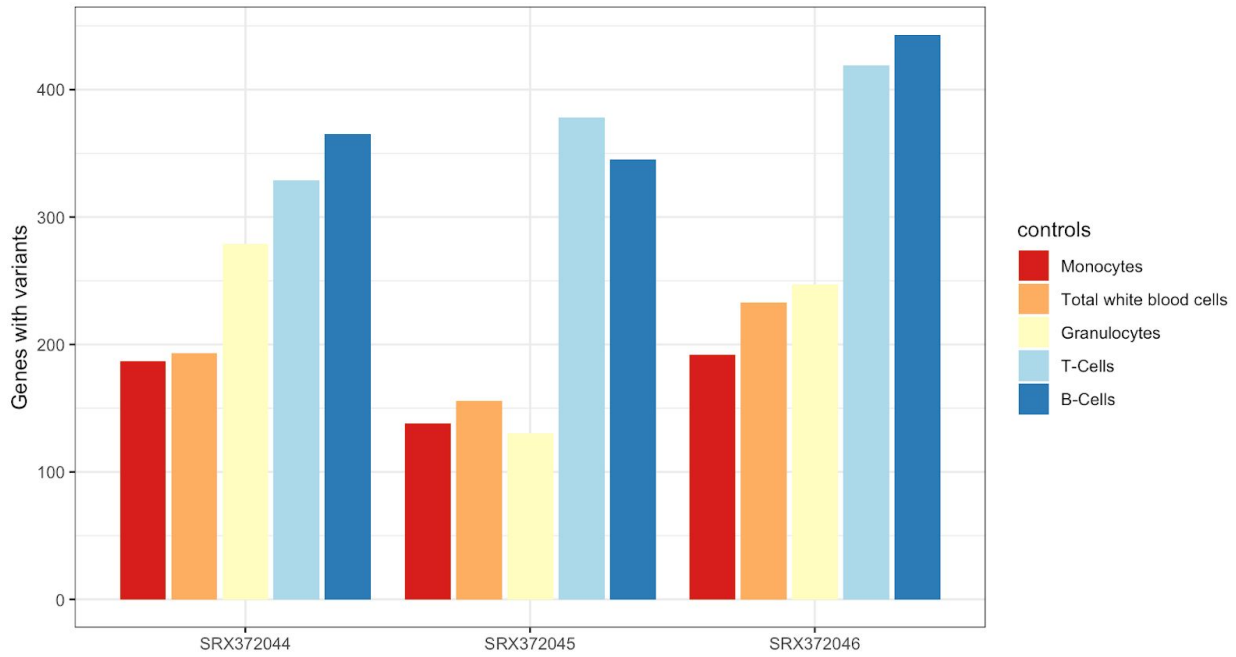
Supplementary Figure 1 | PCA plot of KMT2A-PTD cohort, compared with selected CBF AML controls, and leucegene normals (granulocytes, monocytes, total white blood cells, T-cells and B-cells), derived from Salmon quantification and the top 500 most variable genes. The reduced control set is circled.

Supplementary Table 1 **|** *Validated KMT2A-PTDs from a prior study[37,38] detected by MINTIE in a 24-sample leucegene cohort run against 3 leucegene normals (one each of monocytes, granulocytes and total white blood cells), 13 leucegene normals (5 monocytes, 5 granulocytes and 3 total white blood cells) and a randomly selected cohort of 13 leucegene CBF AMLs with no known KMT2A rearrangement. Coverage obtained from Audemard et al[38]. In cases where multiple PTDs were detected in the same patient, the highest min coverage was used.*

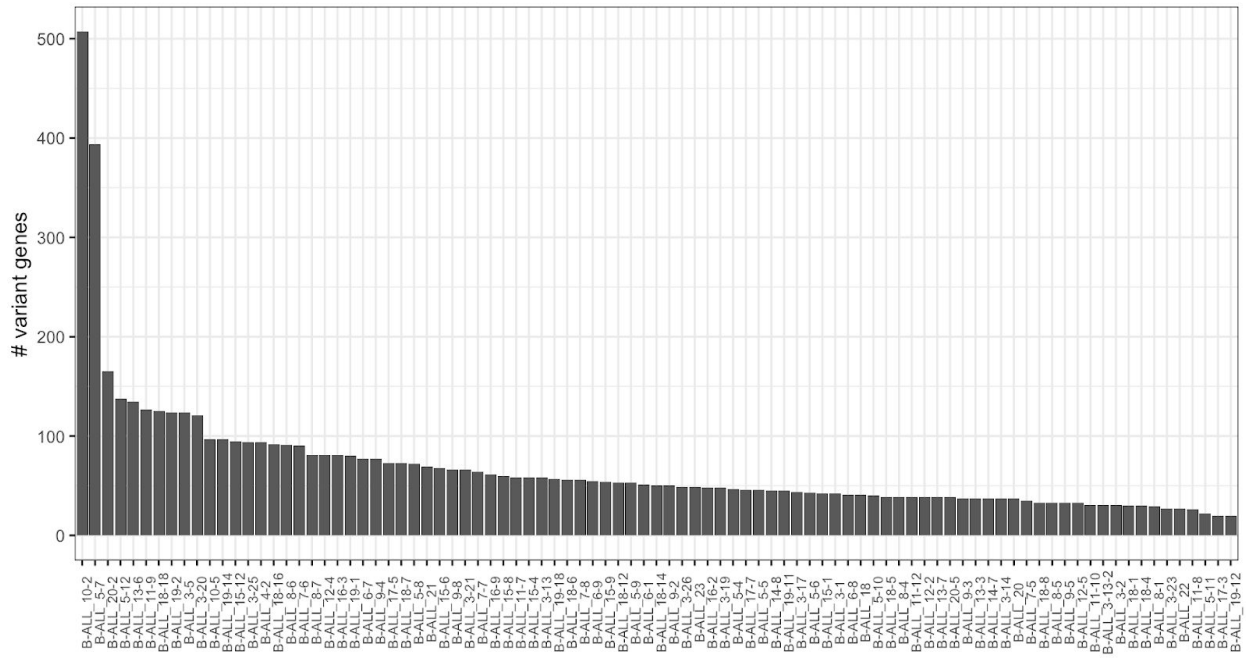
	Variant found using different control groups			
Patient	**3 normals**	**13 normals**	**13 AMLs**	**Coverage**
07H152	Y	Y	Y	158
09H115	Y	Y	Y	125
06H146	Y	Y	Y	87
05H111	Y	Y	Y	79
11H021	Y	Y	Y	63
05H050	Y	Y	Y	58
13H150	Y	Y	Y	58
13H048	Y	Y	Y	57
10H070	Y	Y	Y	53
10H007	Y	Y	Y	50
09H106	Y	Y	Y	49
13H141	Y	Y	Y	45
09H058	Y	Y	Y	29
07H155	Y	Y	Y	23
08H112	Y	Y	Y	22
09H113	Y	Y	Y	17
11H183	N	N	N	16
08H012	N	N	N	15
08H138	N	N	N	15
11H008	N	N	N	13
06H048	N	N	N	10
08H063	N	N	N	6
10H001	N	N	N	6
04H111	N	N	N	3
Found (of 24)	**11**	**16**	**15**	



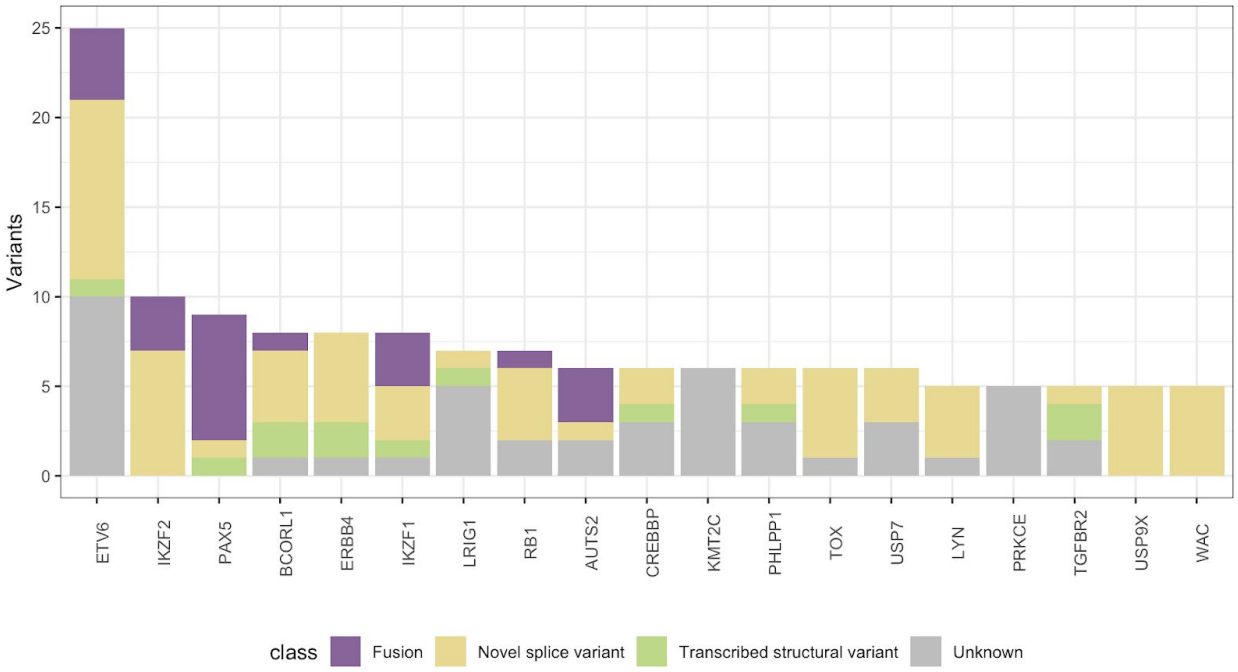
Supplementary Figure 2 | *Number of variant genes found in the KMT2A PTD cohort (24 leucegene samples containing KMT2A alterations) when using three different Leucegene control groups: a randomly selected set of 13 AMLs, 13 normals and 3 normals (subset of the 13).*



Supplementary Figure 3 | *Leucegene total white blood cell samples with different normal cell types as controls. Controls are ordered by total variants found across all three samples.*



Supplementary Figure 4 | *Variant genes found in RCH B-ALL cohort.*



Supplementary Figure 5 | *Variants found in ALL-associated genes in RCH B-ALL cohort.*

Supplementary Table 2 | *Novel variants found in the B-ALL cohort in clinically relevant genes. All locations are hg38.*

Gene	Variant type	Samples affected	Novel splice junction	Sequence queried by SeqOthello
RB1	Unpartnered fusion	3	chr13:48381444-48644164	TAGAACGATGTGAACATCGAATCATGGAATCCCTTGCATGGCTCTCAAGTCAGTTCCTGCCCCACTGCCCCACAGAAGTGTTTTCTGATGTGCT
			chr13:48381444-48647695	TAGAACGATGTGAACATCGAATCATGGAATCCCTTGCATGGCTCTCAAGCCTTGTATACACTCAATATGCAAGAAGCCCTGGAAGTTCCCAAGGT
			chr13:48381444-48530084	TAGAACGATGTGAACATCGAATCATGGAATCCCTTGCATGGCTCTCAAGACAGAGTTTTGCCATGTTGTCCCGGCTGGTCTCTAACTCCTGGGCT
IKZF1	PTD	1	chr7:50382708-50382540	CTGCCGCCGGAGGGACGCCCTCACTGGCCACCTGAGGACGCACTCCGGAGAACGGCCCTTCCAGTGCAATCAGTGCGGGGCCTCATTCACCCAGA
PAX5	PTD	1	chr9:37020805-37002647	ATCACGTCCCCCAGCGCCGACACCAACAAGCGCAAGAGAGACGAAGGACATGGAGGAGTGAATCAGCTTGGGGGGGTTTTTGTGAATGGACGGCC
ETV6	Novel splicing	3	chr12:11752580-11869424	AGCGCTCAGGATGGAGGAAGACTCGATCCGCCTGCCTGCGCACCTGCATAACTGTGTCCAGAGGACCCCCAGGCCATCCGTGGATAATGTGCACC
			chr12:11752580-11890941	AGCGCTCAGGATGGAGGAAGACTCGATCCGCCTGCCTGCGCACCTGCGTTTATGAAAACCCCAGATGAAATCATGAGTGGCCGAACAGACCGTCT
			chr12:11869970-11931173	TGTCTCCCCGCCTGAAGAGCACGCCATGCCCATTGGGAGAATAGCAGGTCCCATCCCATCCGAGTCTCAACAGAAACATCACCTCCCCAGGGAGA

Supplementary Table 3 | *List of Leucegene samples analysed.*

Patient ID	Cohort	Used as control
03H065	Core binding factor	
03H083	Core binding factor	
03H095	Core binding factor	Y
03H109	Core binding factor	
03H112	Core binding factor	
04H030	Core binding factor	Y
04H061	Core binding factor	Y
04H091	Core binding factor	Y
05H042	Core binding factor	Y
05H099	Core binding factor	Y
05H113	Core binding factor	
05H118	Core binding factor	Y
05H136	Core binding factor	Y
05H184	Core binding factor	Y
06H020	Core binding factor	
06H035	Core binding factor	Y
06H115	Core binding factor	Y
07H099	Core binding factor	Y
07H137	Core binding factor	Y
07H144	Core binding factor	
08H034	Core binding factor	
08H042	Core binding factor	
08H072	Core binding factor	

08H081	Core binding factor	
08H099	Core binding factor	
09H016	Core binding factor	
09H040	Core binding factor	
09H066	Core binding factor	
10H008	Core binding factor	
10H030	Core binding factor	
10H119	Core binding factor	
11H022	Core binding factor	
11H104	Core binding factor	
11H107	Core binding factor	
11H179	Core binding factor	
12H042	Core binding factor	
12H044	Core binding factor	
12H045	Core binding factor	
12H098	Core binding factor	
12H165	Core binding factor	
12H166	Core binding factor	
12H180	Core binding factor	
12H183	Core binding factor	
13H066	Core binding factor	
13H120	Core binding factor	
13H169	Core binding factor	
03H041	NUP98-NSD1	
05H034	NUP98-NSD2	

05H163	NUP98-NSD3	
08H049	NUP98-NSD4	
10H038	NUP98-NSD5	
11H027	NUP98-NSD6	
11H160	NUP98-NSD7	
05H050	KMT2A-PTD	
09H113	KMT2A-PTD	
09H115	KMT2A-PTD	
11H021	KMT2A-PTD	
08H012	KMT2A-PTD	
08H112	KMT2A-PTD	
11H008	KMT2A-PTD	
05H111	KMT2A-PTD	
06H146	KMT2A-PTD	
04H111	KMT2A-PTD	
06H048	KMT2A-PTD	
07H152	KMT2A-PTD	
07H155	KMT2A-PTD	
08H063	KMT2A-PTD	
08H138	KMT2A-PTD	
09H058	KMT2A-PTD	
09H106	KMT2A-PTD	
10H001	KMT2A-PTD	
10H007	KMT2A-PTD	
10H070	KMT2A-PTD	
11H183	KMT2A-PTD	
13H048	KMT2A-PTD	
13H141	KMT2A-PTD	
13H150	KMT2A-PTD	

Supplementary Note 1 | *Samples used as controls in the RCH B-ALL analysis.*

B-ALL3-1, B-ALL3-2, B-ALL5-8, B-ALL5-10, B-ALL5-11, B-ALL7-7, B-ALL8-1, B-ALL9-2, B-ALL9-5, B-ALL11-8, B-ALL12-4, B-ALL14-7, B-ALL14-8, B-ALL16-2, B-ALL16-3, B-ALL17-3, B-ALL18-14, B-ALL9-8, B-ALL10-2, B-ALL6-9, B-ALL11-10, B-ALL15-9, B-ALL19-11, B-ALL6-1, B-ALL18-4, B-ALL13-7, B-ALL19-12, B-ALL6-8, B-ALL3-26, B-ALL-22, B-ALL16-9, B-ALL3-20, B-ALL3-25, B-ALL18-7