CRAC: An integrated approach to read analysis

N. Philippe[1,3], M. Salson[2], T. Commes[3], E. Rivals[1]

1 - Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier

2 - Laboratoire d'Informatique Fondamentale de Lille

3 - Institut de Recherche en Biothérapie - CHU de Montpellier

Institut de biologie computationnelle

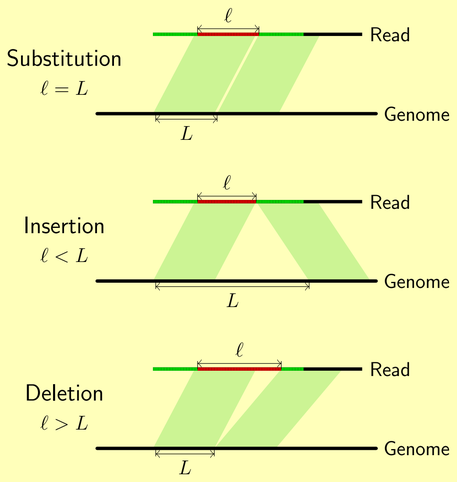
Abstract

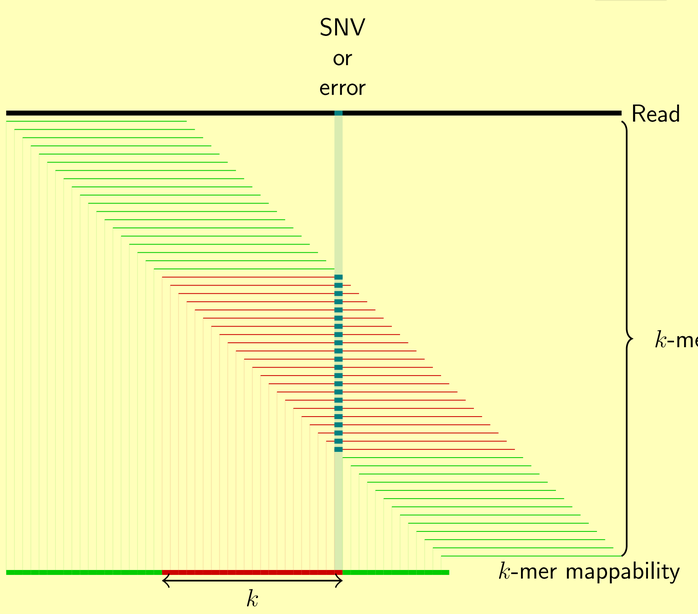
NGS techniques are intensively used to investigate transcriptome complexity using RNA-sequencing assays. Huge read sets are then compared to a reference genome to determine transcribed exons, and unmapped reads are further analyzed for predicting splice junctions. Subsequent analyses seek to reconstruct standard and alternative transcripts. However, transcript inference can only succeed if reads are mapped correctly and splice junctions predicted acurrately. By comparing current solutions we gathered evidence suggesting that read processing was far from fully completing this task. Here, we propose a novel way of analyzing reads that integrates genomic locations and local coverage to distinguish sequence errors from biological mutations, and to infer directly splice junctions within a single read. An evaluation of our program CRAC shows that it improves mapping sensitivity up to 30% compared to state of the art solutions, while being highly specific. Such a sensitivity gain can impact the user's ability to detect rare mutations or splicing variants. Moreover, results indicate that CRAC predicts with bp precision the splice junctions of reads sequenced from non colinear, chimeric RNAs with $[40, 60]\%$ sensitivity and $> 90\%$ specificity, a unique feature to our knowledge. Finally, this integrated read analysis stategy may broaden the scope of the transcriptome that is amenable to discovery using RNA-sequencing approaches.

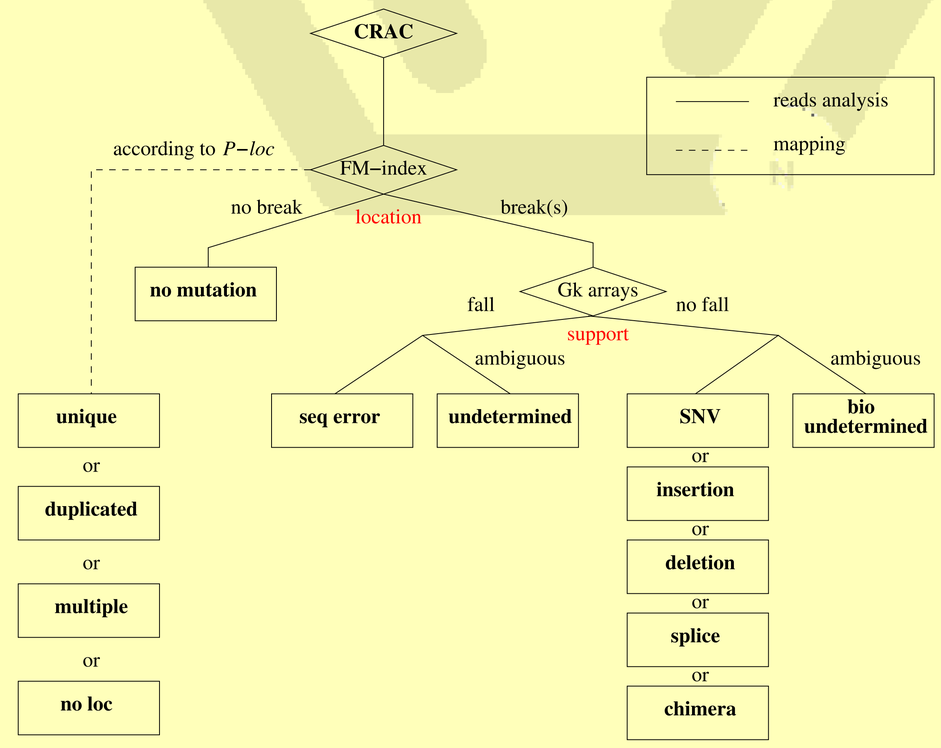
Algorithm

CRAC proceeds each read in turn. For each, it monitors two "signals" that vary with the position in the read sequence of length m. For this, it considers the k-mer starting at every position (ie. $m-k+1$ possible k-mers) in the read and registers:

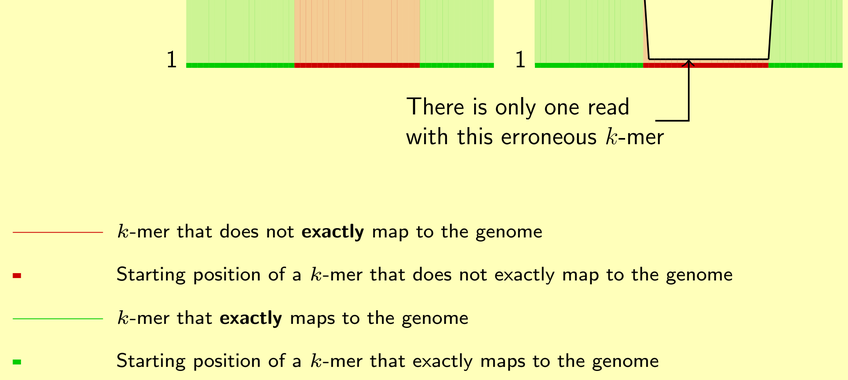
1. **the exact mappability of the k-mer on the reference genome**, its matching locations and their number,
2. **the k-mer support**, which we define as the number of reads sharing this k-mer (ie, the exact same k-mer sequence matches a k-mer from another read). The support value has a minimum value of one since the k-mer exists at least in the current read.







CRAC strategy is to analyse jointly k-mer support and k-mer mappability to predict in a single analysis sequencing errors, as well as potential genetic variations, splice junctions, or chimeras



Results

Biological validations on private AML library:

- $\sim$ 40 millions of unoriented 100 bp reads
- Detection of 511 chimeras

ECHANT NIMES			MYH11-CBFB (INV16 Chimera)			NONE-PAN			ERCC2-SPI1			ZNF217-ZNF48		
NOM	CARYOTYPE		CT	TM	remarque	CT	TM	remarque	CT	TM	remarque	CT	TM	remarque
OS110089	INV16	Applied	27,21/26,96	82,94/83,02	POSITIF	31,29/31,47	79,96/80,20	POSITIF	34,93/36,05	82,02/81,96	POSITIF	40	68,62-84,15	negatif
		Promega	27,21/26,96	82,94/83,02	POSITIF	31,29/31,47	79,96/80,20	POSITIF	34,93/36,05	82,02/81,96	POSITIF	40	68,62-84,15	negatif
OM 090069	INV16	Applied	28,69/28,82	83,09/82,95	POSITIF	neant	none	negatif	39,65/38,40	482,11/75,41	POSITIF	neant/neant	none/none	negatif
		Promega	28,31/28,92	82,82/82,94	POSITIF	neant	none	negatif	38,69/38,83	,23-82,09/82,	POSITIF	neant/neant	none	negatif
OM 110171	INV16	Applied	40/36,59	75,83/82,72	POSITIF	neant	none	negatif	36,53/37,44	82,09/81,94	POSITIF	40/40	79,11/none	negatif
		Promega	35,01/34,89	82,73/83,05	POSITIF	neant	none	negatif	36,73/35,89	82,13/82,24	POSITIF	40/40	86,19/80,78	negatif
OM 110448	INV16	Applied	27,12/27,28	82,86/83,14	POSITIF	neant	none	negatif	35,65/35,56	82,03/82,17	POSITIF	40/40	,29/66,05-79,	negatif
		Promega	26,3/26,18	82,96/82,92	POSITIF	neant/40	76,06	negatif	34,75/33,85	82,14/82,20	POSITIF	38,46/40	81,72/neant	negatif
R2620	INV16	Applied	36,52	82,75	POSITIF	neant	none	negatif	40	none	negatif	neant	none	negatif
		Promega	not tested	not tested	NT	not tested	not tested	not tested	not tested	not tested	NT	not tested	not tested	not tested
OS100380	KN	Applied	40/40	77,4/75,31	negatif	neant	none	negatif	35,19/neant	82,07/none	POSITIF	neant/40	none/76,52	negatif
		Promega	not tested	not tested	NT	not tested	not tested	not tested	not tested	not tested	NT	not tested	not tested	not tested
OM 110564	KN	Applied	neant	none	negatif	neant	none	negatif	neant	none	negatif	36,87	81,98	POSITIF	neant	none	negatif
		Promega	neant	none	negatif	neant	none	negatif	neant	none	negatif	34,64	82,05	POSITIF	40	69,34-84,36	negatif
OM 110522	KN	Applied	37,02	75,97	negatif	neant	none	negatif	neant	none	negatif	37,12	82,1	POSITIF	37,62	81,72	negatif
		Promega	neant	none	negatif	neant	none	negatif	neant	none	negatif	35,59	82,06	POSITIF	40	83,86	negatif
OM 110520	KN	Applied	neant	67,61	negatif	neant	none	negatif	neant	none	negatif	37,32	81,9	POSITIF	40	none	negatif
		Promega	neant	none	negatif	neant	none	negatif	neant	none	negatif	35,59	82,12	POSITIF	40	none	negatif
OM 110424	KN	Applied	neant	none	negatif	40	77,69	proba negatif mais	34,54	82,21	POSITIF	40	66,05-79,40	negatif
		Promega	neant	none	negatif	37,61	75,25	proba negatif mais	33,02	82,07	POSITIF	40	83,47	negatif
OS110518	KN	Applied	neant	none	negatif	neant	none	negatif	33,43	82,46	POSITIF	40	80,44	negatif
		Promega	neant	none	negatif	neant	none	negatif	34,07	82,23	POSITIF	40	78,21	negatif
OM100359	KN	Applied	neant	none	negatif	neant	none	negatif	37,02	82,15	POSITIF	38,12	81,71	negatif
		Promega	40	75,68	negatif	40	77,41	negatif	36,27	82,1	POSITIF	38,21	76,3	negatif
OM100320	KN	Applied	36,61	76,07	negatif	neant	none	negatif	36,01	82,19	POSITIF	40	81,75	negatif
		Promega	40	81,59	negatif	neant	none	negatif	34,68	82,08	POSITIF	neant	none	negatif
OM100311	KN	Applied	neant	none	negatif	neant	none	negatif	36,96	82,15	POSITIF	neant	none	negatif
		Promega	39,36	75,71	negatif	neant	none	negatif	35,55	82,15	POSITIF	40	84,84	negatif
OM100093	KN	Applied	neant	none	negatif	38,78	77,59	proba negatif mais	34,06	82,12	POSITIF	40	78,21	negatif
		Promega	40	none	negatif	neant	none	negatif	33,84	82,13	POSITIF	40	88,88	negatif
Cont		Applied	not tested	not tested	NT	not tested	not tested	not tested	not tested	not tested	NT	not tested	not tested	not tested
		Promega	38,01	75,61	negatif	40	none	negatif	30,52	82,2	NT	39,08	81,83	negatif
BLANC			neant	none	negatif	neant	none	negatif	40	,18/75,69/76,	negatif	neant	none	negatif
			40	75,9	negatif	neant	none	negatif	neant	none	negatif	40	81,67	negatif
VALIDATION			CHIMERA cont INV16			CHIMERA in OM110089 (INV16)			CHIMERA in LAM			CHIMERA ARTEFACTUAL CASE		

Comparative evaluation of mapping sensitivity and precision

Tool	75bp Sensitivity	75bp Precision	200bp Sensitivity	200bp Precision
Bowtie	75.42	99.59	55.72	**99.81**
BWA	79.29	99.13	68.66	96.86
CRAC	**94.51**	**99.72**	**96.02**	**99.92**
GASSST	70.73	99.09	59.43	97.86
GSNAP	**94.62**	**99.88**	**84.84**	99.28
SOAP2	77.6	99.52	56.08	99.78

Comparative evaluation splice junction prediction tools

Tool	75bp Sensitivity	75bp Precision	200bp Sensitivity	200bp Precision
CRAC	79.43	**99.5**	**86.02**	**99.18**
GSNAP	**84.17**	97.03	72.94	97.09
MapSplice	79.89	**97.68**	**84.72**	**98.82**
TopHat	**84.96**	89.59	54.07	94.69

Comparative evaluation of chimeric RNA prediction tools

Tool	75bp Sensitivity	75bp Precision	200bp Sensitivity	200bp Precision
CRAC	**53.89**	**93.84**	**64.86**	**90.18**
MapSplice	2.33	0	**2.63**	0.01
TopHatFusion	**32.73**	42.02		
TopHatFusionPost	12.26	**97.22**		

Conclusions

Highlights:

- Low false positive rate
- Between 60 and 70 % of causes are found (mutations not found are due to a low coverage)
- junctions: more sensitive and more specific than GSNAP, MapSplice, and TopHat

Futur works:

- Transcripts reconstruction (assembly)
- Clinical markers for prognostic and diagnostic
- Chimera variants in myeloid leukemia (normal karyotype)

CRAC is particularly suitable for the data of the future: more massive and longer