Targeted variant detection in leukemia using unaligned RNA-Seq reads

Eric Olivier Audemard[1], Patrick Gendron[1], Vincent-Philippe Lavallée[1,2], Josée Hébert[1,2,4,5], Guy Sauvageau[1,2,4,5], Sébastien Lemieux[1,3*]

[1]The Leucegene project at Institute for Research in Immunology and Cancer, Université de Montréal, Montréal, Québec, Canada.

[2]Division of Hematology, Maisonneuve-Rosemont Hospital, Montréal, Canada

[3]Department of Computer Science and Operations Research, Université de Montréal, Montréal, Canada

[4]Quebec Leukemia Cell Bank, Maisonneuve-Rosemont Hospital, Montréal, Canada

[5]Department of Medicine, Faculty of Medicine, Université de Montréal, Montréal, Canada

Abstract

Mutations identified in each Acute Myeloid Leukemia (AML) patients are useful for prognosis and to select targeted therapies. Detection of such mutations by the analysis of Next-Generation Sequencing (NGS) data requires a computationally intensive read mapping step and application of several variant calling methods. Targeted mutation identification drastically shifts the usual tradeoff between accuracy and performance by concentrating all computations over a small portion of sequence space. Here, we present *km*, an efficient approach leveraging k-mer decomposition of reads to identify targeted mutations. Our approach is versatile, as it can detect single-base mutations, several types of insertions and deletions, as well as fusions. We used two independent AML cohorts (The Cancer Genome Atlas and Leucegene), to show that mutation detection by *km* is fast, accurate and mainly limited by sequencing depth. Therefore, *km* allows to establish fast diagnostics from NGS data, and could be suitable for clinical applications.

Introduction

Extensive molecular characterization of patient samples forms the basis of precision medicine and has proven useful in a number of pathologies, including Acute Myeloid Leukemia (AML)[1]. The use of RNA-Seq for sample characterization is appealing because (i) it is unbiased, as library preparation does not require capture or amplification, (ii) it focuses on a very small portion of the genome (~2%), namely the transcribed one, and (iii) mutations in regulatory regions are expected to leave a trace on the transcriptome, such as altering transcript expression. Unfortunately, the diversity of splicing events as well as incomplete transcriptome annotations make the mapping of RNA-Seq reads to the reference genome a notoriously difficult and computationally intensive task. This mapping step leads to frequent mapping errors, which are all inherited by variant calling methods.

In both clinical and research settings, analysis is often focused on a very limited set of expected variants (e.g., specific genes, positions, fusions or splice variants) for which some predictive value has been established. For instance, the FLT3 internal tandem duplication (FLT3-ITD) represents such a case for AML, since its presence is associated with poor prognosis[2, 3]. Therefore, clinicians routinely assess the presence of FLT3-ITD in patients using PCR-based tests[4, 5] on a targeted region. Alternatively, attempts to detect FLT3-ITD in Next-Gen Sequencing (NGS) data have been hampered by the difficulties of read mappers to properly align reads overlapping repeat junctions. To circumvent this problem, recent tools such as ITDseek[6] and ITDassembler[7], detect ITD from soft-clipped reads and require BAM files specifically generated by BWA-MEM[8], which does not perform an end-to-end alignment. Similarly, the best tools highlighted by Liu *et al.*[9] to detect gene fusions, combine multiple read alignment methods to achieve higher accuracy of both read alignment and fusion breakpoint detection. These methods spend considerable amount of resources mapping reads outside of targeted areas. They might also lose reads due to incorrect mappings. This is especially damaging if the incorrect mapping is due to the presence of a variant. This situation is most acute for variants resulting from sequence rearrangement (e.g., duplications, inversions, fusions, etc.), prompting the development of specialized software tools to detect them from NGS data.

Here we present *km* (https://github.com/iric-soft/km), a general method for targeted variant detection, which bypasses mapping to a reference genome. Instead, *km* looks for evidence of mutations using a compact digest of unaligned reads, similarly to LAVA[10] but without being limited to Single Nucleotide Variants (SNV). We demonstrate the usefulness of our method by detecting a set of several types of variants, which contributes to the identification of AML prognostic subgroups, on two large RNA-Seq cohorts and compare our *in-silico* results to the experimental results of two clinical laboratories. The Leucegene cohort comprises 437 deep transcriptomes (average number of reads: 204 x 10^6) from AML patient samples. The second cohort, consisting of 10,407 samples from 33 cancer types (including 151 AML) from TCGA, is analyzed mostly as a technical demonstration of *km*'s efficiency.

Result

Overview of *km*

Taking advantage of its targeted nature, *km* performs an extensive analysis of a single user-defined sequence, called the *target sequence*, which contains the region of interest. This *target sequence* is broken down into k-mers (subsequences of size k) overlapping by k-1 base pair (bp) to produce a linear directed graph. In this graph, each vertex represents a k-mer and each edge represents two overlapping k-mers (Figure 1(**a**)). Independently, a k-mer count table is prepared, that reports the occurrence of each k-mer in the sequenced sample[11]. A sequenced variant will then appear as an alternative path, connecting the starting and ending k-mers, identified by walking along the linear directed graph and following new overlapping k-mers queried from the count table. This process returns, as a graph, an approximate local assembly, which is then used to identify the presence of simple (e.g., single-base mutation) and more complex sequence variants, e.g., indel and other insertions or deletions (Figure 1(**b**)). In the absence of an alternative path, the target sequence has no variants in sequenced reads.

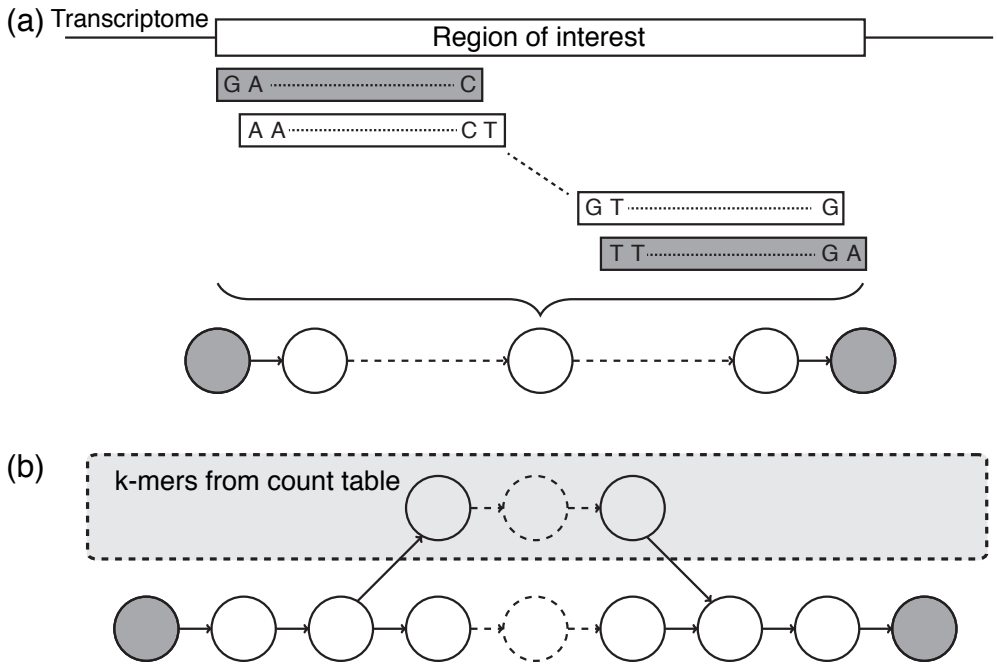

Figure 1: Overview of *km*. The input is a sequence from a region of interest, given by the user. (**a**) This region is segmented in k-mers to create a linear graph, which represent the research space delimited by the first and last k-mers (in grey). (**b**) A variant will be represented by a new path between the two extremities nodes. This path will be found by walking along the linear directed graph and following new overlapping k-mers, queried from the count table.

Analysis time

Table 1 reports our current set of target sequences prepared to annotate some variants of interest to the AML community. This set was designed to represent several types of mutations and illustrate strategies used by *km* to detect them. It includes SNVs in IDH1[12], DNMT3A[13] and MYC[14], insertions in NPM1[15], Internal Tandem Duplication (ITD)[16] and mutations in the Tyrosine Kinase Domain (TKD) of FLT3[17], Partial Tandem Duplication (PTD) in KMT2A and a fusion between NUP98 and NSD1[14, 16]. Moreover, these variants have already been validated, clinically or with other *in-silico* methods, in the two cohorts[14, 18, 19]. To our knowledge, all these variants are specific to AML, excluding the SNV on IDH1 also use in prognostic for Glioma tumour[20].

Gene	Expected type of variant	Gene location	Target sequence length (bp)	Average running time (s/sample)
IDH1	SNP (R132)	Exon 4	65	0.008
DNMT3A	SNP (R882)	Exon 23	65	0.010
NPM1	4bp insertion	Exon 10-11 + UTR	80	0.017
FLT3	ITD	Exon 13-15	345	0.107
FLT3	TKD	Exon 20	68	0.019
MYC	SNP (T58A/P59R)	Exon 2	68	0.018
NUP98-NSD1	Fusion	Exon 11 + 7	62	0.070
NSD1-NUP98	Fusion	Exon 6 + 13	62	0.061
KMT2A	PTD	Exon 8 + 2	62	0.067

Table 1: Current catalog of targeted mutations for AML. Average computation times are reported for Leucegene samples and assume that k-mer count tables are cached in RAM before running *km*. The performance of the caching step is highly dependent on I/O architecture, taking around 25 seconds on a typical system. The approaches used to prepare each *target sequence* for detecting the expected mutations are presented in Methods.

Applying *km* to the 437 samples from the Leucegene cohort, for all *targets sequences* in Table 1, required 3 hours and 11 minutes of CPU time (an average of 26 seconds per sample) considering precomputed count tables. Starting from reads, the Leucegene cohort can thus be annotated for 7 variants under 4 days, on a single workstation (i7-6700K@4.00GHz, using 4 threads, 8 GB of RAM and 31 bp-long k-mers). In contrast, aligning reads for all samples using STAR[21], with the same computational resources and 40 GB of RAM, requires approximately 19 days and is a prerequisite to run all alignment-based variant detection algorithms. In addition, the required storage space is about 4 times smaller for the k-mer count tables compared to aligned reads in BAM format, with an average file size of 3.2 GB per sample (see Supp. Fig. 1).

Cohort annotation

Table 2 presents the outcome of *km* on the catalog of *target sequences* for the Leucegene and TCGA cohorts. Columns “km type” indicate the specific *km* annotations returned for each variant. In

particular, *km* reports insertions using 4 subtypes based on the composition of the inserted sequence. When the insertion is adjacent to a deletion, *km* flags the variant as an Indel. When the inserted sequence is identical to a part of the *target sequence*, *km* returns an ITD (Internal Tandem Duplication). An I&I (Insertion and ITD) is returned when the inserted sequence is not identical but has more than 50% identity. If none of these conditions describes the inserted sequence, *km* returns an Insertion. The TCGA cohort is divided in two subgroups, 151 AML and 10,256 non-AML samples, to show the specificity of each variant found. For more details, output from *km* is available in the Supplementary Material.

To demonstrate the versatility of *km*, we included two types of *target sequences* in our current analysis. The first type of *target sequence* is extracted directly from the reference genome or transcriptome and, in that case, variants are identified as alternative paths. With this approach, *km* can find either known and specific variants (e.g. R882 SNV in DNMT3A) or any variant present within the targeted region (e.g. FLT3-TKD). A second type of *target sequence* is used for the last three targeted variants of Table 1 and 2. Here, *target sequences* are designed to represent the expected mutation, such as the concatenation of exons from two genes involved in an expected fusion (see, Detection of rearrangements). Care must be taken in interpreting *km*'s result when using this type of *target sequence*. In this case, detection of the *target sequence* confirms the presence of the expected mutation and a "variant" (of the *target sequence*) needs to be interpreted as an alternative to the expected mutation. Examples for this type of interpretation are found below for NUP98-NSD1 and KMT2A-PTD.

Data set	Mutation name	Km type						Variant	Target	Number of samples
		Ins	Del	Indel	Sub	ITD	I&I			
Leucegene	IDH1 R132	0	0	0	32	0	0	**32**	437	437
	DNMT3A R882	0	0	0	64	0	0	**64**	436	
	NPM1 4bp ins	22	0	1	0	103	13	**139**	437	
	FLT3-ITD	10	3	3	38	83	54	**162**	429	
	FLT3-TKD	0	4	0	31	0	0	**34**	434	
	MYC T58A/P59R	0	0	0	2	0	0	**2**	437	
	NUP98-NSD1	7 *	0	0	0	0	0	**7**	**6**	
	NSD1-NUP98	0	0	0	0	0	0	**0**	**2**	
	KMT2A-PTD	10 **	0	0	0	0	0	**10**	**15**	
TCGA (AML)	IDH1 R132	0	0	0	11	0	0	**11**	148	151
	DNMT3A R882	0	0	0	12	0	0	**12**	149	
	NPM1 4bp ins	6	0	0	0	28	6	**40**	151	
	FLT3-ITD	76	0	5	22	20	18	**100**	142	
	FLT3-TKD	0	1	0	11	0	0	**12**	149	
	MYC T58A/P59R	0	0	0	2	0	0	**2**	139	
	NUP98-NSD1	0	0	0	0	0	0	**0**	**0**	
	NSD1-NUP98	0	0	0	0	0	0	**0**	**2**	
	KMT2A-PTD	3 **	0	0	0	0	0	**3**	**0**	
TCGA (non-AML)	IDH1 R132	0	0	0	394	0	0	**394**	9850	10256
	DNMT3A R882	0	0	0	0	0	0	**0**	9267	
	NPM1 4bp ins	0	0	0	0	0	0	**0**	10232	
	FLT3-ITD	0	0	0	9	0	0	**9**	163	
	FLT3-TKD	0	0	0	0	0	0	**0**	361	
	MYC T58A/P59R	0	0	0	5	0	0	**5**	9204	
	NUP98-NSD1	0	0	0	0	0	0	**0**	**0**	
	NSD1-NUP98	0	0	0	0	0	0	**0**	**0**	
	KMT2A-PTD	0	0	0	0	0	0	**0**	**0**	

* Fusion with exon 12 found as an insertion in the target sequence

** Tandem duplication extended with exon 9 or 9 and 10

Table 2: Variants identified by *km* using our AML catalog in the leucegene and TCGA. The "target" column reports the number of samples expressing the *target sequence*. The "variant" column shows the number of samples where at least one variant of the *target sequence* is found. As a *target sequence* can represent a reference or a mutated sequence, we have indicated in bold counts representing mutated samples. The columns "km type" identify the specific types of variants detected. Of note, several types of variant can be identified in a given sample. As expected, SNV on IDH1 are found in AML and non-AML samples on Lower Grade Glioma (LGG) (see Supp. Table 1).

Sensitivity and precision

We assessed sensitivity and precision of *km* based on the detection of insertions in NPM1 and FLT3 (see Supp. Table 2). These two variants have been experimentally validated by the Banque de Cellules Leucémique du Québec (BCLQ, www.bclq.org) and TCGA[18], independently from RNA-Seq. For FLT3-ITD, we also compare *km* with existing methods previously proposed to detect ITD (e.g. ITDassembler[7], Pindel[22] and Genomon ITDetector[23]).

On NPM1 variants, *km* successfully identified 117 mutated Leucegene samples of the 118 clinically validated cases by the BCLQ (on 202 tested). All mutated samples identified by *km* were either validated independently by the BCLQ or not part of the subset of samples tested clinically. This high level of performance is also observed with the AML subset of the TCGA cohort[18], where *km* identified all clinically validated mutated samples (36 clinically validated of 150 tested), more 4 mutated samples not detected in clinical settings. Nevertheless, all 4-bp insertions reported by *km* are also known variants from the COSMIC database (see Table 3).

Type	COSMIC ID	Km variant	Km type	Leucegne	TCGA AML
A	COSM17559	TCTG	ITD	103	28
D	COSM17573	CCTG	I&I	11	4
B	COSM17571	CATG	Insertion	10	4
	COSM20809	CCAG	Insertion	3	0
	COSM29814	CAGA	Insertion	2	0
	COSM3356078	CAAG	Insertion	1	0
	COSM29814	CCGA	Insertion	1	0
	COSM3356078	CGCG	Insertion	1	1
	COSM28066	CGGA	Insertion	1	0
	COSM20811	CTTG	Insertion	1	0
	COSM20815	TATG	I&I	1	2
	COSM20813	TCGG	I&I	1	0
	COSM27390	TTCG	Insertion	1	0
	COSM20850	AGAA	Insertion	1	0
	COSM20810	TTGT	Insertion	0	1
	Unknown	gc/CAGGG	Indel	1	0
		Total mutated / Total		138/437	40/151

Table 3: NPM1 mutations identified by *km* in both Leucegene and TCGA cohorts.

Similar performance is observed for FLT3-ITD, where *km* identifies all variants clinically validated from the Leucegene cohort and 31 out of 33 cases for the TCGA AMLs. Upon further investigation of *km*'s output for these missing samples (IDs TCGA-AB-2812 and TCGA-AB-2988), we found that the expected ITD sequences are detected but that the sequences were filtered of the final output, due to part of the *target sequence* having zero coverage (before or after the diverging path, see Methods). We opted to filter out these sequences, as they can arise from the expression of a region made up of k-mers derived from a different transcript but in common with the *target sequence*. However, this is not the case for these two samples where identification of the ITDs is missed due to an homozygous SNV after the ITD (t/C at chr13:28,034,322). Unfortunately, *km* is designed to independently report single mutation events for a *target sequence* in a given sample. Here, the homozygous SNV is detected by *km* and can be flagged as a potential false negative or used to modify the *target sequence* to find the FLT3-ITD in a second *km* analysis. Other strategies are currently being explored to account for the possibility of concurrent mutations for a given target without compromising speed, sensitivity and precision.

Until now, methods specialized to detect ITD were designed for exome sequencing, such as ITDassembler, Pindel and Genomon ITDetector. Consequently, comparison with *km* was performed on a reduced TCGA AML cohort of 28 samples shared by all studies (33 samples reported by ITDassembler [7], less 5 for which no RNA-seq data was available). Of these 28 samples, 22 had clinically validated FLT3-ITD annotations[18]. On this reduced cohort, *km* has a sensitivity of 95% (21 out of 22), while ITDassembler identified 15 samples, Genomon 14 (1 had an incorrect duplication length) and Pindel 11 (5 had incorrect duplication lengths). As mentioned by ITDassembler's author [7], 5 out of the 7 missed mutated samples can be attributed to the length of the duplication, which approaches or exceeds the read length (>60 bp in a 75-bp read). By contrast, the nature of *km*'s algorithm does not limit the duplication size to the read length but instead relies on a user-defined parameter that limits the maximum length of alternative branches to explore (default 500 bp). All details of ITDs found by each software can be found in the Supp. Table 3.

Sensitivity and coverage in RNA-Seq

RNA-Seq specific features need to be accounted for to correctly interpret *km*'s results. First, is the presence of non-spliced transcripts, resulting in reads corresponding to intron sequences being reported by *km* as insertions in the *target sequence*. These sequences are easily detected by their specific genomic positions, lengths and nucleotide sequences across different samples. We encountered this case during the FLT3-ITD analysis, where the *target sequence* overlaps exons 13 to 15 in order to cover all ITD locations (see section "Target sequence" in Methods). Consequently, we have removed from *km*'s output 86-bp and 90-bp insertions, respectively found at locations chr13:28,034,083 and chr13:28,034,304, corresponding to known introns (identified in 76 out of 151 TCGA AML and 10 out of 437 Leucegene samples).

A related subtlety arises when the variant sequence is largely underrepresented compared to the *target sequence*. This can be due to the variant being present in a rare sub-clone or to strongly biased allele-specific expression. In this case, we can lower the coverage ratio threshold (default $p = 0.05$) to avoid ignoring these variants. The missing Leucegene NPM1 mutated sample (13H065) is such a case, where *km* was able to detect a tg/GTCCGA Indel using higher sensitivity thresholds ($p = 0.01$). This variant has a local coverage ratio of 4% in this sample, just below the default threshold of 5%. Similar cases are identified for DNMT3A, where our *km* analysis of the TCGA cohort[18] missed the identification of 7 clinically annotated samples (out of 18 with available RNA-Seq). By adjusting *km*'s parameters for higher sensitivity ($p = 0.01$ and $c = 2$), 3 of those samples are correctly identified.

Detection of rearrangements

Detection of rearrangements bringing together two distant regions of the genome (such as translocations, large inversions, duplications or deletions) requires a slightly different approach to the design of the *target sequence*. In the case of translocations (such as NUP98-NSD1), neither genomic nor transcript sequences can be used as a source for the *target sequence*. Instead we created an artificial sequence that represents an expected transcript of the fusion. This artificial *target sequence* is then used to verify that each k-mer is covered in the sample's count table. With this approach,

support for the *target sequence* is interpreted as a confirmation that this fusion is present in the sample. For large duplications or deletions, the same strategy is applied to avoid capturing several mutations (e.g. common single-base mutations) by using excessively large *target sequences*.

We applied this approach to the detection of the NUP98-NSD1 fusion and report the results in Table 2. We identified 6 samples in the Leucegene cohort that express a fusion transcript joining exon 11 of NUP98 to exon 7 of NSD1[14]. An alternative inserted sequence, that matches exactly exon 12 of NUP98, is also identified for 7 samples (including the previous 6). These results suggest alternative splicing events occurring in the fusion transcript. In the TCGA AML cohort, clinical analyses[18] reports an NSD1-NUP98 fusion on samples TCGA-AB-2930 and TCGA-AB-2856 and a NUP98-NSD1 fusion on TCGA-AB-2930. By adding a second *target sequence* to identify the fusion on the reverse strand, we identified all NSD1-NUP98 fusions but not the one on NUP98-NSD1. This can be explained by a fusion transcript undetectable with this *target sequence* (a fusion involving at least one exon at extremities of *target sequence*, e.g. exon 10 and 8 of NUP98 and NSD1, respectively) or a sequencing coverage too low to catch this variant.

The KMT2A-PTD is a tandem duplication that spans from exon 2 to exon 8 or 10 of KMT2A. Here, applying the strategy used for FLT3-ITD would require a *target sequence* of about 3,712 bp. Instead, we designed our *target sequence* by joining exon 8 directly to exon 2 to capture the specific junctions created by the duplication. Using this sequence, *km* identifies all versions of this duplication by adding an insertion when more exons are included in the duplicated region (Table 2). Surprisingly, further investigation of *km*'s output revealed that among the 24 samples identified as having a KMT2A-PTD in the Leucegene cohort[19], 6 samples showed evidence for more than one transcript. On the TCGA AML cohort, no report of this mutation has been found in cBioPortal or GDC (which combines clinical[18] and computational analyses) but *km* identifies 3 samples with a tandem duplication that includes up to exon 10.

Discussion

To our knowledge, *km* is the first simple and unified method to detect variants arising from various sequence rearrangements (e.g., substitutions, duplications, inversions and fusions). Other methods are designed for the detection of a specific type of rearrangement[6, 7] and report all cases by exploiting characteristics discriminant for the selected type of variant. Alternatively, *km* uses any sequence as an expected reference and reports all alternative forms sharing the reference's extremities. This approach is strictly restricted to the region analyzed but not to the type of variant to be identified. Moreover, this strategy improves the sensitivity for detecting challenging cases such as the FLT3-ITD where position and length are highly variable.

The small amount of computational resources and time required by *km* to process and detect variants in 10,844 samples, is unparalleled. Indeed, since the advent of whole genome and transcriptome sequencing, the default approach has been to perform analyses globally, resulting in very resource-intensive processes. Instead, *km* performs a deeper analysis on focused parts of the genome, preselected based on each user's interest (e.g. valuable for prognosis or targeted therapy). Thereby, *km* avoids consuming resources on *a priori* irrelevant parts of the genome and returns no results outside selected regions. Nonetheless, *km* analyses are incremental and allow to test new regions and enrich previous annotations for new variants. Unsurprisingly, our approach is sensitive to a lack of coverage over the target region, either from allelic imbalance or cellular heterogeneity. In these situations, the design of the RNA sequencing (e.g., depth, number of cells) needs to be adjusted for increased sensitivity.

Finally, this study shows *km*'s potential to establish fast and detailed diagnostics for any given patient, using only one sequencing, further extending the usefulness of these techniques in clinical settings. Indeed, all variants used to evaluate *km* have been shown to have prognostic value[13-17] for Acute Myeloid Leukemia. As shown here, identifying these variants in 437 samples can successfully be done in four days with *km*, using a single workstation. Moreover, as sequencing becomes cheaper and more available, we envision a transition from targeted to whole sequencing, followed by targeted analysis using methods like *km*. This will provide fast clinical prognostics but also gather the

complete data related to a patient's disease for complementary analyses, to uncover other mutations during treatment, as well as for large-scale research. In the future, selecting and testing new target variants to extend the existing catalog would be an interesting step to extend *km*'s clinical (and research) applications to other diseases.

Methods

RNA-Seq sequencing of AML samples

This study is part of the Leucegene project, an initiative approved by the Research Ethics Boards of Université de Montréal and Maisonneuve-Rosemont Hospital. All AML samples were collected with informed consent between 2001 and 2015 according to Quebec Leukemia Cell Bank procedures. RNA-Seq were deposited in the Gene Expression Omnibus (GSE49642, GSE52656, GSE62190, and GSE67040) and exome sequencing in the Short Read Archive (BioProject PRJNA358716). Workflow for sequencing, mutation analysis and transcript quantification has been described previously[19]. Briefly, libraries were prepared with TruSeq RNA Sample Preparation kits (Illumina) and sequencing was performed using an Illumina HiSeq 2000 with 200 cycle paired-end reads. The complete dataset consists of 437 RNA-Seq (204M reads per sample, on average).

TCGA data was obtained from the GDC portal for all 33 cancers types and preprocessed using Picard to obtain raw sequences from the available aligned files, which were then converted to Jellyfish count tables. The complete dataset represents 10,407 RNA-Seq samples of which 151 correspond to AML samples (TCGA-AB-2975 sample was not in the initial TCGA AML cohort[18]). TCGA results shown here are in whole or part based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/.

K-mer count tables with Jellyfish

We used Jellyfish (v2.1.4)[11] to create canonical k-mer count tables of length 31 bp, and opted not to store k-mers that were seen in a single read (parameters: -m 31 -C -L 2). These occurrences either result from sequencing errors or from very low coverage regions, and in both cases, cannot be leveraged to confirm the presence of a mutation. We also filtered k-mers based on sequencing quality

(default: >'+'). In our approach, sequencing errors or low base quality results in the k overlapping k-mers to be ignored during k-mer walking.

km: k-mer walking as a local assembly

The main idea behind the implementation of a k-mer walking algorithm is to identify, in the raw sequencing reads, variant sequences that overlap with the two extremities of a *target sequence* designed to interrogate a region of interest. This *target sequence* can be broken down into k-mers overlapping by *k*-1 bp to produce a linear directed graph where vertices represent k-mers and edges represent overlaps. Using this graph and a k-mer count table, diverging paths can be identified that connect the starting k-mer to the ending one. In the absence of an alternative path, the *target sequence* doesn't have a variant. Consequently, the *target sequence* needs to flank the variant by at least *k* bp to have this common starting and ending k-mer. And *k* must be large enough to linearly decompose the *target sequence*.

The k-mer walking algorithm is implemented as a depth-first search. To limit search space, we ignore branches that do not come back to a k-mer derived from the *target sequence* within *s* steps (default: 500). Also, new branches are explored only if each k-mer has a count greater than a fixed threshold *c* (default: 5) and greater than a fraction *p* (default: 5%) of the alternative k-mer count. These parameters may need to be adjusted to support detection of certain variants (e.g. as we did to find the indel on 13H065, see section "Sensitivity and coverage in RNA-seq") but we have found that the default values perform well in general.

To extract variant sequences from the directed graph, we assign small weights (i.e. 0.01) to edges that are part of the *target sequence* while the others are assigned a unit weight. This ensures the preferential use of paths from the reference sequence, and variants are enumerated by finding the set of unique shortest paths covering all edges of the graph.

The k-mer count table is the core data structure used by our approach and its content, construction time and necessary storage represent key operational characteristics in using *km*. Once the k-mer count tables are prepared, the process of identifying variants is typically IO-bound until the queried

count table is entirely cached in RAM. From an operational perspective, it becomes important to complete the analysis of all *target sequences* for each sample, before moving to the next sample. Alternatively, we have obtained great performance by copying count tables to a RAM disk (or by ensuring that it is cached using a simple scanning command such as "wc -l") before launching a series of targeted analyses. The speedup observed for using *km* is largely dependent on the number of regions to be investigated but in absolute terms, preparation of the k-mer count table requires around 3 seconds per million reads of sequences and 0.1 seconds per samples for each target with length < 300 bp (preloading of the jellyfish database to RAM requires around 25s).

Read dependency

An important caveat to identifying variants based on k-mer count tables is that we ignore read dependency between k-mers. During the k-mer walking, the overlap by k-1 bp between two k-mer is taken as an evidence that there exist at least one read that support the assembled sequence. In a De Bruijn graph[24], the read dependency is explicitly stored in the graph which explains the high memory requirements of this data structure. For *km*, the consequence of ignoring read dependency is that reducing *k* directly impacts the false positive rate by increasing the frequency of incorrect reconstruction, which grows with the frequency of ambiguous regions of length *k-1*. The constraint to begin and end variant paths on k-mers belonging to the *target sequence* acts as a very stringent filter to eliminate these spurious branches that would arise from low complexity regions of the genome. DiscoSNP[25] shares with *km* an internal step where variants are identified through k-mer based local assembly. To provide a safety net, the authors of DiscoSNP opted to include a further step during which they align reads to variant sequences to compute k-read-coherency. Besides being a sound precaution, this read coherency will always remain limited to the read length, thus only protecting against incorrect identifications induced by ambiguous sequences of lengths between *k* and the read length (respectively 31 and 100 bp for the results presented here).

It can be very tempting to increase *k* (up to the read length) with the intention of increasing accuracy, but two factors need to be taken into consideration. First, any sequencing error throws away the contribution of *k* k-mers. Thus, as *k* is increased, the effective depth contributing to the analysis is

decreased. Second, *km*'s algorithm cannot detect a variant that is located less than k bp away from the extremity of the *target sequence*, since the variant branch would not merge back to the last k-mer. Thus, as *k* is increased, longer *target sequences* must be used which can be problematic (see the case of NPM1, in section "Target sequences"). Although in a very different context (genome annotation from NGS), an optimal *k* was proposed to be between 19 and 20 to prepare k-mer count tables of the human genome. Using $k = 22$, a 10^{-4} probability of false location was also estimated[26]. These observations should be used as guiding principles in the rare cases where *k* needs to be adjusted. Based on our experience in analyzing the Leucegene and TCGA cohorts, we have found few cases that prompted us to raise *k* from 21 to 31 as the project progressed. We have been so far satisfied with results obtained at $k = 31$. In a general case, we have no estimate of the frequency with which a variant would be wrongly identified at $k = 31$ and correctly disproved using the full reads. This subject remains to be further explored.

Target sequences

Creation of target sequences by an end user is relatively simple. In general, *k* bp from each side of the targeted variant are required for exact mutation identification and larger regions may be used for more exploratory analyses. Our *target sequence* for the DNMT3A R882 mutation serves as a simple example in which 65 bp (31 + 3 + 31 = 65) represent the minimal *target sequence* flanking the R882 codon.

The FLT3-ITD and the 4-bp insertion in NPM1 present more challenging situations. For NPM1, the expected mutation appears close to the end of the last exon and we recommend extending the sequence as it is essential that the last k-mer of the *target sequence* does not overlap the targeted mutation (see section "k-mer walking as a local assembly"). Consequently, we prepared an 80-bp *target sequence* comprising the end of exon 10, exon 11 and 14 bp of the 3'-UTR. For the FLT3-ITD, the duplication can span over two exons (14 and 15) and can be longer than sequencing reads. To cover all insertion points and lengths, we prepared a 345-bp *target sequence* overlapping exons 13 to 15 of the FLT3 transcript. For detection of rearrangements, we need to include k bp on each side of the junction created by the targeted variant (see section "Detection of rearrangements").

To assess the fraction of the transcriptome that can be targeted using *km*, we computed the fraction of unique sequences identified for various values of k between 10 and 100. With $k = 31$, 96.76% of the human transcriptome is readily represented by a linear k-mer graph and thus accessible as a *target sequence* for *km* (see Supp. Fig. 2).

References

1. Prada-Arismendy, J., Arroyave, J.C. & Röthlisberger, S. Molecular biomarkers in acute myeloid leukemia. *Blood reviews* **31**, 63-76 (2017).
2. Fröhling, S. et al. Prognostic significance of activating FLT3 mutations in younger adults (16 to 60 years) with acute myeloid leukemia and normal cytogenetics: a study of the AML Study Group Ulm. *Blood* **100**, 4372-4380 (2002).
3. Dohner, H. et al. Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. *Blood* **129**, 424-447 (2017).
4. Gale, R.E. et al. The impact of FLT3 internal tandem duplication mutant level, number, size, and interaction with NPM1 mutations in a large cohort of young adult patients with acute myeloid leukemia. *Blood* **111**, 2776-2784 (2008).
5. Stone, R.M. et al. Midostaurin plus Chemotherapy for Acute Myeloid Leukemia with a FLT3 Mutation. *N Engl J Med* **377**, 454-464 (2017).
6. Au, C.H., Wa, A., Ho, D.N., Chan, T.L. & Ma, E.S.K. Clinical evaluation of panel testing by next-generation sequencing (NGS) for gene mutations in myeloid neoplasms. *Diagnostic Pathology* **11**, 11 (2016).
7. Rustagi, N. et al. ITD assembler: an algorithm for internal tandem duplication discovery from short-read sequencing data. *BMC bioinformatics* **17**, 188 (2016).
8. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv.org* **q-bio.GN** (2013).
9. Liu, S. et al. Comprehensive evaluation of fusion transcript detection algorithms and a meta-caller to combine top performing methods in paired-end RNA-seq data. *Nucleic acids research* **44**, e47-e47 (2016).
10. Shajii, A., Yorukoglu, D., William Yu, Y. & Berger, B. Fast genotyping of known SNPs through approximate k-mer matching. *Bioinformatics* **32**, i538-i544 (2016).
11. Marcais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764-770 (2011).
12. Medeiros, B.C. et al. Isocitrate dehydrogenase mutations in myeloid malignancies. *Leukemia* **31**, 272-281 (2017).
13. Yuan, X.-Q. et al. Evaluation of DNMT3A genetic polymorphisms as outcome predictors in AML patients. *Oncotarget* **7**, 60555-60574 (2016).
14. Lavallée, V.-P. et al. Identification of MYC mutations in acute myeloid leukemias with NUP98-NSD1 translocations. *Leukemia* **30**, 1621-1624 (2016).
15. Thiede, C. et al. Prevalence and prognostic impact of NPM1 mutations in 1485 adult patients with acute myeloid leukemia (AML). *Blood* **107**, 4011-4020 (2006).
16. Shiba, N. et al. High PRDM16 expression identifies a prognostic subgroup of pediatric acute myeloid leukaemia correlated to FLT3-ITD, KMT2A-PTD, and NUP98-NSD1: the results of the Japanese Paediatric Leukaemia/Lymphoma Study Group AML-05 trial. *British Journal of Haematology* **172**, 581-591 (2016).
17. Bacher, U., Haferlach, C., Kern, W., Haferlach, T. & Schnittger, S. Prognostic relevance of FLT3-TKD mutations in AML: the combination matters--an analysis of 3082 patients. *Blood* **111**, 2527-2537 (2008).
18. Network, C.G.A.R. et al. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *New England Journal of Medicine* **368**, 2059-2074 (2013).

19. Lavallée, V.-P. et al. The transcriptomic landscape and directed chemical interrogation of MLL-rearranged acute myeloid leukemias. *Nature Genetics* **47**, 1030-1037 (2015).
20. Yan, H. et al. IDH1 and IDH2 mutations in gliomas. *N Engl J Med* **360**, 765-773 (2009).
21. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
22. Ye, K., Schulz, M.H., Long, Q., Apweiler, R. & Ning, Z. 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 (2009).
23. Chiba, K. et al. Genomon ITDetector: a tool for somatic internal tandem duplication detection from cancer genome sequencing data. *Bioinformatics* **31**, 116-118 (2015).
24. Compeau, P.E.C., Pevzner, P.A. & Tesler, G. How to apply de Bruijn graphs to genome assembly. *Nature biotechnology* **29**, 987-991 (2011).
25. Uricaru, R. et al. Reference-free detection of isolated SNPs. *Nucleic acids research* **43**, e11-e11 (2015).
26. Philippe, N. et al. Using reads to annotate the genome: influence of length, background distribution, and sequence errors on prediction capacity. *Nucleic acids research* **37**, e104-e104 (2009).

Acknowledgments

The authors wish to thank Muriel Draoui for project coordination and Sophie Corneau for sample coordination, Isabel Boivin for data validation as well as Marianne Arteau and Raphaëlle Lambert at the IRIC genomics platform for RNA sequencing. The dedicated work of BCLQ staff namely Giovanni d'Angelo, Claude Rondeau and Sylvie Lavallée is also acknowledged. This work was mostly supported by Genome Canada and Genome Quebec with supplementary funds from Amorchem. Contribution from Ministère de l'économie, de l'innovation et des exportations du Québec and Leukemia Lymphoma Society of Canada to is acknowledged. G.S. and J.H. are recipients of research chairs from the Canada Research Chair program and Industrielle-Alliance (Université de Montréal). BCLQ is supported by grants from the Cancer Research Network of the Fonds de recherche du Québec–Santé. RNA-Seq read mapping and transcript quantification were performed on the supercomputer Briaree from Université de Montréal, managed by Calcul Québec and Compute Canada. The operation of this supercomputer is funded by the Canada Foundation for Innovation (CFI), NanoQuébec, RMGA and the Fonds de recherche du Québec - Nature et technologies (FRQ-NT). V.P.L is supported by a fellowship from the Cole Foundation and by a Vanier Canada Graduate Scholarship.

Author contributions

E.O.A co-developed km, prepared figures and tables, and co-wrote the manuscript.

P.G. processed the raw NGS data, edited the manuscript, and co-developed km.

V.P.L. contributed to the development of AML query sequences, validation of identified variants, edited the manuscript.

J.H. provided all the AML samples, and was responsible for experimental validations.

G.S. contributed to project conception and edited the manuscript.

S.L. co-developed km, co-wrote the manuscript, and was responsible for project conception and coordination.

Competing Financial Interests statement

The authors declare no competing financial interests.