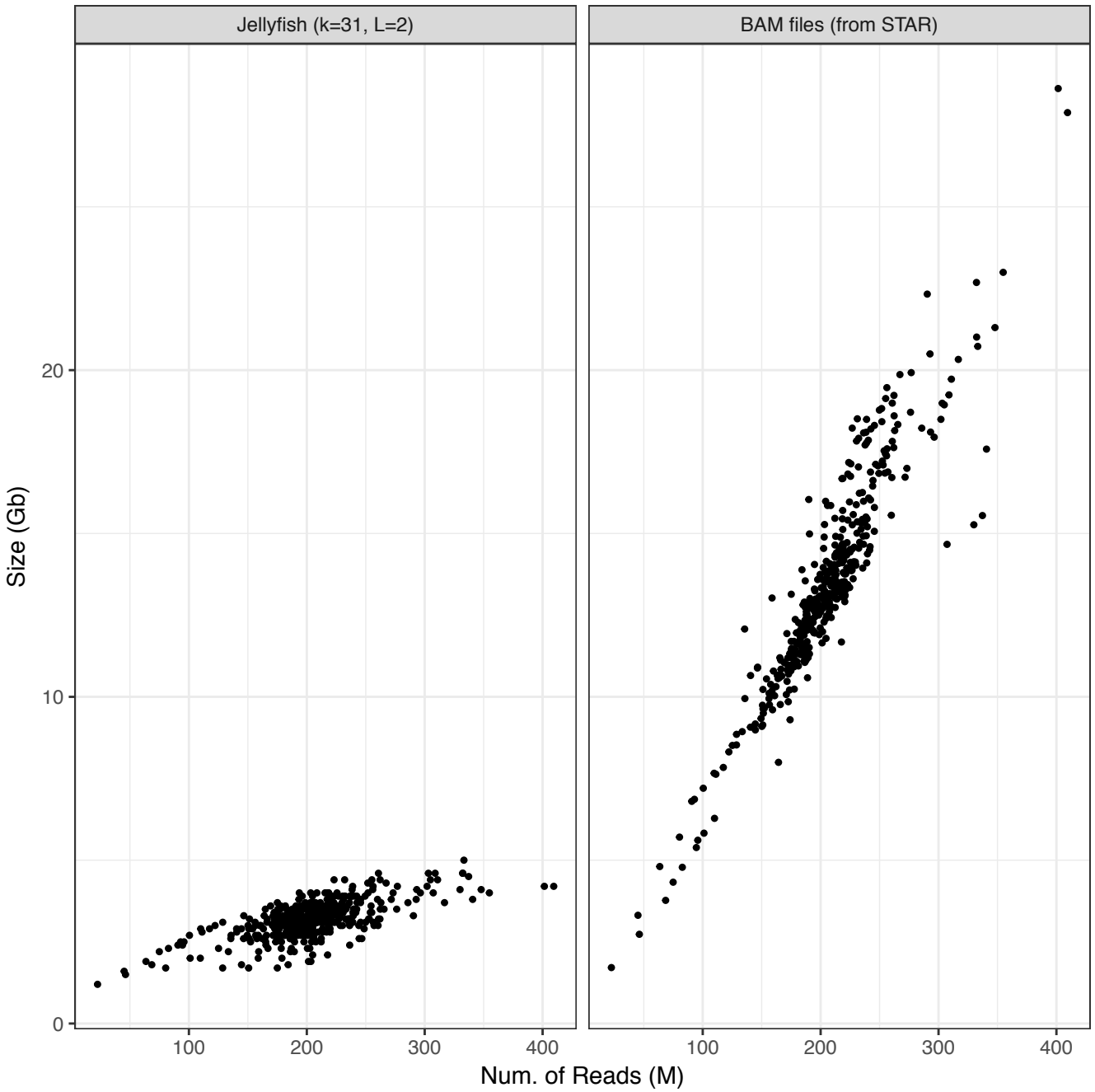
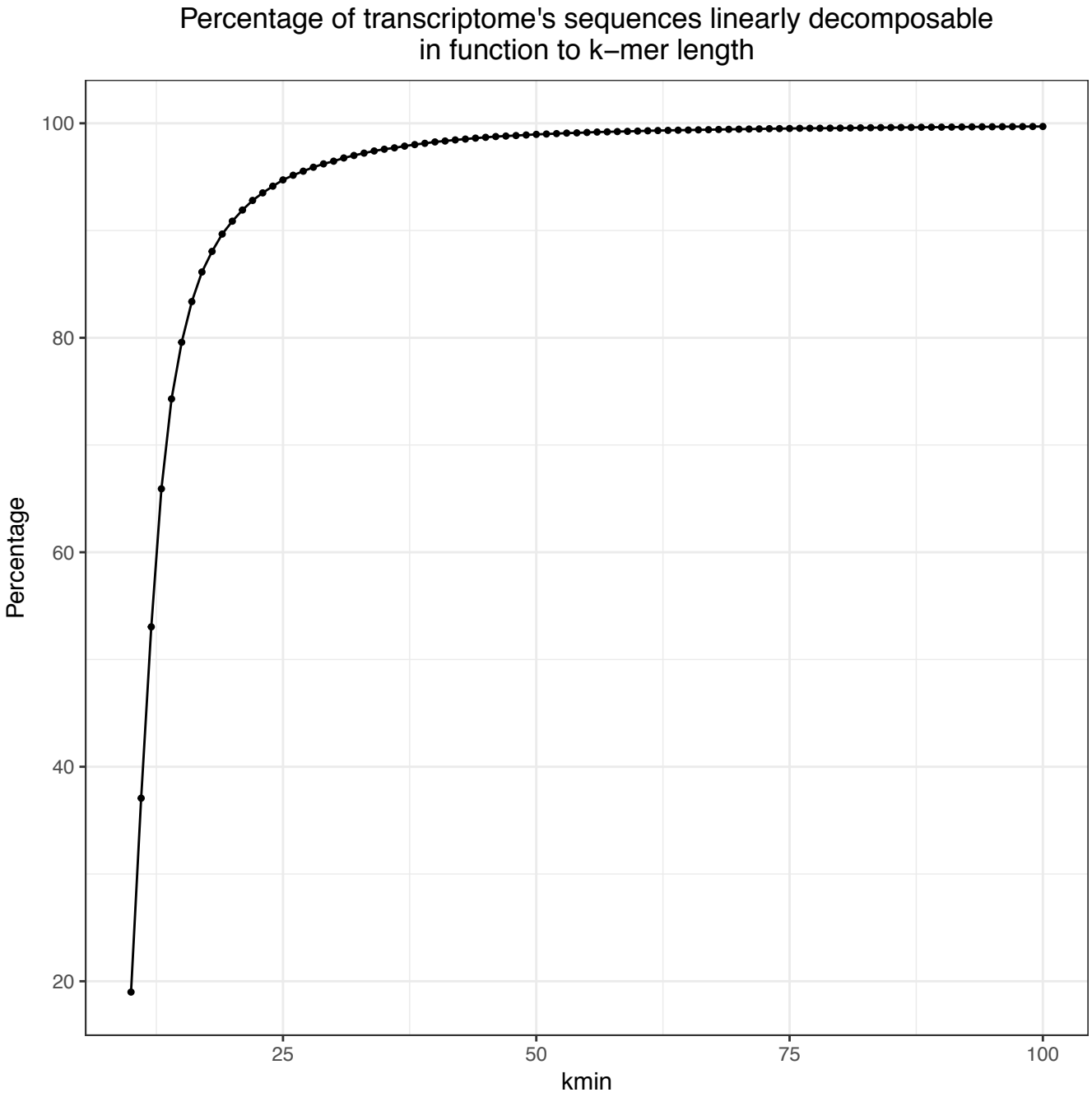


Supplementary figure 1: Comparison of k-mer count tables and BAM files size. K-mer count tables are created by Jellyfish, with k = 31 bp and minimum k-mer count of 2. BAM files are produced following a STAR alignment.



Supplementary figure 2: Percentage of transcriptome sequences from Ensembl annotation (GRCh38.82) that can be represented by a linear directed graph. At k = 31 bp, 96.76% of the transcriptome can be used as a target sequence. The transcript requiring the largest k to achieve a linear representation is ENST00000621744_NBPF19 and would require k = 3,472 bp.

	Tumour	Km	GDC portal
R132C: g/A	AML	8	9
	BLCA	1	1
	BRCA	1	1
	CHOL	3	4
	COAD	2	1
	GBM	0	1
	LGG	11	17
	LIHC	2	2
	LUAD	1	1
	PCPG	0	1
	PRAD	1	2
	SARC	1	1
	SKCM	12	15
	THYM	1	1

	Tumour	Km	GDC portal
R132H: c/T	AML	2	4
	BLCA	1	1
	BRCA	1	1
	CHOL	0	0
	COAD	0	0
	GBM	8	23
	LGG	317	358
	LIHC	0	0
	LUAD	0	0
	PCPG	0	0
	PRAD	1	1
	SARC	0	0
	SKCM	0	0
	THYM	0	0

	Tumour	Km	GDC portal
R132G: g/C	AML	1	0
	BLCA	0	0
	BRCA	0	0
	CHOL	0	0
	COAD	1	1
	GBM	2	1
	LGG	13	10
	LIHC	1	1
	LUAD	0	0
	PCPG	0	0
	PRAD	1	1
	SARC	0	0
	SKCM	0	0
	THYM	0	0

	Tumour	Km	GDC portal
R132L: c/A	AML	0	0
	BLCA	0	0
	BRCA	0	0
	CHOL	0	0
	COAD	0	0
	GBM	0	0
	LGG	0	0
	LIHC	0	0
	LUAD	1	1
	PCPG	0	0
	PRAD	0	0
	SARC	0	0
	SKCM	2	2
	THYM	0	0

	Tumour	Km	GDC portal
R132S: g/T	AML	0	0
	BLCA	0	0
	BRCA	0	0
	CHOL	1	1
	COAD	0	0
	GBM	0	0
	LGG	8	9
	LIHC	0	0
	LUAD	0	0
	PCPG	0	0
	PRAD	0	0
	SARC	0	0
	SKCM	0	0
	THYM	0	0

Supplementary table 1: Number of TCGA samples with a mutation on amino acid 132 of IDH1, found by *km* compared to data available on the GDC portal. Variant calling on the GDC portal is done in-silico on Exome sequencing, which can explain the small differences with *km*'s results.

NPM1		Leucegene	
		positive	negative
km	positive	117	0
	negative	1	85

FLT3-ITD		Leucegene	
		positive	negative
km	positive	101	0
	negative	0	259

NPM1		TCGA(AML)	
		positive	negative
km	positive	36	4
	negative	0	114

FLT3-ITD		TCGA(AML)	
		positive	negative
km	positive	31	7
	negative	2	117

Supplementary table 2 Contingency tables on NPM1 insertion and FLT3-ITD, using common samples between km and experimental validations.

Sample	Length	Type	km	cBio	ITDassembly	pIndel	Genomon
2812	51	-		✓	✓	✓	
2823	57	I&I	✓		✓	✓	
2823	72	ITD	✓				
2825	102	I&I	✓	✓		✓	
2830	42	-					✓
2830	56	-				✓	
2830	69	I&I	✓	✓	✓		
2836	33	ITD	✓	✓	✓		
2840	18	ITD	✓	✓	✓		
2840	105	Indel	✓				
2844	87	I&I	✓	✓		✓	
2853	18	ITD	✓	✓	✓		
2853	105	Indel	✓				
2856	22	Indel	✓				
2862	69	-				✓	✓
2863	105	Indel	✓				
2869	54	ITD	✓	✓	✓		✓
2870	12	Indel	✓	✓			
2871	63	ITD	✓	✓			
2875	30	I&I	✓	✓	✓		✓
2877	18	ITD	✓	✓	✓		✓
2880	21	ITD	✓	✓	✓		✓
2895	45	ITD	✓	✓	✓	✓	✓
2895	50	-				✓	
2895	51	I&I	✓		✓		
2896	63	ITD	✓				
2896	153	-				✓	✓
2913	66	ITD	✓	✓		✓	✓
2915	49	-				✓	
2915	51	I&I	✓	✓	✓		✓

Sample	Length	Type	km	cBio	ITDassembly	pIndel	Genomon
2918	88	-				✓	✓
2918	90	I&I	✓				
2919	93	ITD	✓			✓	✓
2921	24	I&I	✓	✓			✓
2921	57	I&I	✓				✓
2925	42	ITD	✓	✓	✓	✓	✓
2930	42	I&I	✓	✓			✓
2931	70	-				✓	
2931	75	I&I	✓	✓			
2934	56	-				✓	
2934	57	I&I	✓	✓	✓		✓
2942	24	ITD	✓	✓	✓		✓
2949	39	I&I	✓		✓		
2959	118	-				✓	
2959	132	I&I	✓				
2965	72	I&I	✓	✓			
2970	57	I&I	✓	✓			
2976	18	I&I	✓				
2976	24	I&I	✓	✓			
2980	21	ITD	✓	✓			
2980	57	ITD	✓				
2981	54	I&I	✓	✓			
2986	48	ITD	✓	✓			
2988	36	-		✓			
2994	21	ITD	✓	✓			
2998	21	-		✓			
2998	51	ITD	✓				
3007	21	-		✓			
3007	24	ITD	✓				
3007	57	ITD	✓				

Supplementary table 3: Summary of all variants found by km, ITDassembler, Pindel and Genomon ITDetector on 28 TCGA AML samples for which exome and RNA sequencing were available.