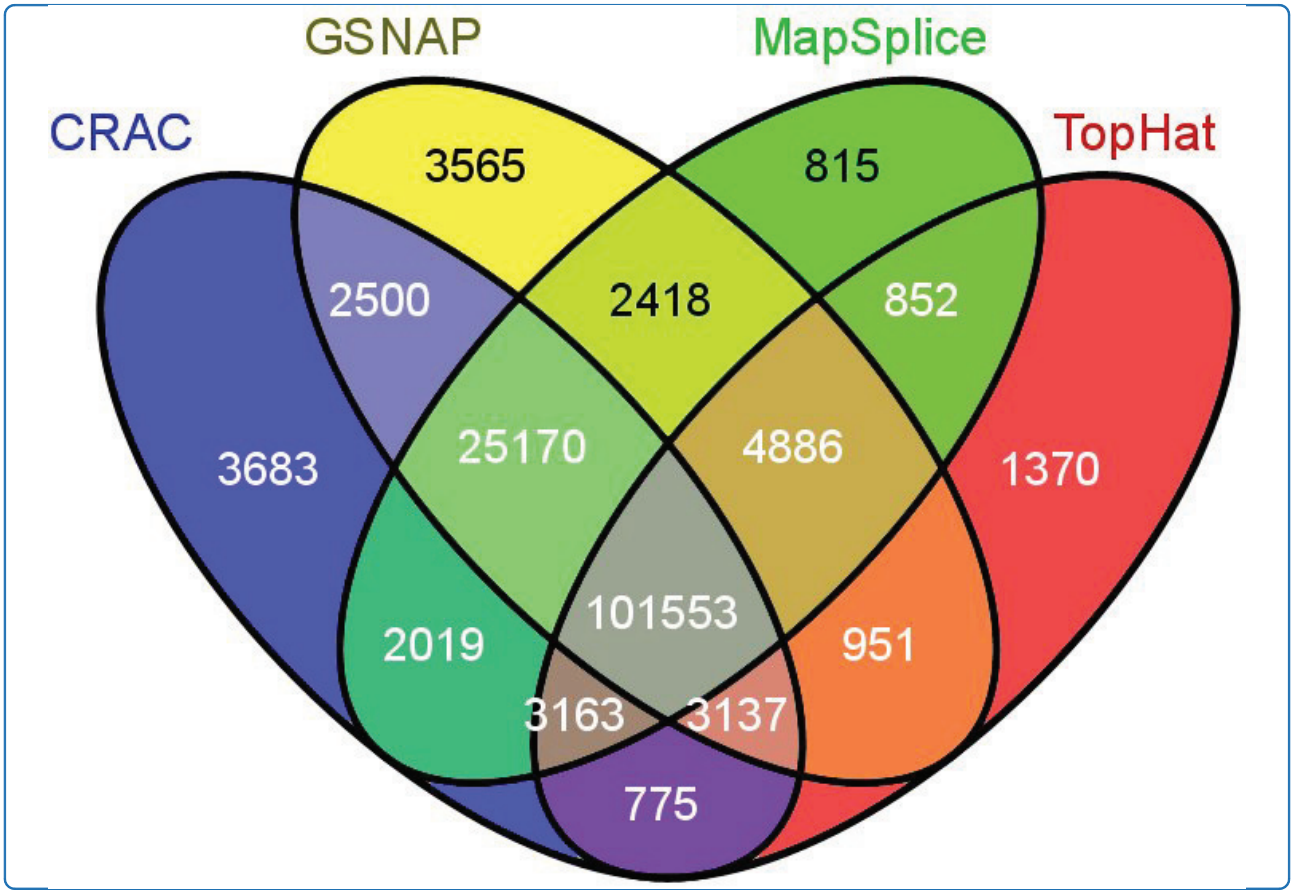
Genome Biology



CRAC: an integrated approach to the analysis of RNA-seq reads

Philippe *et al.*

BioMed Central

Philippe *et al. Genome Biology* 2013, **14**:R30
http://genomebiology.com/content/14/3/R30 (28 March 2013)

Genome Biology

SOFTWARE Open Access

CRAC: an integrated approach to the analysis of RNA-seq reads

Nicolas Philippe[1,2,5†], Mikaël Salson[3,4†], Thérèse Commes[2,5] and Eric Rivals[1,5*]

Abstract

A large number of RNA-sequencing studies set out to predict mutations, splice junctions or fusion RNAs. We propose a method, CRAC, that integrates genomic locations and local coverage to enable such predictions to be made directly from RNA-seq read analysis. A *k*-mer profiling approach detects candidate mutations, indels and splice or chimeric junctions in each single read. CRAC increases precision compared with existing tools, reaching 99:5% for splice junctions, without losing sensitivity. Importantly, CRAC predictions improve with read length. In cancer libraries, CRAC recovered 74% of validated fusion RNAs and predicted novel recurrent chimeric junctions. CRAC is available at http://crac.gforge.inria.fr.

Rationale

Understanding the molecular processes responsible for normal development or tumorigenesis necessitates both identifying functionally important mutations and exploring the transcriptomic diversity of various tissues. RNA sequencing (RNA-seq) provides genome-scale access to the RNA complement of a cell with unprecedented depth, and has therefore proven useful in unraveling the complexity of transcriptomes [1,2]. The analyses of RNA-seq reads aim at detecting a variety of targets: from transcribed exons and classical splice junctions with canonical splice sites, to alternatively spliced RNAs, RNAs with non-standard splice sites, read-through and even non-colinear chimeric transcripts [3]. Moreover, RNA-seq also gives access to those somatic mutations and genetic polymorphisms that are transcribed. Chimeric RNAs result from the transcription of genes fused together by chromosomal rearrangements [4], especially in cancer [5], and they can also be induced by trans-splicing between mature messenger RNAs (mRNAs) [6]. RNA-seq can also capture these complex, non-colinear transcripts, whose molecular importance is still poorly assessed and which may provide new diagnostic or therapeutic targets [7,8].

As next generation sequencing (NGS) improves and becomes cheaper, bioinformatic analyses become more critical and time consuming. They still follow the same paradigm as in the first days of NGS technologies: a multiple step workflow - mapping, coverage computation, and inference - where each step is heuristic, concerned with only a part of the necessary information, and is optimized independently from the others. Consequently analyses suffer from the drawbacks inherent to this paradigm: (a) pervasive erroneous information, (b) lack of integration, and (c) information loss, which induces re-computation at subsequent steps and prevents cross-verification. An example of the lack of integration is that the mapping step cannot use coverage information, which prevents it from distinguishing biological mutations from sequencing errors early in the analysis.

Here, we design a novel and integrated strategy to analyze reads when a reference genome is available. Our approach extracts information solely from the genome and read sequences, and is independent of any annotation; we implemented it in a program named CRAC. The rationale behind it is that an integrated analysis avoids re-computation, minimizes false inferences, and provides precise information on the biological events carried by a read. A peculiarity of CRAC is that it can deliver computational predictions for point mutations, indels, sequence errors, normal and chimeric splice junctions, in a single run. CRAC is compared with state-of-the-art tools for mapping (BWA, SOAP2, Bowtie, and GASSST) [9-13], and both normal (GSNAP, TopHat, and MapSplice)

* Correspondence: rivals@lirmm.fr
† Contributed equally
[1]Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM), UMR 5506, CNRS and Université de Montpellier 2, 161 rue Ada, 34095 Montpellier Cedex 5, France
Full list of author information is available at the end of the article

[3,14,15] and chimeric (TopHat-fusion) [16] splice junction predictions. The results show the relevance of the approach in terms of efficiency, sensitivity, and precision (which is also termed specificity in the literature). We also provide true assessments of the sensitivity of each method by analyzing complex simulated data.

Availability: CRAC is distributed under the GPL-compliant CeCILL-V2 license and is available as source code archive or a ready-to-install Linux package from the CRAC project website [17] or the ATGC bioinformatics platform [18]. It includes two programs: crac-index to generate the index of the genome, and crac for analyzing the reads.

Algorithm

Overview

CRAC is a method for analyzing reads when a reference genome is available, although some procedures (for example, the support computation) can be used in other contexts as well. CRAC analysis is solely based on the read collection and on the reference genome, and is thus completely independent of annotations. CRAC disregards the sequence quality information of reads. Here, analyzing reads means detecting diverse biological events (mutations, splice junctions, and chimeric RNAs) and sequencing errors from a RNA-seq read collection.

CRAC analysis is based on two basic properties: P1 and P2.

P1: For a given genome size, a sequence of a specific length will match on average to a unique genomic position with high probability. This length, denoted k, can be computed and optimized [19]. Thus, in a read any k-mer (a k-long substring) can be used as a witness of the possible read matching locations in the genome. A k-mer may still have a random match to the reference genome. However, in average over all k-mers, the probability of getting a false location (FL) is approximately 10^{-4} with $k = 22$ for the human genome size [19].

P2: As reads are sequences randomly sampled from biological molecules, several reads usually overlap a range of positions from the same molecule. Hence, a sequencing error that occurs in a read should not affect the other reads covering the same range of positions. In contrast, a biological variation affecting the molecule should be visible in many reads overlapping that position.

CRAC processes each read in turn. It considers the k-mers starting at any position in the read (that is, $m - k + 1$ possible k-mers). It computes two distinct k-mer profiles: the location profile and the support profile.

- The **location profile** records for each k-mer its exact matching locations on the genome and their number.
- The **support profile** registers for each k-mer its support, which we define as the number of reads sharing this k-mer (that is, the k-mer sequence matches exactly a k-mer of another read). The support value has a minimum value of one since the k-mer exists in the current read.

CRAC's strategy is to analyze these two profiles jointly to detect multiple events and predict sequencing errors in a single analysis, as well as potential genetic variations, splice junctions, or chimeras (Additional file 1). The genomic locations of a k-mer are computed using a compressed index of the reference genome, such as a compressed Burrows-Wheeler transform [20], while the support of a k-mer is obtained on-the-fly by interrogating a specialized read index, called a Gk arrays [21]. CRAC ignores the pairing information of paired end reads. Each read in a pair is processed independently of the other.

Clearly, the support is a proxy of the coverage and allows property P2 to be exploited for distinguishing sequencing errors from variations, and gaining confidence in predictions. As illustrated below, the location profile delivers a wealth of information about the mapping, but the originality of CRAC is its ability to detect the concordance of variations in the two profiles.

Description of the algorithm

In a collection, some reads will exactly match the reference genome, while others will be affected by one or more differences (with a probability that decreases with the number of differences). Here, we describe how a read is processed and concentrate on reads that differ from the reference. For clarity, we make simplifying assumptions: (a) k-mers have no false genomic locations, (b) the read is affected by a single difference (substitution, indel, or splice junction), and (c) this difference is located $>k$ nucleotides away from the read's extremities (otherwise, we say it is a border case). These assumptions are discussed later.

Consider first a substitution, which may be erroneous (a sequencing error) or of biological origin (an SNP, single nucleotide variant (SNV), or editing). Say the substitution is at position h in the read. All k-mers overlapping position h incorporate this difference and will not match the genome. Thus, the location profile will have zero location for k-mers starting in the range $[h - k + 1, h]$. In contrast, k-mers starting left (respectively right) of that range will have one location in the genome region where the RNA comes from. Moreover, locations of the k-mers starting in $h - k$ and $h + 1$ are $k + 1$ nucleotides apart on the genome. We call the range of k-mers having zero location, a break (Figure 1a). This allows the location of the difference in both the read and the genome to be found, but does not distinguish erroneous from biological differences. The support profile will inform us on this matter.

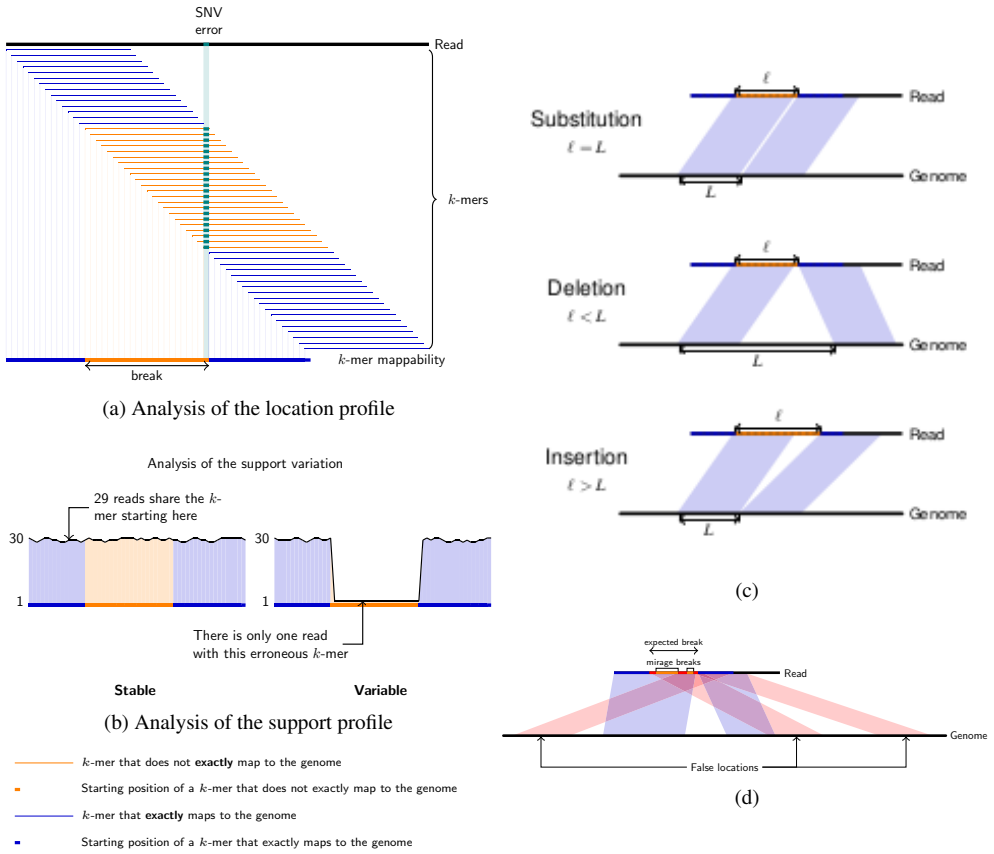


Figure 1 The CRAC algorithm. **(a)** Illustration of a break in the location profile. We consider each *k*-mer of the read and locate it exactly on the genome. In all figures, located *k*-mers are shown in blue, and unmapped *k*-mers in light orange. If the read differs from the genome by, for example an SNV or an error, then the *k*-mers containing this position are not located exactly on the genome. The interval of positions of unmapped *k*-mers is called a break. The end position of the break indicates the error or SNV position. **(b)** The support profile. The support value of a *k*-mer is the number of reads from the collection in which this *k*-mer appears at least once. The two plots show the support profile as a black curve on top of the location profile (in blue and orange). The support remains high (left plot) over the break if many reads covering this region are affected by a biological difference (for example, a mutation); it drops in the region of the break when the analyzed read is affected by a sequencing error; in this case, we say the support is dropping. **(c)** Rules for differentiating a substitution, a deletion, or an insertion depending on the break. Given the location profile, one can differentiate a substitution, a deletion, or an insertion by computing the difference between the gap in the genome and the gap in the read between *k*-mers starting before and after the break. **(d)** False locations and mirage breaks. When false locations occur inside or at the edges of a break they cause mirage breaks. False locations are represented in red. The break verification and break merging procedures correct for the effects of false locations to determine the correct break boundaries (and for example the correct splice junction boundaries) to avoid detecting a false chimera (Rule 2a) instead of a deletion. SNV: single nucleotide variant

If the substitution is a sequencing error, it is with high probability specific to that read. Hence, the *k*-mers overlapping the substitution occur in that read only: their support value is one (minimal). If the substitution is biological, a sizeable fraction of the reads covering this transcript position share the same *k*-mers in that region. Their support remains either similar to that of *k*-mers outside the break or at least quite high depending on the homozygosity or heterozygosity of the mutation. An erroneous difference implies a clear drop in the support profile over the break (Figure 1b). Thus, the ranges of the location break and the support drop will coincide for an error, while a biological difference will not specifically alter the support profile over the break. To detect this drop we compare the average support inside versus outside the break using a separation function (Figure 1b and Additional file 2). Using this procedure, support profiles are classified as undetermined if the support is too low all along the read, and otherwise as either dropping or non-dropping. Reads with a dropping support profile are assumed to incorporate sequencing errors, and those with a non-dropping support to accurately represent sequenced molecules.

This procedure can be generalized to differences that appear as long indels; all cases are summarized by a detection rule. We can apply a similar location/support profile analysis to predict such events.

Rule 1 (Figure 1c): Consider a read affected by a single difference (substitution, indels) compared to the genome.

Let $j_b < j_a$ (where b stands for before and a after) be the positions immediately flanking the observed break in the location profile (that is, the break is in the range $[j_b + 1, j_a - 1]$). Let $l := j_a - j_b$. L denotes the offset between the genomic locations of the k-mers starting in j_b and j_a, so that $L := \text{loc}(j_a) - \text{loc}(j_b)$. (1) If $l = L = k + 1$ the difference consists of a single substitution at position $j_a - 1$ in the read and $\text{loc}(j_a) - 1$ in the genome. (2) If $l = k$ and $L = k + p$ for some integer p, then this is a p nucleotide deletion with respect to the reference genome, which is located between position $j_a - 1$ and j_a in the read, and between $\text{loc}(j_a) - p$ and $\text{loc}(j_a) - 1$ on the genome. (3) Symmetrically, if $l = k + p$ and $L = k$ for some integer p, the difference is a p nucleotide insertion with respect to the reference.

We call the k-mer concordance the condition that loc (j_a) and $\text{loc}(j_b)$ are on the same chromosome, the same strand, and that $\text{loc}(j_a) - \text{loc}(j_b)$ equals $j_a - j_b$ plus or minus the inferred difference (that is, 0 for a substitution and p for indels). This notion can be extended to all k-mer pairs on each side of the break (that is, not merely j_b, j_a).

The observed missing part in the read can be due to a polynucleotidic deletion or the removal of intronic or intragenic regions by splicing. Without annotations, only the expected length (that is, the value of p) can distinguish these cases. CRAC uses arbitrary, user-defined thresholds to classify such biological deletions into short deletions and splice junctions. CRAC does not use splice site consensus sequences.

Rule 2: Other reads may present profiles not considered in Rule 1. In particular, some reads will have a break but the genomic locations at its sides are either on distinct chromosomes or not colinear on the same chromosome. We term these chimeric reads (by chimeric we mean made of a non colinear arrangement of regions rather than unreal), and consider three subcases corresponding to possible known combinations [4]: (a) same chromosome, same strand but inverted order, (b) same chromosome but different strands, and (c) different chromosomes. (For chimeric RNAs, CRAC can even distinguish five subclasses; see Additional file 2 for details). CRAC can handle such cases with the profile analysis. These cases resemble that of deletions (Rule 1, case 2), except that the genomic locations are not colinear. Indeed, CRAC checks the break length $l = k$, as well as the coherence of adjacent k-mers left or right of the break. Coherence means that, for some (small) integer δ, k-mers in the range $[j_b - \delta, j_b]$ (respectively, $[j_a, j_a + \delta]$) have adjacent locations on the genome. Reads satisfying these criteria and harboring a non-dropping support profile are primarily classified as chimeric reads, which may reveal artifactual or sheer chimeric RNAs (chRNAs) (see Discussion).

CRAC processes reads one by one, first by determining the location breaks, then analyzing the support profile, and applying the inference rules whenever possible. A read is classified according to the events (SNV, error, indels, splice, or chimera) that are predicted, and its mapping unicity or multiplicity. Additional file 1 gives an overview of the classification. The CRAC algorithm is described for the analysis of an individual read, but its output can be parsed to count how many reads led to the detection of the same SNV, indel, splice, or chimera; this can serve to further select candidates. CRAC accepts the FASTA and FASTQ formats as input, and outputs distinct files for each category, as well as a SAM formatted file with mapping results.

In describing CRAC's method above, we first assumed simplifying conditions: especially the absence of false locations (FLs) and border cases. Some details will clarify how the actual procedure handles real conditions.

Differences with the genome at a read's extremities (border cases)

Border cases are not processed with a specific procedure by CRAC; instead, the sequencing depth of NGS data indicates border cases. While processing a read, if an event (say, a splice junction) generates a break at one of the read's extremities, the coverage ensures that the same event is likely located in the middle of other reads, and will be detected when processing these. The border case read is classified either as undetermined or biologically undetermined depending on its support profile, and it is output in the corresponding files.

False locations (Figure 1d)

Our criterion to set k ensures a low average probability of a random k-mer match on the genome [19], but it does not prevent random matches, which we term false locations. Compared to true (unique or multiple) locations, FL of a k-mer will generally not be coherent with those of neighboring k-mers. It may also alter the break length in an unexpected manner, making the break length another criterion of verification (Rule 1). When a read matches the genome, CRAC considers ranges of k-mers having coherent locations to infer its true genomic position. In the case of a break, CRAC faces two difficulties. First, when a FL happens at the end of a break, CRAC may incorrectly delimit the break. When a FL occurs inside a break, it makes adjacent false breaks, termed mirage breaks. In both cases, the FL may cause CRAC to avoid Rule 1, apply Rule 2, and predict a false chimeric read. To handle a FL at a break end, CRAC uses a break verification procedure, and it applies a break fusion procedure to detect and remove mirage breaks.

These procedures are detailed in Additional file 2, which also includes explanations of the distinction of dropping and non-dropping supports around a break, on read

mapping at multiple locations, on the subclassification of chimeric reads, and on the simulation protocol.

Results

We evaluated CRAC for mapping reads, predicting candidate SNVs, indels, splice junctions, and chimeric junctions, and compared it to other tools. Simulated data are needed to compute exact sensitivity and accuracy levels, while real data enable us to study predictions with biologically validated RNAs. For simulating RNA-seq, we first altered a reference genome with random substitutions, indels, and translocations to derive a mutated genome, then reads were sequenced *in silico* using FluxSimulator [22], the annotated RefSeq transcripts, and a realistic distribution of random expression levels (Additional file 2). As read lengths will increase, we used two simulated datasets to assess different strategies: one (hs75) with a typical read length of 75, another (hs200) with reads of 200 nt representing the future.

Mapping with current (75 nt) and future (200 nt) reads

Mapping, that is, the process of determining the location of origin of a read on a reference genome, provides critical information for RNA-seq analysis. Currently used mappers (Bowtie, BWA, SOAP2 and Bowtie2) compute the best continuous genome-read alignments up to a certain number of differences [9,11,12,23]. CRAC and GSNAP [14], also consider discontinuous alignments to search for the locations of reads spanning a splice junction: they can find both continuous and spliced alignments.

An overview of mapping results with 75 nt reads (Table 1) indicates a high level of precision, but strong differences in sensitivity among tools. All achieve a global precision >99%, meaning that output genomic positions are correct. Bowtie, BWA, and SOAP2 are similar by design, and all look for continuous alignments with a few substitutions and small indels. Although its approach differs, GASSST also targets these (and is better for longer indels). Even within this group, the sensitivity varies significantly: from 70% for GASSST to 79% for BWA. These figures are far from what can be achieved on RNA-seq data since GSNAP and CRAC, which also handle spliced reads, reach 94% sensitivity: a difference of at least 15 points compared to widely used mappers (Bowtie2 included). As only uniquely mapping reads were counted, the sensitivity cannot reach 100%: some reads are taken from repeated regions and thus cannot be found at a unique location.

Table 1 Comparative evaluation of mapping sensitivity and precision

	75 bp		200 bp	
Tool	**Sensitivity**	**Precision**	**Sensitivity**	**Precision**
Bowtie	75.42	99.59	55.72	**99.81**
Bowtie2	76.64	99.26	62.31	98.78
BWA/BWA-SW	79.29	99.13	68.66	96.86
CRAC	*94.51*	*99.72*	**95.9**	*99.79*
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

We compared the sensitivity and precision of different tools on the human simulated RNA-seq (42M, 75 nt and 48M, 200 nt) against the human genome for mapping. The sensitivity is the percentage of correctly reported cases over all sequenced cases, while the precision is the percentage of correct cases among all reported cases. Values in bold in the three tables indicate the maximum of a column, and those in italics the second highest values. For all tasks with the current read length, CRAC combines good sensitivity and very good precision. Importantly, CRAC always improves sensitivity with longer reads, and delivers the best sensitivity while keeping a very high precision.

One gets a clearer view by considering the subsets of reads that carry an SNV, an indel, an error, a splice, or a chimeric junction (Figure 2). Strikingly, CRAC is the only tool that achieves similar performance, a sensitivity of 94% to 96%, in all categories. For instance with indels, GSNAP yields 65% and 83% sensitivity on insertions and deletions respectively, Bowtie2 yields 70% sensitvity for both insertions and deletions, while the other tools remain below 30%. BWA, GASSST, Bowtie, and SOAP2 output continuous alignments for 9% to 19% of spliced reads, and Bowtie2 up to 35%. Although their output locations are considered correct, for they are in one exon, their alignments are not. Such reads are considered as mapped and thus not reanalyzed by tools like TopHat or MapSplice in a search for splice junctions, which may lead to missing junctions.

Analyzing longer reads (200 nt) is another challenge: the probabilities for a read to carry one or several differences (compared to the reference) are higher. In this dataset, 36% of the reads cover a splice junction, and 50% carry an error. Compared to the 75 nt data, while their precision remains >99%, BWA, GASSST, Bowtie, Bowtie2, SOAP2, and GSNAP, have lower sensitivity (approximately 10 points less for BWA-SW, GASSST, and GSNAP, 14 for Bowtie2, and 20 for Bowtie). Only CRAC remains as precise and gains 1.5 points in sensitivity (Table 1). The detail by category confirms this situation (Figure 2), showing CRAC is better than current tools. CRAC's *k*-mer profiling approach can accurately handle reads altered by distinct categories of biological events, and importantly adapts well to longer reads.

The same analyses have been performed on *Drosophila* datasets and these show that all tools perform better, but the differences between tools remain (Additional file 3). The run times and memory usage of all tools are given in Additional file 3, Table S3. CRAC requires a large memory and its run time for analyzing reads ranges between that of Bowtie and TopHat, which are practical tools. Indexing the human genome with crac-index takes two hours on an x86_64 Linux server on a single thread and uses 4.5 gigabytes of memory.

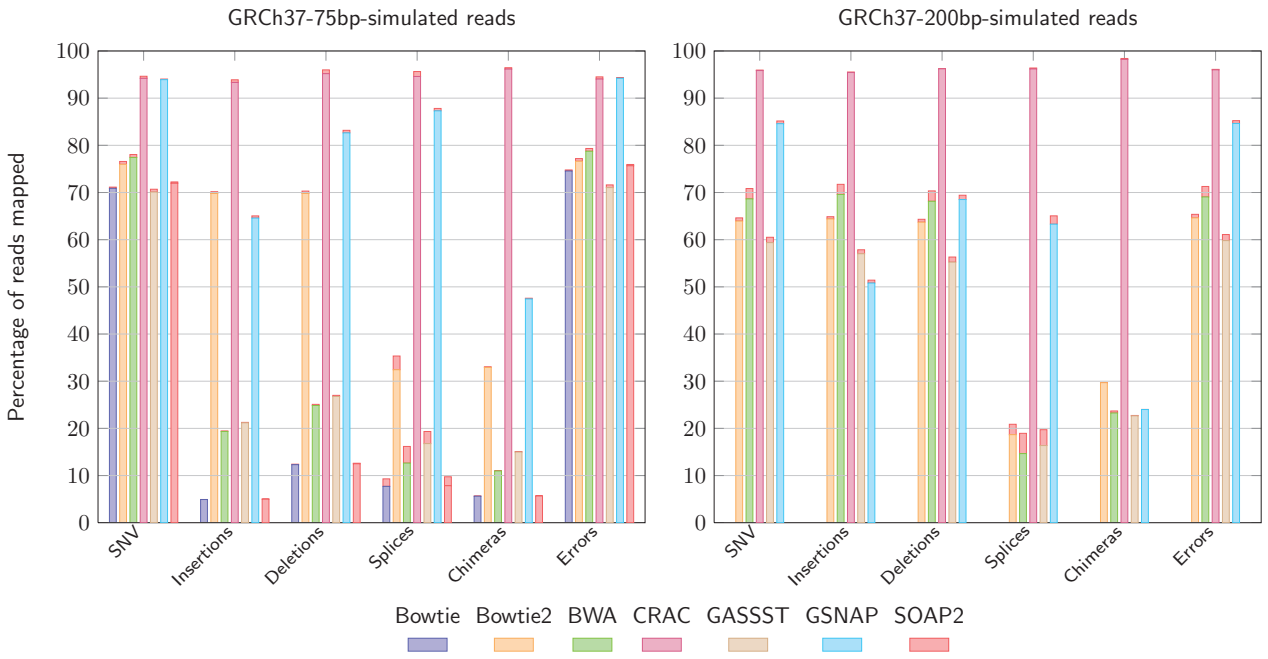


Figure 2 Comparison of mapping results by category for seven tools. The figure shows the mapping by event category on simulated RNA-seq against the human genome on datasets with short and long reads (left 42M, 75 nt; right 48M, 200 nt) for seven different mapping tools: Bowtie, Bowtie2, BWA/BWA-SW, CRAC, GASSST, GSNAP, and SOAP2. We consider six categories of reads depending on whether they contain an SNV, an insertion, a deletion, a junction, a sequence error, or a chimeric splice junction (a chimera). In each category, the bar is the percentage of those reads mapped at a unique location by the corresponding tool. The red tip at the top of the bar is the percentage of incorrectly mapped reads. With 75 nt reads, CRAC is better than the other tools, reaching a sensitivity >90% and a precision >95% whatever the category. The other tools except GSNAP are below 50% sensitivity for mapping reads in categories where spliced alignments are needed (for which they are not intended) and for reads containing insertions or deletions. With 200 nt reads, CRAC remains by far the most sensitive and specific tool; the difference between CRAC and GSNAP and Bowtie2 increased in all categories. Compared to short reads, the other tools had a better mapping of insertion and deletion containing reads. SNV: single nucleotide variant

Predicting distinct categories of biological events

Mapping is not a goal per se, but only a step in the analysis; the goal of read analysis is to detect candidate biological events of distinct categories (SNVs, indels, and splice and chimeric junctions) from the reads. The question is: if, for example, there is an SNV or splice junction that has been sequenced, can it be predicted and not buried under a multitude of false positives (FPs)? Here, sensitivity and precision are relative to the number of events, not to the number of reads covering them. We assessed CRAC's prediction ability and compared it to splice junction prediction tools on our simulated datasets.

Figure 3 gives CRAC's precision and sensitivity for each category of events and for sequencing error detection. For SNVs and indels (<15 nt), CRAC achieves a sensitivity in the range [60,65]% and a precision in the range [96.5,98.5]% (Figure 3), making it a robust solution for such purposes. Typically, CRAC missed SNVs that either have low coverage (42% of them appear in ≤2 reads) or are in reads carrying several events (66% of missed SNV reads also cover a splice junction). For the splice junction category, CRAC delivers 340 false and 67,372 true positives (TPs).

An overview and the effect of read length on sensitivity and precision are shown in Table 2. With 75 nt, all splice detection tools achieve good sensitivity, ranging from 79% for CRAC to 85% for TopHat, but their precision varies by more than 10 points (range [89.59,99.5]). CRAC reaches 99.5% precision and thus outputs only 0.5% FPs; for comparison, MapSplice and GSNAP output four times as many FPs (2.32% and 2.97%), while TopHat yields 20 times more FPs (10.41%). With 200 nt reads, tools based on *k*-mer matching, that is CRAC and MapSplice, improve their sensitivity (6.5 and 5 points respectively), while mapping-based approaches (GSNAP and TopHat) lose, respectively, 12 and 30 points in sensitivity, and TopHat2 gains 6.4 points in sensitivity. With long reads, CRAC has the second best sensitivity and the best precision (>99%). It also exhibits a better capacity than MapSplice to detect junctions covered by few reads: 15,357 vs 13,101 correct junctions sequenced in ≤4 reads.

A comparison using chimeric RNAs shows that CRAC already has an acceptable balance between sensitivity and precision with 75 nt reads (53% and 93%, respectively), while the sensitivities of TopHat-fusion and MapSplice remain below 32% (Table 3). With 200 nt

(A)

		SNV	Ins.	Del.	Splices	Errors	Chimera
200bp	True positives	37,833	3,347	3,290	125,530	31,023,122	1,185
	False positives	1,899	151	68	1,027	139,676	129
75bp	True positives	18,670	1,641	1,637	67,372	10,324,528	624
	False positives	609	57	26	340	37,660	41

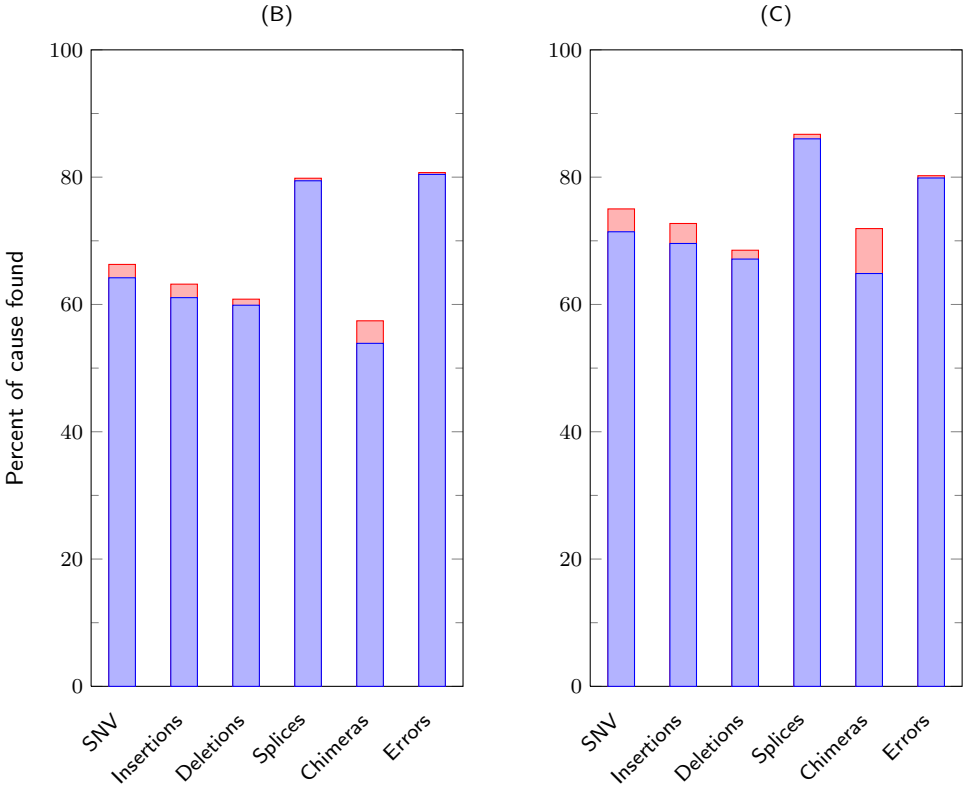


Figure 3 Sensitivity and precision of CRAC predictions by category for human simulated data. **(A)** Absolute numbers of true and false positives reported by CRAC. These figures are the number of distinct events, say SNVs, reported by CRAC, not the number of reads containing the same SNV. False positives represent a small fraction of its output, thereby indicating a high level of precision. **(B)** and **(C)** For each category, the figure shows the proportion of events found by CRAC for the 75 nt and 200 nt datasets. The blue bars are the true positives, while the red bars on top are the false positives. The height of a blue bar gives CRAC's sensitivity, and the relative height of the red part of the bar gives the precision. For the two read lengths, for all categories the sensitivity increases with longer reads, while the precision in each category varies only a little. SNV: single nucleotide variant

reads, only CRAC is able to predict chimeric splice junctions with acceptable precision, and sensitivity is improved compared to shorter reads (Table 3 and Additional file 3).

As with mapping, for all categories of event, CRAC's prediction performance improves with longer reads (Figure 3).

Predicting distinct categories of biological events on real data

Splice junction prediction

To evaluate CRAC's ability to detect splice junctions in real RNA-seq data, we compared it to state-of-the-art tools (TopHat, GSNAP, and MapSplice) on a dataset of 75 million stranded 100 nt reads (ERR030856; see Additional file 4 Table S1). Splice junctions were searched for using each tool and then compared to human RefSeq transcripts. Each found junction consists of a pair of genomic positions (that is, the exons 3' end and 5' start) and we considered that it matches a RefSeq junction if the positions were equal within a 3 nt tolerance. Found junctions were partitioned into *known*, *new*, and *other junctions* (KJs, NJs, and OJs, respectively). Known junctions are those already seen in a RefSeq RNA, new ones involve RefSeq exons but in a combination that has not yet been observed in RefSeq, while the remaining junctions go into

Table 2 Comparative evaluation of splice junction prediction tools

	75 bp		200 bp	
Tool	**Sensitivity**	**Precision**	**Sensitivity**	**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
TopHat2	82.25	92.71	**88.65**	91.35

We compared the sensitivity and precision of different tools on the human simulated RNA-seq (42M, 75 nt and 48M, 200 nt) against the human genome for splice junction prediction. The sensitivity is the percentage of correctly reported cases over all sequenced cases, while the precision is the percentage of correct cases among all reported cases. Values in bold in the three tables indicate the maximum of a column, and those in italics the second highest values. For all tasks with the current read length, CRAC combines good sensitivity and very good precision. Importantly, CRAC always improves sensitivity with longer reads, and yields the best precision (that is the fewer false positives) over all solutions, even against specialized tools like TopHat.

the class other. Note that known RefSeq junctions include both junctions between neighboring exons and alternative splicing cases, mostly caused by exon skipping or alternative splice sites [24]. Novel junctions will provide new alternative splicing candidates, while junctions in class other are totally new candidate RNAs.

For each tool, the distribution of junctions in the classes, and the number of detected RefSeq RNAs and genes (those having at least one KJ or NJ) are given in Figure 4a. The agreement on known junctions (KJs) among the tools is shown as a Venn diagram (Figure 4b); see Additional file 4 for the corresponding figures and a Venn diagram on novel junctions (NJs). Clearly, MapSplice, GSNAP, and CRAC find between [140,876;144,180] known junctions and all three agree on 126,723 of them. GSNAP and CRAC share 93% of CRAC's reported known junctions. TopHat reports about 25,000 junctions fewer than the other tools, and only 1,370 of its junctions are not detected by any of them. For instance, CRAC covers 93% of TopHat's KJs. As known junctions likely contain truly expressed junctions of well-studied transcripts, these figures assess the sensitivity of each tool and suggest that in this respect CRAC equals state-of-the-art tools. Logically, the numbers vary more and the agreements are less pronounced among novel junctions. A marked difference appears within the class other: CRAC yields only 20.36% of other junctions, while with the other tools find [25;27]% of detected junctions.

Table 3 Comparative evaluation of chimeric RNA prediction tools

	75 bp		200 bp	
Tool	**Sensitivity**	**Precision**	**Sensitivity**	**Precision**
CRAC	*53.89*	*93.84*	*64.86*	**90.18**
MapSplice	2.33	0	2.63	0.01
TopHat2	**77.72**	7.32	**70.72**	*12.50*
TopHat-fusion	32.73	42.02		
TopHat-fusion-post	12.26	**97.22**		

We compared the sensitivity and precision of different tools on the human simulated RNA-seq (42M, 75 nt and 48M, 200 nt) against the human genome for chimeric junction prediction. The sensitivity is the percentage of correctly reported cases over all sequenced cases, while the precision is the percentage of correct cases among all reported cases. Values in bold in the three tables indicate the maximum of a column, and those in italics the second highest values. For all tasks with the current read length, CRAC combines good sensitivity and very good precision. Importantly, CRAC always improves sensitivity with longer reads, and has the best balance between sensitivity and precision. TopHat-fusion could not process 200 nt reads.

To further test CRAC with negative controls, we created a set of 100,000 random junctions by randomly associating two human RefSeq exons, and for each we built a 76 nt read with the junction point in the middle of the read (see Additional file 4). These 100,000 reads were processed by CRAC with k = 22 and it predicted no splice junctions.

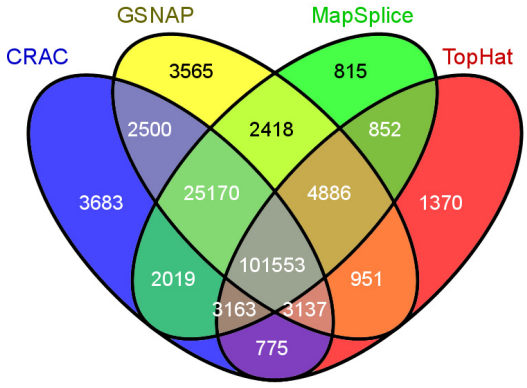
Are the junctions in classes New and Other interesting candidates? To check predicted junctions, we extracted a 50 nt sequence around each inferred junction point and aligned it with BLAST against the set of human mRNAs/ESTs (for details and results see Additional file 4). A 50 nt sequence can either match over its entire length on an EST or match only one side of the junction but not both exons. The former confirms the existence of that junction in the ESTs and yields a very low E-value ($\leq 10^{-15}$), while the latter has a larger value ($\geq 10^{-10}$). As expected, at least 95% of KJs have very low E-values against ESTs, whatever the tool. Among new and other junctions, BLAST reports good alignments for respectively 68% and 69% of CRAC's junctions. The corresponding figures are 47% and 47% for GSNAP, 49% and 50% for MapSplice, 51% and 44% for TopHat. The percentages of OJs and NJs confirmed by mRNAs are >13% for CRAC and <8% for all other tools (excepted for OJs with TopHat, which was 17%, the same as CRAC). If we consider all junctions, 93% of CRAC junctions align entirely to an EST with a good hit. Whatever the class of the junctions, CRAC predicts more unreported junctions that are confirmed by mRNAs or ESTs than the other tools. This corroborates the precision rates obtained by these tools on simulated data.

Regarding expressed transcripts, all tools detect >18,000 transcripts and agree on 17,131 of them (Additional file 4 Figure S1). GSNAP and CRAC agree on 97% (19,431) of CRAC's detected transcripts, expressed in 15,589 distinct genes, which represents 87% of the 17,843 multi-exon RefSeq genes.

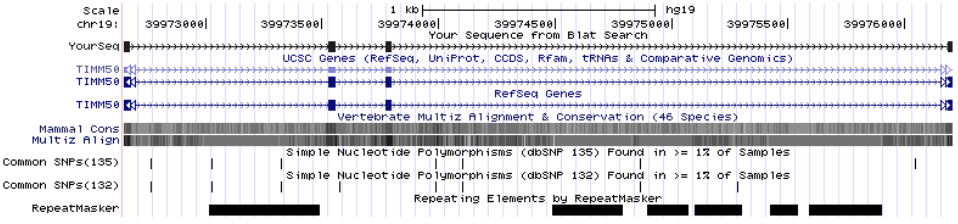
By simultaneously exploiting the genomic locations and support of all k-mers gives CRAC some specific abilities for junction detection. CRAC reports 752 junctions with an intron larger than 100 knt. The other tools find fewer of these junctions: 695, 589, and 470 for GSNAP, MapSplice, and TopHat, respectively, but both MapSplice and TopHat find fewer than expected by chance according to the global

ERR030856	CRAC		MapSplice		TopHat		GSNAP	
	%	#	%	#	%	#	%	#
known SJ	77.63	142,000	68.67	140,876	71.02	116,687	68.12	144,180
new SJ	2.01	3,671	4.35	8,921	3.62	5,956	5.13	10,861
other SJ	20.36	37,254	26.98	55,349	25.35	41,667	26.76	56,626
RefSeq RNAs		19,998		19,549		18,326		20,313
RefSeq genes		15,868		15,825		15,223		15,935

(a)



(b)



(c)

Figure 4 Splice junction detection using human real RNA-seq: comparison and agreement. The figure shows the detection of splice junctions by CRAC, MapSplice, TopHat, and GSNAP for a human six-tissue RNA-seq library of 75M 100 nt reads (ERR030856). **(a)** Number and percentage of known, new, and other splice junctions detected by each tool with +/−3 nt tolerance for ERR030856. **(b)** Venn diagram showing the agreement among the tools on known RefSeq splice junctions (KJs). Additional file 4 has pending data for novel junctions (NJs) and RefSeq transcripts. **(c)** A read spanning four exons (2 to 5) and three splice junctions of the human TIMM50 gene displayed by the UCSC genome browser. The included exons, numbers 3 and 4, measure 32 and 22 nt, respectively. So exon 3 has exactly the *k*-mer size used in this experiment. KJ: known splice junction; SJ: splice junction

agreement between these tools (Additional file 4). CRAC also reveals 69,674 reads that cover exactly two known RefSeq junctions, that is, that cover three distinct exons and include one of them. An example of a double junction covering a 29 nt exon of the CALM2 gene is shown in Additional file 4. Moreover, of 9,817 of such junctions, GSNAP, MapSplice, and TopHat, find respectively 8,338, 9,167, and 7,496, which for GSNAP and TopHat is less than expected by taking a random sample of junctions (Additional file 4). CRAC even maps reads spanning 3 successive junctions (4 exons), and finds an additional 89 junctions, which are not all reported by current tools. For instance, GSNAP does not map such reads. An example for the TIMM50 gene is shown in Figure 4c. Altogether, these results suggest that numerous new splice junctions, even between known exons, remain to be discovered [25], but other predicted junctions that would correspond to completely new transcripts may also be due in part to the

inaccuracy of splice junction prediction tools. In this respect, CRAC seems to ally sensitivity and precision, which should help discriminate true from false candidates, while it has good potential for detecting multiple junctions occurring within the same read. Such reads with multiple junctions will be more abundant with longer reads, and are useful for the reconstruction of transcripts, which is done on the basis of detected junctions [26].

Comparisons of chimeric splice junction prediction

Edgren et al. used deep RNA-sequencing to study chimeric gene fusions in 4 breast cancer cell lines (BT-474, KPL-4, MCF-7, and SK-BR-3; see Additional file 4 Table S1); they found 3 known cases and validated 24 novel intergenic fusion candidates (that is, involving 2 different genes) [27]. As CRAC, TopHat-fusion can predict both intragenic and intergenic chRNA candidates and identify a chimeric junction in a spanning read [16]. For evaluation purposes, we processed each library with TopHat-fusion and CRAC, and compared their results. TopHat-fusion exploits both the read sequence and the read pairs, while CRAC uses only the single read sequence. Otherwise, TopHat-fusion per se[1] and CRAC both select potential chRNAs based on computational criteria. We further filtered out all candidate chimeric reads for which an alternative, colinear alignment was found by GSNAP (Additional file 4). Then, filtered predictions were compared with valid chRNAs. A post-filtering script, called TopHat-fusion-post, based on biological knowledge, can be applied to TopHat-fusion results, but in [16] its parameters were chosen 'using the known valid fusions as control', and may have biased the comparison. So, we recalculated all predictions using TopHat-fusion with and without TopHat-fusion-post.

The numbers of distinct candidate chimeric junctions (chRNA for short) and chimeric single reads detected by both tools in each library are given in Table 4.

The 50 nt reads, which are well suited for Bowtie and TopHat, are unfavorable for CRAC, which performs better with longer reads. Globally after filtering with GSNAP, TopHat-fusion reports a total of 193,163 chRNAs, while CRAC outputs 455: a 600-fold difference. Compared to the results obtained above for a six-tissue library (ERR030856), TopHat-fusion reports about as many chimeric junctions as CRAC, GSNAP, or MapSplice for normal splice junctions. Such a set likely includes a majority of false positives as already noted [16], and cannot help in estimating the quantity of non-colinear RNAs in a transcriptome. In comparison, CRAC's output is a practical size and allows an in-depth, context-dependent investigation for promising candidates for validation.

In CRAC's output, intragenic and intergenic chRNAs account for 58% and 42% respectively, and are partitioned into five subclasses (Methods, Additional file 5). Looking at the intersection, TopHat-fusion also outputs 76% (346) of the chRNAs found by CRAC, therefore providing additional evidence in favor of their existence, since the presence of some supporting read pairs is a mandatory criterion in TopHat-fusion [16] (Additional file 5).

When compared with the set of validated chimeras of Edgren et al. [27], TopHat-fusion and CRAC detected 21 and 20 out of 27, and agreed on 17 of them (Table 5).[2] The first 20 cases were found by CRAC, and the 7 remaining ones were not predicted by CRAC; however, for the final 2, we could not detect any read matching the 15 to 20 nt over the junction. Among the seven cases CRAC misses, only one (BCAS4-BCAS3) is a false negative, four are uncertain with not enough expressed candidates (CPNE1-P13, STARD3-DOK5, WDR67-ZNF704, and PPP1R12A-SEPT10), and no read seems to match the junction of the two remaining ones (DHX35-ITCH and NFS1-PREX1). As the BCAS4-BCAS3 junction includes a substitution near the splice site, the reads carry two events (SNV plus junction): CRAC does not exactly position the junction and outputs them in the BioUndetermined file, whose exploration could extract BCAS4-BCAS3 as a candidate (future work). For the four uncertain cases, the *k*-mer support over the junction break equals one, meaning that only one read matches the junction exactly; hence CRAC identifies a chimeric junction, but classifies them as uncertain candidates

Table 4 Chimeric RNA detection in breast cancer libraries

Edgren libraries	CRAC				TopHat-fusion			
	Raw		After GSNAP		Raw		After GSNAP	
	Number of chRNAs	Number of reads	Number of chRNAs	Number of reads	Number of chRNAs	Number of reads	Number of chRNAs	Number of reads
BT-474	692	9,661	153	460	109,711	349,801	81,327	189,523
KPL-4	407	5,157	60	199	32,412	98,330	23,075	53,165
MCF-7	466	3,475	90	180	42,738	121,544	27,267	57,676
SK-BR-3	703	9,354	152	577	86,249	241,219	61,494	130,682

TopHat-fusion reports approximately 200 times more raw candidates than CRAC; this ratio increases after filtering. Comparison with the set of validated chRNAs by Edgren et al. [27] shows that both the filtered and unfiltered predictions of CRAC and TopHat-fusion include respectively 20 and 21 true chRNAs and they agree for 17 of them.

(Undetermined file). Three out of four are nevertheless detected by TopHat-fusion, but with two or one spanning reads (2,1,1) and few supporting pairs (6,5,0), thereby corroborating CRAC's view and confirming these are expressed at very low levels in this dataset.

Considering validated intergenic chRNAs [27], the sensitivity over the 27 valid chRNAs is comparable between TopHat-fusion (77% = 21/27) and CRAC (74% = 20/27), while the precision over the total number of candidates is markedly in favor of CRAC ($21/143,003 \simeq 0.01\%$ vs $20/192 \simeq 10.4\%$;[3] Table 5, Additional file 5). Clearly, some experimentally validated chRNAs (like DHX35-ITCH or NFS1-PREX1), happen to have no read spanning their junction, and thus should not be computationally predicted as candidates on the basis of this read data. This important statement illustrates how difficult computational chRNA prediction is, thereby emphasizing the quality of CRAC's analysis. Moreover, the evidence suggests that other promising candidate chRNAs populate CRAC's results.

Numerous chRNAs are predicted in classes 3/5, where the RNA non-colinearity appears as an inversion. CRAC detects three such chRNAs within the MAN1A2 gene, which recur in up to three out of four breast cancer libraries, and in a K562 library. These specific inversions in MAN1A2 are described as post-transcriptional exon-shuffling RNAs and found highly expressed in several acute lymphoblastic leukemia samples [28]. Our results support the existence of such mRNA-exhibiting shuffled

Table 5 CRAC and TopHat-fusion predictions for the set of validated chimeric junctions from breast cancer libraries

Library	Fused genes	Chromosomes	5′ position	5′ strand	3′ position	3′ strand	Average support[a]	CRAC[b]	TopHat-fusion[c]
BT-474	SNF8-RPS6KB1	17-17	47,021,337	1	57,970,686	-1	36	Yes	Yes
BT-474	CMTM7-GLB1	3-3	32,483,329	-1	33,055,545	1	2	Yes	Yes
BT-474	SKA2-MYO19	17-17	57,232,490	-1	34,863,351	-1	6	Yes	Yes
BT-474	ZMYND8-CEP250	20-20	45,852,968	-1	34,078,459	1	9	Yes	Yes
BT-474	VAPB-IKZF3	20-17	56,964,572	1	37,934,021	-1	6	Yes	Yes
BT-474	ACACA-STAC2	17-17	35,479,452	-1	37,374,427	-1	46	Yes	Yes
BT-474	DIDO1-TTI1	20-20	61569147	-1	36,634,800	-1	2	Yes	Yes
BT-474	RAB22A-MYO9B	20-19	56,886,178	1	17,256,205	1	9	Yes	Yes
BT-474	MCF2L-LAMP1	13-13	11,371,8616	-1	113,951,811	-1	2	Yes	No
KPL-4	NOTCH1-NUP214	9-9	139,438,475	-1	134,062,675	1	2	Yes	Yes
KPL-4	BSG-NFIX	19-19	580,782	1	13,135,832	1	9	Yes	Yes
MCF-7	RPS6KB1-TMEM49	17-17	57,992,064	1	57,917,126	1	5	Yes	Yes
MCF-7	ARFGEF2-SULF2	20-20	47,538,548	1	46,365,686	-1	10	Yes	Yes
SK-BR-3	PKIA-RARA	8-17	79,485,042	-1	38,465,537	-1	7	Yes	Yes
SK-BR-3	TATDN1-GSDMB	8-17	125,551,264	-1	38,066,177	-1	334	Yes	Yes
SK-BR-3	KCNB1-CSE1L	20-20	47,956,856	-1	47,688,990	-1	6	Yes	No
SK-BR-3	CYTH1-EIF3H	17-8	76,778,283	-1	117,768,258	-1	11	Yes	Yes
SK-BR-3	SUMF1-LRRFIP2	3-3	4,418,012	-1	37,170,640	-1	4	Yes	Yes
SK-BR-3	SETD3-CCDC85C	14-14	99,880,273	1	100,002,353	1	3	Yes	No
SK-BR-3	PCDH1-ANKHD1	5-5	141,234,002	1	139,825,559	-1	2	Yes	Yes
BT-474	CPNE1-P13	20-20	34,243,123	NA	43,804,501	NA	1	No	Yes
BT-474	STARD3-DOK5	17-17	37,793,479	NA	53,259,992	NA	1	No	Yes
SK-BR-3	WDR67-ZNF704	8-8	124,096,577	NA	81,733,851	NA	1	No	Yes
MCF-7	BCAS4-BCAS3	20-17	49,411,707	NA	59,445,685	NA	3	No	Yes
KPL-4	PPP1R12A-SEPT10	12-2	80,211,173	NA	11,034,3414	NA	1	No	No
SK-BR-3	DHX35-ITCH	20-20	Unknown	NA	Unknown	NA	NA	No	No
SK-BR-3	NFS1-PREX1	20-20	Unknown	NA	Unknown	NA	NA	No	No

NA: not applicable

[a] Average support value over the junction *k*-mers

[b] Detected by CRAC

[c] Detected by TopHat-fusion

CRAC and TopHat-fusion predictions on the set of validated chimeric junctions from four breast cancer libraries [27]. The first 20 cases were found by CRAC, and the 7 remaining ones were not predicted by CRAC; however, for the final 2, we could not detect any read matching the 15 to 20 nt over the junction. A short read length penalizes CRAC: indeed, with $k = 22$, only the 6 (= 50 - 2 × 22) middle positions of a read could be used to locate any event (splices or mutations) exactly. Hence we expect that the spanning reads by which a chRNA is amenable to detection by CRAC to be rare. NA: not applicable. Columns: library, fused genes ID, annotation of the junction points, chromosomes, 5′ position and strand, 3′ position and strand, average support value over the junction k-mers, detection by CRAC and by TopHat-fusion (THF).

exons, as well as cases where the inversion is short, sometimes inducing a repeat within the read (see an example in the LONP1 gene given in Additional file 4).

Notably, among 455 chRNAs, CRAC reports 36 chRNAs that appear to recur in two, three, or even all four breast cancer libraries (Additional file 5). Among these 36 chRNAs: 24 are intra- and 12 are inter-chromosomal, 20 are intragenic, while 16 fuse different genes. Moreover, 35 out of 36 (including the MAN1A2 and LONP1 cases) harbor exactly the same junction point in all libraries in which they were detected. Previous investigations of these libraries [16,27] did not report any recurrent chRNAs. However, when we ran TopHat-fusion, it also output 23 of these chRNAs among 193,163 candidates.

For instance, we found a HSPD1-PNPLA4 chRNA in both KPL-4 and SK-BR-3 libraries: PNPLA4 (GS2) is highly expressed in human SW872 liposarcoma cells [29], while HSPD1, the heat shock protein Hsp60, shows a broad antiapoptotic function in cancer [30]. Among the intragenic chRNAs, we observed in all four libraries a non-colinear chRNA within GNAS, a gene coding for the G-protein alpha subunit, which is known to be associated with multiple human diseases including some cancers [31], and was recently found to be recurrently mutated in cystic pancreatic lesions related to invasive adenocarcinomas [32], as well as amplified in breast cancers [33]. Moreover, we also found the same CTDSPL2-HNRNPM chimeric RNA in the BT-474, MCF-7, and SK-BR-3 libraries. Both genes belong to the heterogeneous nuclear ribonucleoprotein family and play a pivotal role in pre-mRNA processing. Importantly, HNRNPM regulates the alternative splicing of carcinoembryonic antigen-related cell adhesion molecule-1 (CEACAM1) in breast cancer cells [34].

Discussion

CRAC is a multi-purpose tool for analyzing RNA-seq data. In a single run it can predict sequencing errors, small mutations, and normal and chimeric splice junctions (collectively termed events). CRAC is not a pipeline, but a single program that can replace a combination of Bowtie, SAMtools, and TopHat/TopHat-fusion, and can be viewed as an effort to simplify NGS analysis. CRAC is not simply a mapper, since it uses local coverage information (in the support profile) before computing the genomic position of a read. In contrast to the current paradigm, mapping and post inferences are not disjoint steps in CRAC. Instead, it implements a novel, integrated approach that draws inferences by simultaneously analyzing both the genomic locations and the support of all *k*-mers along the read. The support of a *k*-mer, defined as the number of reads sharing it, approximates the local read coverage without having the reads mapped. The combined *k*-mers location and support profiles enable CRAC to infer precisely the read and genomic positions of an event, its structure, as well as to distinguish errors from biological events. Integration is not only the key to an accurate classification of reads (Additional file 1), but it avoids information loss and saves re-computation, and is thereby crucial for efficiency. Indeed, CRAC takes more time than state-of-the-art mappers, but is considerably faster than splice junction prediction tools (for example, Bowtie plus TopHat). The other key to efficiency is the double-indexing strategy: a classical FM-index (where FM stands for Ferragina - Manzini) for the genome and the Gk arrays for the reads [21]. This makes CRAC's memory requirement higher than that of other tools, but fortunately computers equipped with 64 gigabytes of memory are widespread nowadays. Experiments conducted on simulated data (where all answers are known), which are necessary for assessing a method's sensitivity, have shown that for each type of prediction CRAC is at least competitive or surpasses current tools in terms of sensitivity, while it generally achieves better precision. Moreover, CRAC's performances further improve when processing longer reads: for example on 200 nt reads, it has 85% sensitivity and 99.3% precision for predicting splice junctions.

CRAC analyzes how the location and support profiles vary and concord along the read. Hence *k*-mers serve as seeds (in the genome and in the read set), and *k* is thus a key parameter. Its choice depends on the genome length [19], and quite conservative values - $k = 22$ for the human genome - have been used in our experiments. Smaller *k* values are possible with smaller genomes (like bacterial ones). *k* affects the number of false genomic locations (FLs) that occur in the profile; a FL indicates a wrong location for a *k*-mer, which differs from the location of origin of the sequenced molecule. This tends to induce a false location for the read (mapping) or a false location for a junction border (normal and chimeric junction prediction). However, CRAC uses two criteria to avoid these pitfalls: the coherence of locations for adjacent *k*-mers over a range and the concordance of locations for the *k*-mers around the break (especially in the break verification and fusion procedures; see Additional File 2). When *k*-mers surrounding the break have a few, but several, locations, CRAC examines all possible combinations, and as FL occurrences are governed mainly by randomness, this eliminates discordant positions. FLs have a larger effect on the prediction of chimeras. Overall, the results on both simulated and real data, like the improved mapping sensitivity (+15 points compared to Bowtie, BWA, and SOAP2), show that CRAC makes accurate predictions with conservative values. *k* controls the balance between sensitivity (shorter seeds) and precision. The breast cancer libraries we used

have 50 nt reads, but CRAC could still find 74% of the chimeric RNAs validated by Edgren et al. [27]. Of course, the *k* value has two limitations: first, the minimal exon size detectable in a read is ≥*k*, second, reads must be long enough (>40 nt with *k* = 20 for the human genome). However, NGS is progressing towards longer reads, which should become standard, and Figure 4c illustrates well CRAC's ability to detect short exons within single reads. The *k*-mer profiling approach detects events located near the read extremities, but cannot exactly determine their position in the read. Hence the inference rules cannot be fully applied, and CRAC classifies such reads as incompletely determined (Undetermined and BioUndetermined files). However, the position of an event in a read is random, and thus, the high coverage delivered by NGS nearly ensures that the same event occurs in the middle of other reads covering it. Consequently, *border cases* do not hinder CRAC from detecting mutations, splice junctions, etc. Only errors escape this rule since they are mostly read specific. A more complex drawback of *k*-mer profiling is when two events are located <*k* positions apart on the genome (see the BCAS4-BCAS3 chimera); again such cases even with a high support are not fully resolved and end up in the BioUndetermined file. A post-processing of reads in this file, for example by an alignment program, could clearly save such cases. Obviously, such cases are rare, and we keep this as future work. As briefly mentioned, *k*-mer profiling also detects when reads span a repeat border region, which should help in inferring the locations of mobile genetic elements, duplications, or copy number variations; this suggests further developments and CRAC's usefulness for analyzing genomic data.

Determining the correct genomic location of reads is crucial information for any NGS data analysis and especially for cataloging all transcripts of a cell with RNA-seq. Generally, a mapping step computes this information using efficient, well-known tools (BWA, Bowtie, and SOAP2), but the mapping sensitivity is rarely questioned. We performed extensive mapping tests on simulated data, which showed that sensitivity can truly be improved and that CRAC makes a significant step in this direction. Of course by considering discontinuous alignments (as do CRAC and GSNAP) many reads covering splice junctions can be mapped, which BWA, Bowtie/Bowtie2, and SOAP2 cannot detect. However, the mapping results for categories of reads carrying one mutation, a short indel, or even errors indicate that classical mappers missed between 15 to 20 points in sensitivity, thereby confirming that the difference due to splice junction reads is critical even for other events, while CRAC performs equally well (>90%) whatever the category (Figure 2). The other way around, those tools are able to map 10% to 35% of reads containing a splice junction. This can negatively affect downstream analyses depending on the type of events under investigation. For instance to predict splice junctions, in the current strategy (TopHat, MapSplice, or TopHat-fusion), reads are first mapped with Bowtie to divide the collection into: (a) reads having a continuous alignment on the genome and (b) unmapped reads. The former serve further to delimit exons, and the latter are then processed again to search for spliced alignments. If a read that requires a discontinuous alignment is mapped by Bowtie, it is not considered by TopHat, MapSplice, or TopHat-fusion as potentially containing a junction, and they will not find a spliced alignment for it. In contrast, CRAC's *k*-mer profiling approach is flexible, reliable in this respect (Figure 3), and importantly, adapts well to longer reads (for example, 200 nt). This last point is key since longer reads will be available soon. They will much more likely incorporate not one, but several events - errors, mutations, splice junctions, etc. - and thus be harder to map. Whatever the class of required predictions, CRAC's sensitivity is always improved with longer reads. This is crucial for detecting multiple exons within single reads, and CRAC exhibits a better ability in this as exemplified by a transcript of TIMM50 gene (Figure 4c).

An issue in transcriptomics is to reliably extract the complete set of splice junctions with a minimal number of false positives [24]. In this regard, our results (Table 2) demonstrate that *k*-mer profiling approaches (MapSplice and CRAC) profit greatly in sensitivity from longer reads, and that CRAC is the tool with the highest precision whatever the read length. They also indicate that CRAC handles difficult cases with higher sensitivity, like long-distance splices, multi-exon reads, or RNA expressed at a low level. The analysis of a multi-tissue library shows that CRAC, GSNAP, and MapSplice have a very large (>90%) agreement on the set of reported known junctions (>140,000 distinct junctions), RefSeq transcripts, and genes, thereby providing evidence of their ability to extract splice junctions of well-annotated transcripts (Figure 4b and 4a). In contrast, TopHat misses 21% of these known RefSeq junctions. Comparatively, CRAC reports fewer novel or unknown junctions than other tools, and tends to be more conservative, which likely reflects its precision. Altogether, CRAC is a solution for exploring qualitatively the transcriptome of a sample with high sensitivity and precision, and thus provides the primary material for determining all transcript structures, which is indispensable for estimating the expression levels of all RNA isoforms [3,26].

Recent investigations have suggested that non-colinear RNAs are quantitatively more abundant in human transcriptomes than previously thought, underlining the structural diversity of these chimeric RNAs and their occurrence in cancers [8,27,28,35,36]. Predicting chimeric RNAs (chRNAs) is the most difficult and error-prone

computation when analyzing RNA-seq. The combinatorial possibilities of aligning a read partly to two distinct regions on the same or different chromosomes [4] increase the likeliness of predicting FPs. It explains why filtering for suboptimal but colinear alignments of an apparent chimeric read may still help, and also partly why TopHat-fusion per se yields so many more chRNA candidates compared to CRAC (Table 4). Paired end reads are not sufficient: analyzing single reads by splitting them is inevitable for predicting the chimeric junction point; hence *k*-mer profiling also suits this purpose. Nevertheless, paired end reads are useful for performing a complementary consolidation of chRNA candidates, which we may develop in the future. However, chRNAs can occur at low expression levels and be much less expressed than their parental genes; this impels CRAC to rely less on the support profile than for mutation prediction. In addition, transcriptional noise or template switching during library preparation may generate true chimeric reads from biologically irrelevant chRNAs. Thus, subsequent criteria are definitely needed to prioritize chRNA candidates: the consistent finding of the same junction point has been suggested as an important one [27,36,37]. Notably, CRAC predicted for the four breast cancer libraries 36 recurrent chRNAs that were not reported previously [16,27], and 35/36 always harbor the same junction point in the different libraries and among the distinct reads predicting them. Several of these involve genes known to be implicated in tumorigenesis or tumor maintenance, like GNAS [31] or HSPD1 [30]. As CRAC outputs also included 74% of validated chRNAs with a single clear false negative, this shows that CRAC consistently reports interesting chRNA candidates based on the read data. As already mentioned, CRAC distinguishes between five chRNA classes, included those exhibiting small-scale sequence inversions, as illustrated by a chRNA within the LONP1 gene, which recurs in normal and tumoral libraries. We also reported cases of chRNAs, which although validated, do not constitute good candidates *for the computational inference step*, since not enough reads in the data support their existence. The latter point is critical and strengthens how difficult chimeric RNA prediction is.

Here, the *in silico* experiments focus on transcriptomic data, but the method is also applicable to genomic sequencing. For instance, the counterparts of splice junctions and chimeras in RNA-seq are large deletions and rearrangements (translocation, inversion, and displacement of a mobile element) in DNA. Thus, CRAC may also prove useful for genomic analyses.

Endnotes

[a] TopHat-fusion without the extra post-filtering script.

[b] If TopHat-fusion-post is applied to TopHat-fusion's results with default parameters, it reports 27 chimera, 11 of them being validated chimeras, which is about half those reported by TopHat-fusion alone.

[c] Only intergenic chRNAs are counted here.

Additional material

Additional file 1: Figure with read classification performed by CRAC.

Additional file 2: Additional description of the CRAC algorithm, the simulation of RNA-seq data, and the tools used for comparison.

Additional file 3: Results for simulated RNA-seq data.

Additional file 4: Results for real RNA-seq data.

Additional file 5: Table with chimeric RNAs predicted for four breast cancer libraries.

List of abbreviations

chRNA: chimeric RNA; EST: expressed sequence tag; FL: false location; indel: insertion or deletion; FP: false positive; KJ: known splice junction; NGS: next generation sequencing; NJ: new splice junction; nt: nucleotide; OJ: other splice junction; RNA-seq: RNA sequencing; SJ: splice junction; SNP: single nucleotide polymorphism; SNV: single nucleotide variant; TP: true positive.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

NP, MS and ER devised the algorithm. NP and MS developed the source code. All authors devised and analyzed the software comparisons and evaluations, as well as the analysis of real datasets. NP and MS performed the experiments. NP and MS prepared the figures. ER wrote the manuscript with contributions from all authors. ER and TC supervised the research. All authors read and approved the final manuscript.

Acknowledgements

The authors thank Alban Mancheron for help in packaging the CRAC software, Gwenael Piganeau for critically reading this manuscript. NP is supported by Fondation ARC pour la Recherche sur le Cancer (grant PDF20101202345), Ligue Contre le Cancer (grant JG/VP 8102). NP, TC, and ER are supported by a CNRS INS2I (grant PEPS BFC: 66293), the Institute of Computational Biology, Investissement d'Avenir. TC and ER acknowledge the support from the Region Languedoc Roussillon (grant Chercheur d'Avenir, grant GEPETOS). MS is partially supported by the French ANR-2010-COSI-004 MAPPI Project. TC is supported by the Ligue régionale contre le cancer and the University of Montpellier 2. We acknowledge funding from Agence Nationale de la Recherche (grant Colib'read ANR-12-BS02-008), from the NUMEV Labex, and from the CNRS Mastodons Program. All experiments were run on the ATGC bioinformatics platform [38].

Authors' details

[1]Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM), UMR 5506, CNRS and Université de Montpellier 2, 161 rue Ada, 34095 Montpellier Cedex 5, France. [2]Institut de Recherche en Biothérapie (IRB), U1040 INSERM, CHRU Montpellier Hôpital Saint-Eloi 80, av. Augustin Fliche, 34295 Montpellier Cedex 5, France. [3]Laboratoire d'Informatique Fondamentale de Lille (LIFL), (UMR CNRS 8022, Université Lille 1) and Inria Lille-Nord Europe, Cité scientifique-Bâtiment M3, 59655 Villeneuve d'Ascq Cedex, France. [4]LITIS EA 4108, Université de Rouen, 1 rue Thomas Becket, 76821 Mont-Saint-Aignan Cedex, France. [5]Institut de Biologie Computationnelle, 95 Rue de la Galéra, 34095 Montpellier Cedex 5, France.

Received: 18 November 2012 Revised: 28 February 2013
Accepted: 28 March 2013 Published: 28 March 2013

References

1. Ozsolak F, Milos PM: **RNA sequencing: advances, challenges and opportunities.** *Nat Rev Genet* 2011, **12**:87-98.

2. Meyerson M, Gabriel S, Getz G: **Advances in understanding cancer genomes through second-generation sequencing.** *Nat Rev Genet* 2010, **11**:685-696.
3. Trapnell C, Pachter L, Salzberg S: **TopHat: discovering splice junctions with RNA-seq.** *Bioinformatics* 2009, **25**:1105-1111.
4. Gingeras T: **Implications of chimaeric non-co-linear transcripts.** *Nature* 2009, **461**:206-211.
5. Mitelman F, Johansson B, Mertens F: **Mitelman database of chromosome aberrations and gene fusions in cancer.** 2013 [http://cgap.nci.nih.gov/Chromosomes/Mitelman].
6. Li H, Wang J, Mor G, Sklar J: **A neoplastic gene fusion mimics trans-splicing of RNAs in normal human cells.** *Science* 2008, **321**:1357-1361.
7. Rabbitts TH: **Commonality but diversity in cancer gene fusions.** *Cell* 2009, **137**:391-395.
8. Kannan K, Wang L, Wang J, Ittmann MM, Li W, Yen L: **Recurrent chimeric RNAs enriched in human prostate cancer identified by deep sequencing.** *Proc Natl Acad Sci* 2011, **108**:9172-9177.
9. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25**:1754-1760.
10. Li H, Durbin R: **Fast and accurate long-read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2010, **26**:589-595.
11. Li R, Li Y, Kristiansen K, Wang J: **SOAP: short oligonucleotide alignment program.** *Bioinformatics* 2008, **24**:713-714.
12. Langmead B, Trapnell C, Pop M, Salzberg S: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** *Genome Biol* 2009, **10**:R25.
13. Rizk G, Lavenier D: **GASSST: global alignment short sequence search tool.** *Bioinformatics* 2010, **26**:2534-2540.
14. Wu TD, Nacu S: **Fast and SNP-tolerant detection of complex variants and splicing in short reads.** *Bioinformatics* 2010, **26**:873-881.
15. Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, He X, Mieczkowski P, Grimm SA, Perou CM, MacLeod JN, Chiang DY, Prins JF, Liu J: **MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery.** *Nucleic Acids Res* 2010, **38**:e178.
16. Kim D, Salzberg SL: **TopHat-Fusion: an algorithm for discovery of novel fusion transcripts.** *Genome Biol* 2011, **12**:R72.
17. **CRAC project website..** [http://crac.gforge.inria.fr/].
18. **ATGC Bioinformatics Platform, CRAC.** [http://www.atgc-montpellier.fr/crac].
19. Philippe N, Boureux A, Tarhio J, Bréhélin L, Commes T, Rivals E: **Using reads to annotate the genome: influence of length, background distribution, and sequence errors on prediction capacity.** *Nucleic Acids Res* 2009, **37**:e104.
20. Ferragina P, Manzini G: **Opportunistic data structures with applications.** *Proceedings of FOCS* 2000, 390-398.
21. Philippe N, Salson M, Lecroq T, Leonard M, Commes T, Rivals E: **Querying large read collections in main memory: a versatile data structure.** *BMC Bioinf* 2011, **12**:242.
22. Griebel T, Zacher B, Ribeca P, Raineri E, Lacroix V, Guigó R, Sammeth M: **Modelling and simulating generic RNA-seq experiments with the flux simulator.** *Nucleic Acids Res* 2012, **40**:10073-10083.
23. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nat Methods* 2012, **9**:357-359.
24. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB: **Alternative isoform regulation in human tissue transcriptomes.** *Nature* 2008, **456**:470-476.
25. Brawand D, Soumillon M, Necsulea A, Julien P, Csardi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M, Albert FW, Zeller U, Khaitovich P, Grätzner F, Bergmann S, Nielsen R, Paabo S, Kaessmann H: **The evolution of gene expression levels in mammalian organs.** *Nature* 2011, **478**:343-348.
26. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: **Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation.** *Nat Biotech* 2010, **28**:511-515.
27. Edgren H, Murumagi A, Kangaspeska S, Nicorici D, Hongisto V, Kleivi K, Rye IH, Nyberg S, Wolf M, Borresen-Dale A, Kallioniemi O: **Identification of fusion genes in breast cancer by paired-end RNA-sequencing.** *Genome Biol* 2011, **12**:R6.
28. Al-Balool HH, Weber D, Liu Y, Wade M, Guleria K, Nam PLP, Clayton J, Rowe W, Coxhead J, Irving J, Elliott DJ, Hall AG, Santibanez-Koref M, Jackson MS: **Post-transcriptional exon shuffling events in humans can be evolutionarily conserved and abundant.** *Genome Res* 2011, **21**:1788-1799.
29. Jenkins CM, Mancuso DJ, Yan W, Sims HF, Gibson B, Gross RW: **Identification, cloning, expression, and purification of three novel human calcium-independent phospholipase A2 family members possessing triacylglycerol lipase and acylglycerol transacylase activities.** *J Biol Chem* 2004, **279**:48968-48975.
30. Ghosh JC, Siegelin MD, Dohi T, Altieri DC: **Heat shock protein 60 regulation of the mitochondrial permeability transition pore in tumor cells.** *Cancer Research* 2010, **70**:8988-8993.
31. Wood LD, Parsons DW, Jones S, Lin J, Sjoblom T, Leary RJ, Shen D, Boca SM, Barber T, Ptak J, Silliman N, Szabo S, Dezso Z, Ustyanksky V, Nikolskaya T, Nikolsky Y, Karchin R, Wilson PA, Kaminker JS, Zhang Z, Croshaw R, Willis J, Dawson D, Shipitsin M, Willson JKV, Sukumar S, Polyak K, Park BH, Pethiyagoda CL, Pant PVK, *et al*: **The genomic landscapes of human breast and colorectal cancers.** *Science* 2007, **318**:1108-1113.
32. Wu J, Matthaei H, Maitra A, Molin MD, Wood LD, Eshleman JR, Goggins M, Canto MI, Schulick RD, Edil BH, Wolfgang CL, Klein AP, Diaz LA, Allen PJ, Schmidt CM, Kinzler KW, Papadopoulos N, Hruban RH, Vogelstein B: **Recurrent GNAS mutations define an unexpected pathway for pancreatic cyst development.** *Sci Trans Med* 2011, **3**:92ra66.
33. Kan Z, Jaiswal BS, Stinson J, Janakiraman V, Bhatt D, Stern HM, Yue P, Haverty PM, Bourgon R, Zheng J, Moorhead M, Chaudhuri S, Tomsho LP, Peters BA, Pujara K, Cordes S, Davis DP, Carlton VEH, Yuan W, Li L, Wang W, Eigenbrot C, Kaminker JS, Eberhard DA, Waring P, Schuster SC, Modrusan Z, Zhang Z, Stokoe D, de Sauvage FJ, Faham M, *et al*: **Diverse somatic mutation patterns and pathway alterations in human cancers.** *Nature* 2010, **466**:869-873.
34. Dery KJ, Gaur S, Gencheva M, Yen Y, Shively JE, Gaur RK: **Mechanistic control of carcinoembryonic antigen-related cell adhesion molecule-1 (CEACAM1) splice isoforms by the heterogeneous nuclear ribonuclear proteins hnRNP L, hnRNP A1, and hnRNP M.** *J Biol Chem* 2011, **286**:16039-16051.
35. Maher CA, Palanisamy N, Brenner JC, Cao X, Kalyana-Sundaram S, Luo S, Khrebtukova I, Barrette TR, Grasso C, Yu J, Lonigro RJ, Schroth G, Kumar-Sinha C, Chinnaiyan AM: **Chimeric transcript discovery by paired-end transcriptome sequencing.** *Proc Natl Acad Sci* 2009, **106**:12353-12358.
36. Maher C, Kumar-Sinha C, Cao X, Kalyana-Sundaram S, Han B, Jing X, Sam L, Barrette T, Palanisamy N, Chinnaiyan A: **Transcriptome sequencing to detect gene fusions in cancer.** *Nature* 2009, **458**:97-101.
37. Houseley J, Tollervey D: **Apparent non-canonical trans-splicing is generated by reverse transcriptase *in vitro*.** *PLoS ONE* 2010, **5**:e12271.
38. **ATGC Bioinformatics Platform, Next generation Sequencing.** [http://www.atgc-montpellier.fr/ngs].

doi:10.1186/gb-2013-14-3-r30
Cite this article as: Philippe *et al.*: CRAC: an integrated approach to the analysis of RNA-seq reads. *Genome Biology* 2013 **14**:R30.

CRAC: An integrated approach to analyse RNA-seq reads

Additional File 1

Overview of the read classification performed by CRAC.

Nicolas Philippe and Mikael Salson and Thérèse Commes and Eric Rivals

February 13, 2013

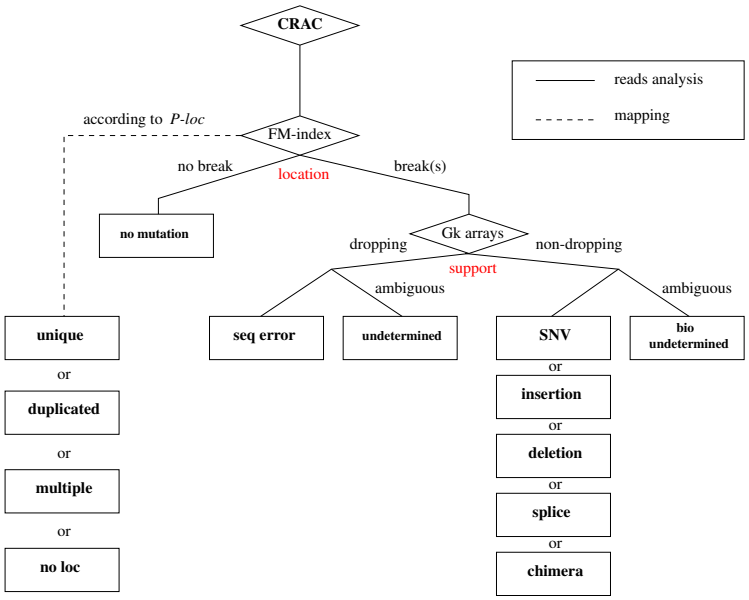


Figure 1: Overview of the read classification performed by CRAC. CRAC processes each read in turn and performs several predictions regarding its genomic locations, sequencing errors, the presence of mutations (SNV, insertion, and deletion), as well as normal or chimeric splice junctions. For each question the read can be assigned to one or several classes. Mapping: depending on the possible genomic locations of its k-mers obtained from the FM-index, CRAC decides whether the read has a unique (**unique**), or either a few (**duplicated**) or many genomic locations (**multiple**). When too many k-mers cannot be located the read is considered as having no location (**no loc**). Break: when the k-mer location profile contains a break, the Gk arrays are interrogated to analyze the support profile and decide whether it is due to a sequencing error or a biological event. In each case, the profiles may still be ambiguous and the read is then classified as not fully determined (**undetermined** or **bio undetermined**). Otherwise, according to the rules that distinguish different events the read is assigned to the relevant categories (**seq error**, **SNV**, **insertion**, **deletion**, **splice** or **chimera**).

CRAC: an integrated approach to the analysis of RNA-seq reads
Additional File 2
Supplementary Methods: CRAC, simulation, and other tools.

Nicolas Philippe and Mikael Salson and Thérèse Commes and Eric Rivals

February 28, 2013

1 CRAC algorithm: details and real conditions

In the description of CRAC's method, we first assumed simplifying conditions; now we explain how the actual procedure deals with real conditions.

Differences with the genome at the read's extremities (border cases). This aspect does not influence CRAC's algorithm; rather, the sequencing depth of NGS data saves border cases. Indeed, the random fragmentation of molecules implies that a sequence difference will be uniformly located within the reads that cover these positions. The coverage ensures that these positions will likely be located in the middle of some other reads. Hence, if a biologicial event is missed in some border case read, it will be detected in other reads if the transcript is not rare. The results on simulated data illustrate well this situation. Reads with a difference near their extremities have a break lacking a start or an end (*i.e.* the location before or after the break cannot be determined): this prevents CRAC from identifying the type of difference and from finding its exact position in the read. As some characteristics of the difference are left undetermined, the read is classified either as "undetermined" or "biologically undetermined" depending on its support profile.

Genomic repeats Many reads are sequenced from repeated genomic regions. This translates in k-mers having multiple genomic locations. However, these locations are globally coherent. If a majority of located k-mers ($> 80\%$) are in this case, CRAC classifies reads as duplicated (≤ 5 matching locations) or repeated (> 5 locations). To apply Rule 1, CRAC needs to check the concordance of locations on each side of the break. When processing entirely or partly duplicated reads (not repeated ones), CRAC searches systematically each combination of locations and privileges coherent and concordant ones to reduce the risk of false inferences.

False locations (FL) Our criterion to set k ensures a low average probability of a random k-mer match on the genome [1], but it does not prevent random match, which we term false locations (FL). Compared to true (unique or multiple) locations, the FL of a k-mer will generally not be coherent with those of neighboring k-mers. It may also alter the break length in an unexpected manner: another criterion of verification (see Rule 1). When a read matches the genome, CRAC considers ranges of k-mers having coherent locations to infer its true genomic position. In case of a break in the location profile, CRAC faces two difficulties. First, when a FL happens at a border of a break, it may lead to an incorrect border. When a FL occurs inside a break, it makes up adjacent false breaks, termed *mirage breaks*. In both cases, if the FL is considered true, as it likely disagrees with the location of the other break border, it leads to avoid Rule 1, apply Rule 2, and to predict a false chimeric read. To handle FL at a break border, CRAC uses when necessary a *break verification* procedure, which checks the coherence of locations in the range $[j_b - \delta, j_b]$ (resp. $[j_a, j_a + \delta]$), the concordance of locations across the break, and the break length. This leads to discard the FL and identify the correct borders. Note that verifying the coherence of locations over δ consecutive k-mers is equivalent to considering a single $k+\delta$-mer and therefore diminishes the probability of FL. To detect and remove mirage breaks, CRAC applies a *break fusion* procedure. It checks the concordance of locations across each break, and also across a fused break, *i.e.* looking whether locations before the first break agree with those after the second break. This procedure, which handles duplicated locations by searching for all possible combinations, favors solutions with one or two breaks that avoids predicting chimeric reads.

Multiple differences When a read is affected by several differences (*i.e.* events *e.g.*, a substitution and a splice junction), two cases arise. Either these are k nucleotides apart, then two distinct breaks are found in the location profile (after verification and possible fusion to filter out false locations). Most of the time, k-mers locations across each break and across the two breaks are concordant, and CRAC detects two events. When the differences are too near ($< k$ nucleotides, which is less likely), a single break of unexpected length is found, and it hinders the precise inference of the types and positions of the differences (although the support analysis can still distinguish an error from biological event). Such reads are flagged as having a too large break, hiding several events, and are classified as "undetermined" or "biologically undetermined" depending on the support variation. A notable exception is the case of two substitutions, where we have $L = \ell > k$ and $\ell - k$ gives the distance between the two substitutions.

Rare splice junctions We mentioned that when the support is too low, the read is classified as "undetermined" because we cannot determine whether we are facing a sequencing error or a biological event. The situation is easier for splice junction, for it is unlikely that a sequencing error consists of a large insertion. Hence, when we detect a large insertion (given by the break type) with low support, we can confidently flag the read as being a "weak splice". This particularly highlights the need for integrating support analysis, with mapping and indel detection. Considering at once those informations enables CRAC to detect rare splices in a sample.

2 Simulation of RNA-seq data

To evaluate the sensitivity and precision of mapping programs and of tools for SNV, splice junction, or chimera predictions, we have produced benchmarks of simulated data for various genomes, read lengths, and amounts of sequencing. Using the program FluxSimulator, one can simulate RNA-sequencing from an annotated genome and control, through parameters, the length and number of reads, the amount of sequencing errors, and expression levels of annotated transcripts. However, this incorporates neither genetic mutations (substitutions and indels), nor translocations that make the sampled genome different from the reference genome used for mapping. Translocations are important for they occur in cancer cells and generate chimeric RNAs, which CRAC aims at predicting. To fill this gap, we developed a complementary tool to FluxSimulator called GenomeSimulator, which generates from the input reference genome, a mutated genome that is altered by random mutations and translocations. This program yields a mutated genome with modified annotations and the list of all alterations with their positions compared to the original reference. These files are stored and given to FluxSimulator, which then generates RNA-seq reads. Our system records all biological and erroneous differences compared to the reference, their exact positions on it and in the simulated reads, to allow the verification of mapping and prediction results.

Our goal is to compare several tools on their sensitivity and precision of predictions on RNA-seq data. With RNA-seq data, both events occuring at the genomic (single point mutations, indels, translocations) and at the transcriptomic levels (splicing events) are amenable to detection. Thus, to mirror real conditions our simulation protocol must incorporate the fact that the individual genome whose transcriptome is assayed differs from the reference genome the reads will be aligned to. The simulation procedure comprises two steps (Figure 1): 1/ a **genome alteration step**, in which the reference genome is randomly modified to account for individual genetic differences, 2/ a **RNA-seq simulation step**, where randomly chosen annotated genes are expressed and sequenced to yield reads. In the end, the simulation delivers a read collection and files recording the positions of genomic mutations, splice sites, chimeric junctions, in the reads and on the reference genome. One difficulty is to link the genomic alterations on the simulated reads in terms of positions on the reference genome.

To perform the RNA-seq simulation (step 2) we used the program FluxSimulator [2], which decomposes the simulation in gene expression, library construction and sequencing, and incorportates their systematic biases. However, for simulating an altered genome (step 1), we developed our own program that ensures an easy interconnection with FluxSimulator. Here, we provide a short overview of these computational simulation steps.

2.1 The genome simulation

This procedure modifies the sequence of the input reference genome by introducing random point mutations (SNV), insertions and deletions (or indels), as well as translocations. Substitutions and indels are introduced at random genomic locations at rates chosen by the user. By default, one every $1,000$ nucleotides will be substituted (a rate of 0.1%), while at $1/10,000$ positions, an indel will be introduced (a rate of 0.01%). At a substituted position, the new nucleotide is chosen at

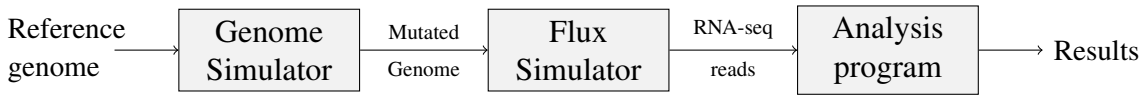


Figure 1: Overview of the simulation and analysis protocol. To generate a RNA-seq dataset, the Reference genome is first mutated by the Genome Simulator (step 1), the modified genome and annotations are input into the RNA-seq simulator (Flux Simulator step 2), which yields a collection of RNA-seq reads. This dataset is then analyzed by any desired tool (step 3). Its results can be confronted to the true errors, mutations, splice sites, chimeras, and its sensitivity and precision evaluated for each of these predictions.

random with equal probability among the three other possibilities. For indels, the length is chosen uniformly within a range $[1,15[$ nucleotides and the inserted sequence chosen randomly. Regarding translocations, whose goal is to generate gene fusions, the exchanged chromosomal locations are chosen within annotated genes. The genome simulator takes a gene annotation file as input to get the positions of all exons. For a translocation, two genes are chosen at random with an input probability, for each gene a breakpoint is chosen at random within its intronic regions, then we perform a bidirectional exchange of the start of one gene with that of the other gene, thereby creating two chimeric, or fusion genes. One starts with the sequence of gene 1 and ends with that of gene 2, and conversely. Exons are never splitted by this process, as we hypothesized that fusion genes with disrupted exons are counterselected by evolution. For the sake of simplicity, the simulator generates chimeras on the forward strand only. RNA-seq reads covering the fused genes will expressed chimeric RNAs; it is worth noting that both fused genes generated by a translocation are, as any other gene, not necessarily expressed, nor covered by reads.

2.2 The RNA sequencing simulation

FluxSimulator provides a RNA-seq simulation that includes all steps impacting the final reads: gene expression, library preparation, and sequencing. Besides the parameters, the input consists in a transcript annotation file in GFF format, which allows FluxSimulator to generate alternatively spliced RNAs for any single gene. We provided FluxSimulator with the RefSeq transcripts from the chosen species as extracted from Ensembl [3]. Another key parameter for testing read analysis tools is the sequencing error model. For substitutions, we used an error model issued from an analysis of the Illumina® sequencing technology with 75 nt reads [4] and we extrapolated this model for 200 nt long reads. In this model, indels are short (in the range $[1,5]$ nt) and their probabilities are much lower than that of substitutions. We also controlled the read length and the read numbers, and asked for single read sequencing only. We set up the parameters to obtain distributions of expression levels similar to that of real experiments [5] (see the RPKM graph for the Human datasets in Figure 2).

Detailed explanations on FluxSimulator parameters are available at http://fluxcapacitor.wikidot.com/simulator.

Sammeth, M., Lacroix, V., Ribeca, P., Guigó, R. The FLUX Simulator. http://flux.sammeth.

Name	Species	Number of reads in million	Read length in nt
hs-75	Human	42	75
hs-200	Human	48	200
dm-75	Drosophila melanogaster	45	75
dm-200	Drosophila melanogaster	48	200

Table 1: Simulated RNA-seq data-sets

net http://flux.sammeth.net/simulator.html

Of course, the simulated RNA-seq reads cover neither the whole sampled genome, nor all genes. Only "expressed" genes will have reads associated with their transcripts. Hence, some genomic mutations produced by the genome simulator will not be covered by any reads in fine. When describing the datasets, we indicate how many events of one categories has been seen by some reads. Intermediate scripts compute the positions and nature of all events (error, SNV, indels, normal and chimeric splice junctions) in the genome and all reads, so as to enable a precise evaluation of all predictions.

3 Simulated RNA-seq datasets

Here, we describe the simulated RNA-seq datasets used for comparing various tools (Table 1), and provide the numbers of alterations generated in each simulated genome, as well as those that are visible in the final RNA-seq reads (Table 2 and 3)

Type	SNV	Insertion	Deletion	Chimera
Genome-wide	3,139,937	287,336	287,502	1,002
Sequenced (75bp)	29,084	2,687	2,734	647
Sequenced (200bp)	52,971	4,810	4,901	914

Table 2: Number of mutations randomly generated in the simulated Human genome, and the numbers among those that were effectively sequenced in the Human simulated RNA-seq datasets from Table 1.

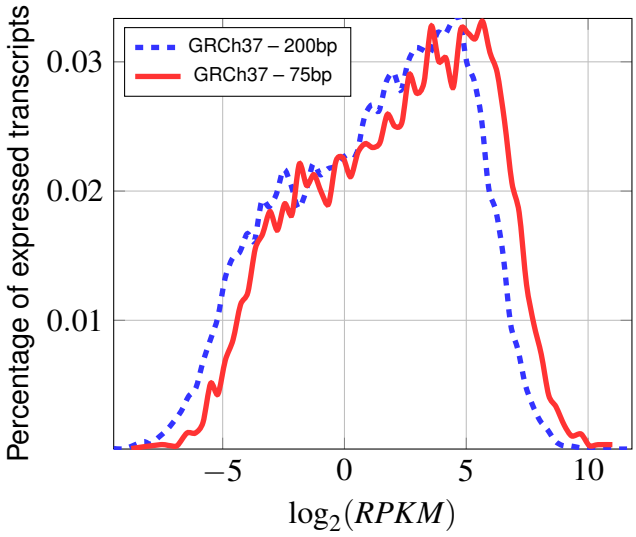
4 Tools used for comparison: version and parameters

We compared CRAC with other tools on its ability to map reads, to identify splice sites, to identify chimeric RNAs. The table 4 lists the software we used, their version, and parameters.

Type	SNV	Insertion	Deletion	Chimera
Genome-wide	132,193	12,146	12,168	676
Sequenced (75bp)	28,397	2,512	2,644	651
Sequenced (200bp)	30,549	2,698	2,861	668

Table 3: Number of mutations randomly generated in the simulated drosophila genome, and numbers of those that were effectively sequenced in the Drosophila simulated RNA-seq datasets from Table 1.

Figure 2: Distribution of $\log_2$(RPKM) for simulated Human data sets: *simulatedHuman75nt-42M* (in red) and *simulatedHuman200nt-48M* (in blue).



5 Score analysis to discriminate between sequence error and biological events

When CRAC analyzes a read, it needs to distinguish sequence errors from biological events using the support profile. Here, we detail the method we employed to derive a discrimination function that performs this distinction in the program. For this sake, we used a machine learning technique, called Support Vector Machines (SVM), which belongs to supervised classification approaches. Two SVMs were trained using simulated data for distinguishing: 1/ errors from mutations, 2/ errors from splice junctions. We explained how this was done in the first case; the same method has been used for the second.

Let us denote by f the discrimination function we want to learn. To do this, we use i/ the *support profile* (described in Algorithm section of the MS - section) to define two variables

Tool	Version	Parameters
Bowtie	0.12.7	–end-to-end –very-sensitive -k 2
Bowtie2	2.0.2	–best -n 3 -y -t -k 2
BWA	0.5.9	
BWA-SW	0.5.9	-b 5 -q2 -r1 -z10
CRAC	1.0.0	-k 22 -m 75
GASSST (short)	1.27	-w 18 -p 94 -s 3 -g 4
GASSST (long)	1.27	-w 18 -p 89 -s 3 -g 9
SOAP2	2.20	
GSNAP	2011-03-28	-N 1 –novel-doublesplices
MapSplice	1.15	–fusion -L 22
TopHat	1.2.0	
TopHat2	2.0.6	–b2-very-sensitive –fusion-search
TopHat-fusion	0.1.0 (BETA)	

Table 4: Parameters used for launching the tools. For CRAC, the `-m` parameter corresponds to the read length and is set according to the read set.

S_{in} and S_{out} of f, ii/ the *simulated dataset* (described Supplementary Section 2) to separate two labelled classes: "errors" and "biological events", iii/ a technique of supervised classification to compute f.

Support profile in a break As explained in the article, CRAC proceeds each read in turn and computes its k-mers *location profile* and k-mer *support profile*. Both a sequence error or a biological mutation (eg, substitution or indel) constitute a difference in sequence between the read and the genome, and consequently, both generate a *break* in the location profile. (see Figure 1a in the MS). To determine whether the source of this event is biological or erroneous, we must focus on the *support profile* (see Figure 1b). We compute two values: the average of the k-mers support outside the break, denoted by S_{out}, and the average of the k-mers support inside the break, denoted by S_{in}. The goal is to compare S_{in} and S_{out} with the following hypothesis: most reads that cover a biological event share the mutation, whereas an error occurs in a small number of reads.

Two labelled classes: "errors" and "biological events" On one hand, from the simulation protocol (supplementary section 2), we know which reads are affected by an error and which reads are affected by a biological event. On the other hand, using CRAC on simulated dataset (supplementary section 3), we can extract reads affected by an event by searching all breaks. As we know the answer for each event, we can define two labelled classes "errors" or "biological events" and save all pairs (S_{out}, S_{in}) for each class (one corresponding to sequence errors coordinates and the other corresponding to biological events coordinates).

Design of the separation To discriminate between the two classes, we use a technique of supervised classification called *Support Vector Machines* (SVM) [6]. It is a learning procedure that tries to find a discrimination function; this function can then be used to assign a new observation to the labelled classes. Here, we want to separate all pairs (S_{out}, S_{in}) for the two labelled classes "errors" and "biological events" according to a function f defined by $f(S_{out}) = S_{in}$. We used a SVM implementation in the R language (the package can be found at http://www.duclert.org/Aide-memoire-R/Apprentissage/SVM.php) with parameters set to:

`mode=CLASSIFY, kernel=POLYNOMIAL, degree=1/3`

to separate the two different vectors. Note that we used a polynomial kernel because the predicted separation was found to be a curve. We have only designed the function on the dataset hs-75 and we used the same for all dataset. In hs-75, we used an error model issued from an analysis of the Illumina® sequencing technology with real 75 nt reads [4] (see supplementary section 2.2). The function f computed by SVM to discriminate between sequence error and biological events (substitution or short indel) is the following:

$$f(S_{out}) = -2.40850 + 2.15859 \times S_{out}^{\frac{1}{3}} + 0.15670 \times S_{out}^{\frac{5}{6}} \quad (1)$$

Because a sequence error affects only a few nucleotides in a read, it can be easily distinguished from a splicing event (junction exon/intron) also characterized by a break. The predicted separation between sequence errors and splice events was found to be a different, less stringent curve. Thus, we decided to learn another specific SVM for splice junctions events with parameters set to (mode=CLASSIFY, kernel=POLYNOMIAL, degree=1/2). The function g computed by SVM to discriminate between sequence error and splicing events is:

$$g(S_{out}) = 0.51081 + 0.16758 \times S_{out}^{\frac{1}{2}} \quad (2)$$

Classification We can define several statistics:

TP: the true positives, *i.e.* the proportion of biological events which are classified correctly

FP: the false positives, *i.e.* the proportion of biological events which are misclassified

TN: the true negatives, *i.e.* the proportion of sequence errors events which are classified correctly

FN: the false negatives, *i.e.* the proportion of sequence errors events which are misclassified

Figure 3 illustrates the distribution of all points (S_{out}, S_{in}) for the two simulated Human datasets hs-75 and hs-200 (described in section 3). The black curves in Figure 3 (A) and Figure 3 (B) show the separations between errors and biological events calculated with SVM. Note that in case of sequence errors, we have a drop in the support profile (section) so it is natural that sequence errors are found under the curve. We visualize the formation of two separate clouds on each side of the curve (in both figures 3): TP (blue dots above the curve) and TN (blue dots below

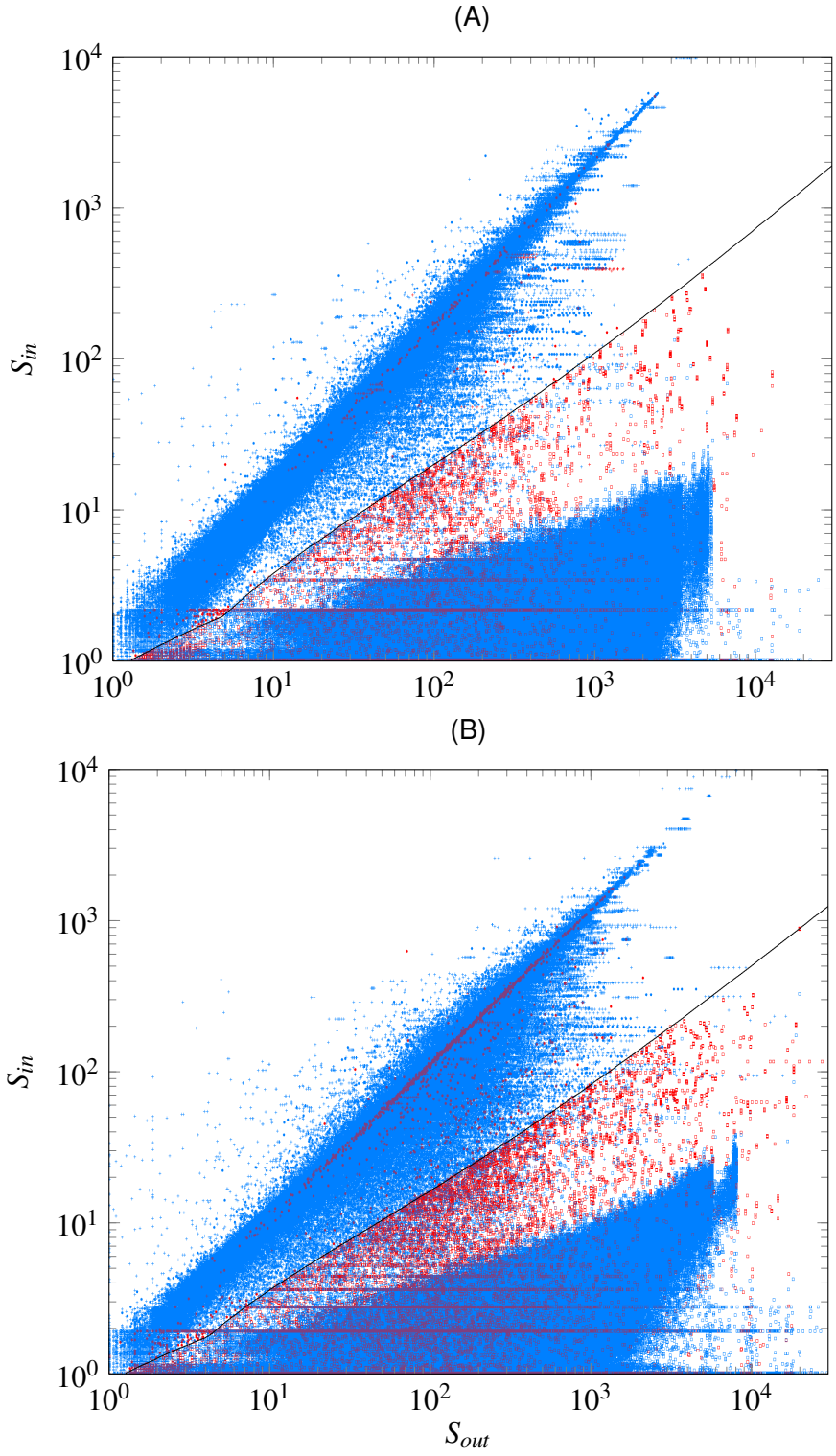


Figure 3: Computation of the separation between sequence errors and biological events for two read datasets: hs-75 (A) et hs-200 (B). Each point illustrates an event which is define in CRAC algorithm by a break. S_{out} is the average of the k-mers support outside the break and S_{in} is the the average of the k-mers support inside the break. The black curve is the separation, and was computed using a SVM approach. The red dots are misclassified points: FP are below the separation and the FN are above the separation. The blue points are well classified points: TP are above the separation and TN are below the separation.

the curve). On the contrary, the red dots represent the events that are misclassified: respectively FP for biological events and the FN for sequence errors. We can observe between the two clouds a mixture of blue dots and red dots that form an ambiguous zone where errors and biological events are indistinguishable.

However, our goal is to increase the precision in "biological events" class, *i.e.* to avoid sequence errors to be detected as biological events. In order to minimize the number of false positives in biological events, we used in the SVM model a variable precision parameter to increase precision for "biological events" class. We set that parameter with a probability of 0.98 even if in return it increases the number of FN for the class "errors".

Score in CRAC We establish a score for each event and for each break, using (S_{in}, S_{out}) . We compute $f(S_{out})$ and the vertical distance relatively to the separation $d = S_{in} - f(S_{out})$. Accordingly, we consider a sequence error when $d \leq 0$ and a biological event when $d > 0$. Indeed, a point which is close to the separation is less likely to be valid than a remote point. For example, a point with a $S1_{out} = 2.6$ and $S1_{in} = 1.5$ has a low score of 0.59 ($S1_{in} - f(S1_{out}) = 1.5 - (-2.40850 + 2.15859 \times 2.6^{\frac{1}{3}} + 0.15670 \times 2.6^{\frac{5}{6}}) = 1.5 - 0.91$). On the contrary, a point with a $S2_{out} = 26$ and $S2_{in} = 25$ will have a score of 23.85.

We see in Table 1 that CRAC has a better sensitivity in hs-200 while reads are longer (section), *i.e.* we find more errors and biological events. We can observe in Figure 3 more true positives in hs-200 than in hs-75 (blue dots above the curve). Because we have increased the variable precision parameter for the biological class, we can see in Figure 3 (A) and (B) that there are more FN than FP. In other words, TP, TN and FP are still the same between hs-75 and hs-200 (blue dots of each side and red dots above the curves) but not FN (red dots below the curve (B)). The results in Table 1 validate our approach because the precision of CRAC remains the same for hs-75 and hs-200.

In conclusion, we defined two functions: i/ one to distinguish sequence errors from point mutations (substitutions, short indels); ii/ another to discriminate splicing events in gray zone from the non-ambiguous zone. We have explained before how we design the function for i/ but the approach is the same for ii/.

6 Partition of chimeric RNAs

CRAC predicts candidate chimeric RNAs. These are splice junctions in which the 5' and 3' "exons" are either 1/ not located one after the other on the same chromosomal strand, these are said to be non colinear, or 2/ are too far apart on the chromosome to belong to the same gene. Obviously, the second case can only be determined by looking at annotations, which we forbid in CRAC. Without annotation, the decision between a splice inside one gene or across two genes is made arbitrarily depending on the distance on the chromosome. For simplicity, we use the term "exon" although the transcribed region may be located outside a known gene, in an intron, in antisense of a gene. By exon, we mean an unspliced part of the RNA.

CRAC further partitions all chRNA in five classes depending on the exon organization; this partition resembles that depicted in [7, Figure 1]. The five classes are as follows:

1. The exons are located on different chromosomes
2. The exons are colinear but (likely) belong to different genes; this must be checked with annotation.
3. The exons are on the same chromosome and same strand, but not in the order in which they are found on DNA, and they do not overlap each other.
4. The exons are on the same chromosome but on different strands.
5. Exactly as in class 3, but the exons overlap each other by at least one nucleotide.

In class 1, the splicing joins pieces from distinct chromosomes, while in classes $2-5$ the exons are on the same chromosome. In summary, class 2 is the only colinear case.

We create class 5 to distinguish cases truely due to large scale inversions (class 3) from those likely due to local inversions or repeats inside genes. When analyzing the breast cancer libraries, we found many such cases.

To investigate more closely these candidates, we confront them to Ensembl annotations [3] and could determine whether the involved "exons" are in annotated exons, introns, or outside genes.

References

[1] Philippe, N., Boureux, A., Tarhio, J., Bréhélin, L., Commes, T., and Rivals, E. Using reads to annotate the genome: influence of length, background distribution, and sequence errors on prediction capacity. *Nucleic Acids Res.* **37**(15), e104 (2009). 2

[2] Griebel, T., Zacher, B., Ribeca, P., Raineri, E., Lacroix, V., Guigó, R., and Sammeth, M. Modelling and simulating generic rna-seq experiments with the flux simulator. *Nucleic Acids Res.* **40**(20), 10073–10083 (2012). 3

[3] Flicek, P., Aken, B. L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., Down, T., Dyer, S. C., Eyre, T., Fitzgerald, S., Fernandez-Banet, J., Graf, S., Haider, S., Hammond, M., Holland, R., Howe, K. L., Howe, K., Johnson, N., Jenkinson, A., Kahari, A., Keefe, D., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Megy, K., Meidl, P., Overduin, B., Parker, A., Pritchard, B., Prlic, A., Rice, S., Rios, D., Schuster, M., Sealy, I., Slater, G., Smedley, D., Spudich, G., Trevanion, S., Vilella, A. J., Vogel, J., White, S., Wood, M., Birney, E., Cox, T., Curwen, V., Durbin, R., Fernandez-Suarez, X. M., Herrero, J., Hubbard, T. J. P., Kasprzyk, A., Proctor, G., Smith, J., Ureta-Vidal, A., and Searle, S. Ensembl 2008. *Nucleic Acids Res.* **36**(S1), D707–714 (2008). 4, 11

[4] Aury, J., Cruaud, C., Barbe, V., Rogier, O., Mangenot, S., Samson, G., Poulain, J., Anthouard, V., Scarpelli, C., Artiguenave, F., and Wincker, P. High quality draft sequences for prokaryotic genomes using a mix of new sequencing technologies. *BMC Genomics* **9**, 603 (2008). PMID: 19087275. 4, 8

[5] Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**(7), 621–628 (2008). 4

[6] Crammer, K. and Singer, Y. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research* **2**, 265–292, December (2001). 8

[7] Gingeras, T. Implications of chimaeric non-co-linear transcripts. *Nature* **461**, 206–11 (2009). 11

CRAC: An integrated approach to analyse RNA-seq reads
Additional File 3
Results on simulated RNA-seq data.

Nicolas Philippe and Mikael Salson and Thérèse Commes and Eric Rivals

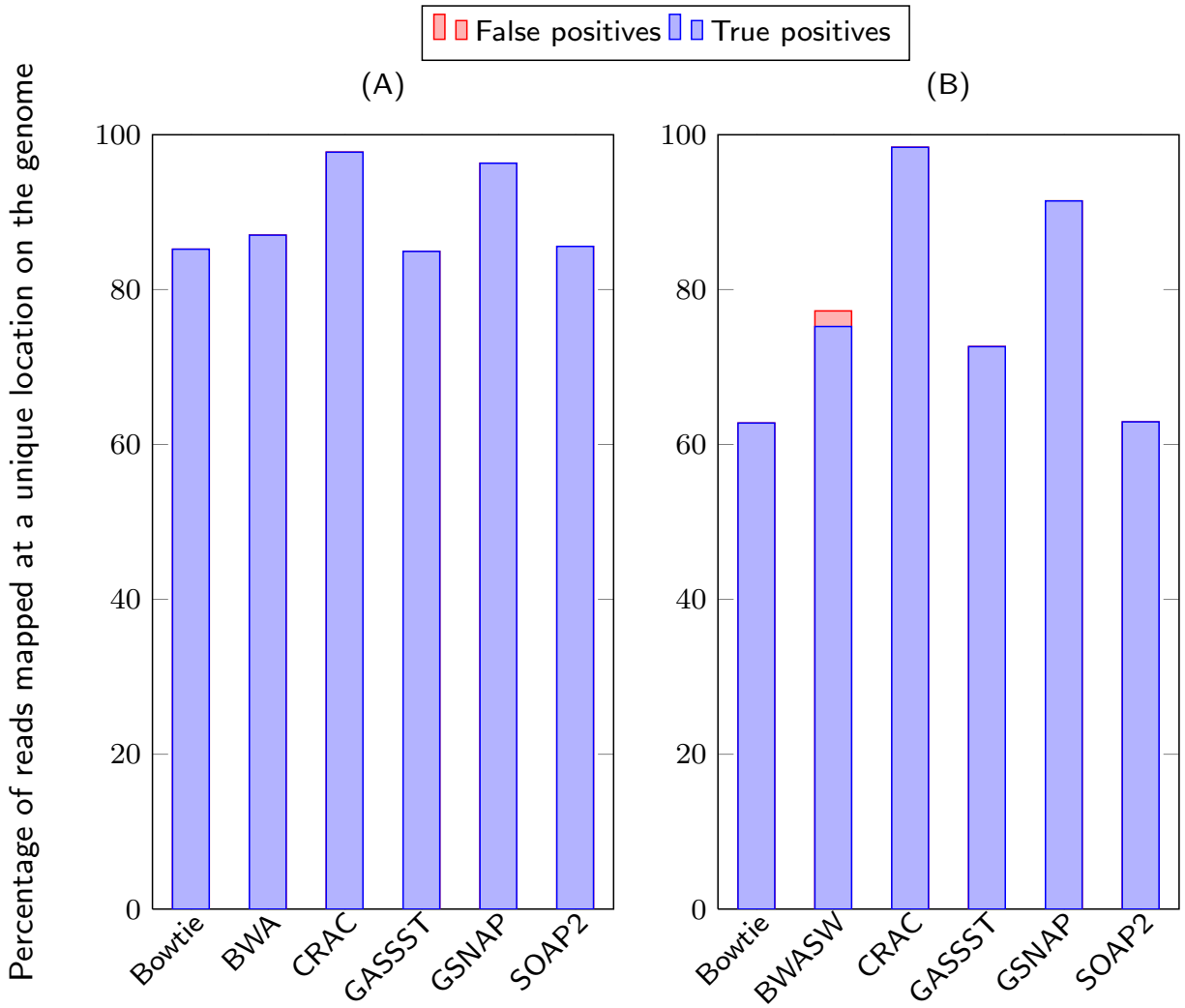
February 13, 2013

1 Results on Drosophila simulated RNA-seq data

All analyses were performed on sets of Human and Drosophila simulated RNA-seq data to assess the impact of the reference genome. Results on Human data are presented in the manuscript, while all pendant results on Drosophila datasets are given here. Although Drosophila and Human genomes differ in length, gene density as well as in number of short introns, the results are similar between the two species for mapping, splice junction or chimeric RNA predictions.

Mapping Figure 1 compares the sensitivity and precision of mapping between Bowtie, BWA / BWA-SW, CRAC, GASSST, GSNAP and SOAP2 [1, 2, 3, 4, 5]. The version and parameters used for these programs are given in Additional File 2. As for Human data, the percentages of uncorrectly mapped reads (in red) are almost invisible except for BWA-SW on 200 nt reads, meaning that almost all output genomic locations are correct. However, the difference in sensitivity remains and shows that CRAC exhibits both high sensitivity and precision. Again, its behavior improves with longer reads.

Figure 1: Comparison of **sensitivity** and **precision** on simulated RNA-seq against the drosophila genome for (A) *simulatedDroso75nt-45M* and (B) *simulatedDroso200nt-48M*.



Normal and chimeric splice junctions detection. Table 1 shows the sensitivity and precision of splice junction prediction on *D. melanogaster* simulated data. CRAC is compared to TopHat, MapSplice, and GSNAP [6, 7, 8]. Again CRAC is highly sensitive, even if TopHat achieves between +2 to +4 points in sensitivity, but CRAC remains the most precise among all tools. For instance, TopHat yields 10 to 20 times more false positive junctions than CRAC.

Table 1: Sensitivity and precision of detection of splices among different softwares. TP is the number of true positives and FP the number of false positives.

	75bp				200bp			
Tool	Sensitivity	Precision	TP	FP	Sensitivity	Precision	TP	FP
CRAC	**87.31**	**99.78**	39,637	84	**91.15**	**99.59**	42,835	178
GSNAP	80.67	99.05	36,623	350	79.7	98.8	37,453	454
MapSplice	86.19	**99.54**	39,127	182	89.31	**99.42**	41,971	244
TopHat	**91.04**	95.94	41,329	1,749	**93.89**	94.93	44,123	2,354

Table 2 shows the sensitivity and precision of chimeric junction prediction on *D. melanogaster* simulated data. CRAC is compared to MapSplice [7], TopHat-fusion [9], and TopHat-fusion-Post (*i.e.*, TopHat-fusion followed by a post-processing script).

Here, both CRAC and TopHat-fusion achieve better sensitivity than on Human data. However, CRAC reaches much higher precision than any other tool, at the exception of TopHat-fusion-Post which has 100% precision but delivers only 2 candidate chimeric junctions, that is $< 1\%$ sensitivity.

Table 2: Sensitivity and precision of detection of chimera among different softwares. TP is the number of true positives and FP the number of false positives.

	75bp				200bp			
Tool	Sensitivity	Precision	TP	FP	Sensitivity	Precision	TP	FP
CRAC	**75.94**	**99.8**	1,069	2	**68.29**	**99.1**	1,217	11
MapSplice	3.63	36.45	51	89	**3.2**	**0.19**	57	29,784
TopHatFusion	**82.35**	47.13	1,157	1,298				
TopHatFusionPost	0.14	**100**	2	0				

2 Additional results on Human simulated RNA-seq data

2.1 Comparison of 11 vs 42 million reads

We assessed the impact on mapping results of the size of the dataset in terms of number of reads, and hence of coverage. We performed the same analysis with a subset of 11 million reads and with the whole set of 42 million reads. The read length is 75 nt. The results for each set and for all tools are displayed in Figure 2 (A) for 11 millions and (B) for 42 millions reads. The impact is negligible, except for BWA that yields more false locations (small red bar on top of the blue one in A) with the medium size set (96.28 vs 99.13%). Especially, CRAC sensitivity and precision are not impacted by the number of reads, although this number changes the support values. For comparison, as shown in the manuscript, using longer reads impacts much deeply all mapping tools (Figure 3 in the MS).

2.2 Comparison of running times and memory usages

We give in Table 3 the running times and memory usages observed for mapping and splice junction prediction with various programs for processing the 42 million of 75 nt reads (Human simulated data). Times can be in days (d), hours (h) or even minutes (m), while the amount of main memory is given in Gigabytes (Gb). Although CRAC performs several prediction tasks - for point mutations, indels, splice junction and chimeric RNAs - its running time is longer than those of mapping tools and shorter than those of splice junction prediction tools. Its memory consumption is larger due to the use of a read index, the Gk arrays. This index is indispensable to query the support profile of each read on the fly.

Programs	Bowtie	BWA	GASSST	SOAP2	CRAC	GSNAP	MapSplice	TopHat
Time (dhm)	7h	6h	5h	40m	9h	2d	4h	12h
Memory (Gb)	3	2	43	5	38	5	3	2

Table 3: Running times and memory usages observed for mapping or splice junction prediction with various programs.

3 Cases of failures

For some simulated datasets, we experienced failures while running other tools in our comparisons, as mentioned in the Results of the article. For instance, TopHat-fusion did not deliver results on the 200 nt read datasets [9]. TopHat-fusion was unable to process the 200 nt simulated reads for a yet unknown reason. On that input, TopHat-fusion ran during about one month, while still filling temporary files but it stopped without any error message. We tried a few times and always obtained the same results. Finally, we contacted TopHat-fusion's contributors twice *via* their mailing list, but did not obtain any reply.

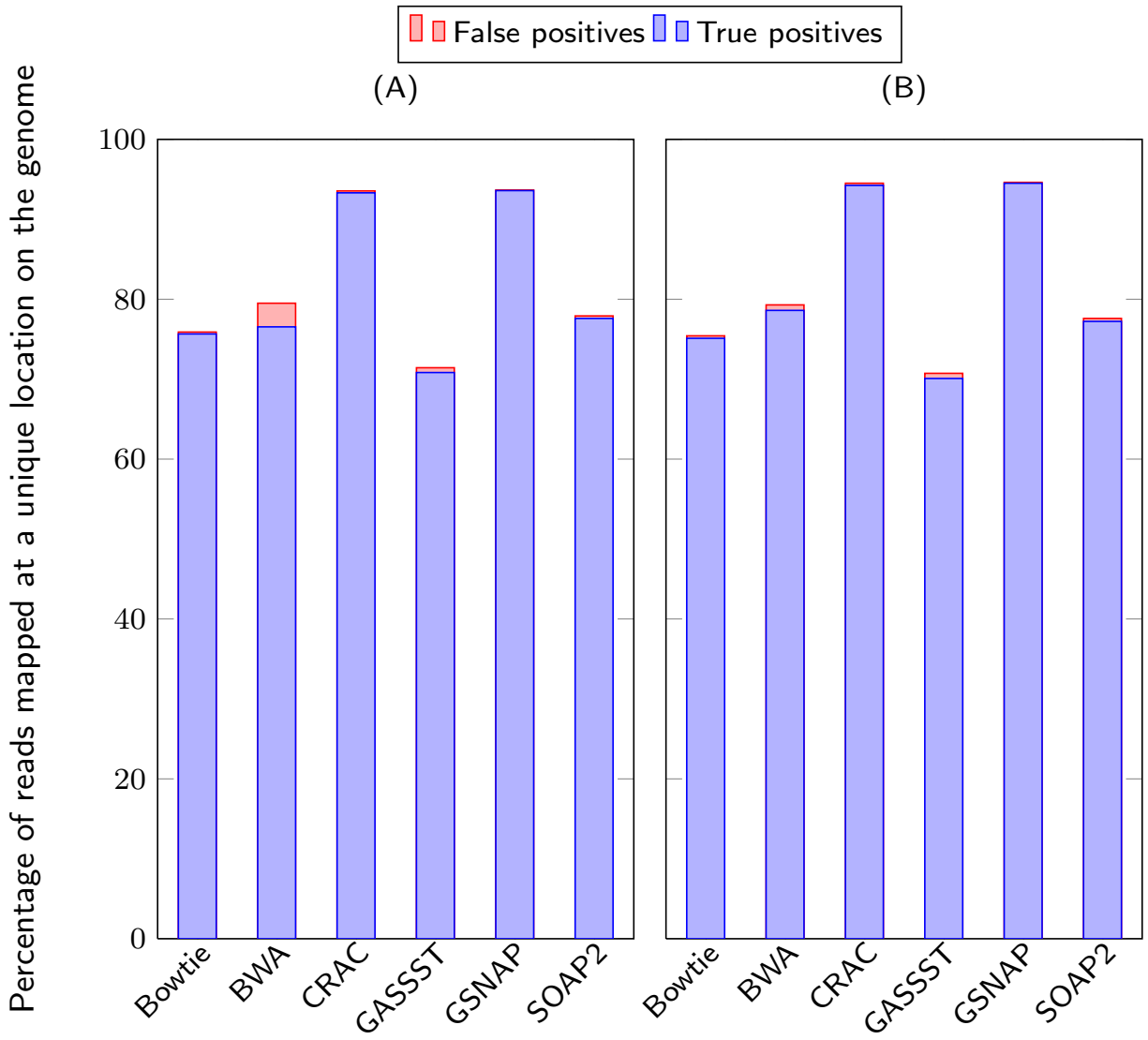


Figure 2: Impact on mapping results of medium (A) versus large (B) dataset. Comparison of **sensitivity** and **precision** on simulated RNA-seq against the Human genome on medium and large size datasets (11M-75 nt vs 42M-75 nt).

References

[1] Li, H. and Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**(14), 1754–1760 (2009). 1

[2] Li, H. and Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**(5), 589–595 (2010). 1

[3] Li, R., Li, Y., Kristiansen, K., and Wang, J. SOAP: short oligonucleotide alignment program.

Bioinformatics **24**(5), 713–714 (2008). 1

[4] Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**(3), R25 (2009). 1

[5] Rizk, G. and Lavenier, D. GASSST: global alignment short sequence search tool. *Bioinformatics* **26**(20), 2534 –2540 (2010). 1

[6] Trapnell, C., Pachter, L., and Steven L. Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**(9), 1105–1111 (2009). 3

[7] Wang, K., Singh, D., Zeng, Z., Coleman, S. J., Huang, Y., Savich, G. L., He, X., Mieczkowski, P., Grimm, S. A., Perou, C. M., MacLeod, J. N., Chiang, D. Y., Prins, J. F., and Liu, J. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.* **38**(18), e178 (2010). 3

[8] Wu, T. D. and Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**(7), 873 –881 (2010). 3

[9] Kim, D. and Salzberg, S. L. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol.* **12**(8), R72 (2011). 3, 4

CRAC: An integrated approach to analyse RNA-seq reads

Additional File 4

Results on real RNA-seq data.

Nicolas Philippe and Mikael Salson and Thérèse Commes and Eric Rivals

February 13, 2013

1 The real RNA-seq data sets

Five distinct Human RNA-seq datasets were used for assessing the capacity of predicting splice junctions and chimeric RNAs from CRAC and other tools. The main characteristics of these data sets are summarized in Table 1. The first four lines are breast cancer libraries sequenced using unstranded paired-end RNA-seq from Edgren *et al.* [1]. The last line, ERR030856, corresponds to a normal multi-tissue library sequenced using stranded RNA-seq.

Data source	Library	Read type	Fragment length	Read length	Number of fragments (or reads)
Breast cancer libraries [1]	BT474	Paired	100-200	50	21,423,697
	SKBR3	Paired	100-200	50	18,140,246
	KPL4	Paired	100	50	6,796,443
	MCF7	Paired	100	50	8,409,785
ERR030856	16 normal tissue mixtures	Single	-	100	75,000,000

Table 1: Real Human RNA-seq data used to compare splice and chimeras detection tools: four breast cancer libraries of [1] of unoriented 50 nt reads, sequenced with 1G Illumina Genome Analyzer 2X, and accessible at NCBI Sequence Read Archive [SRA:SRP003186]; one collection of 100 nt oriented reads sequenced with HiSeq 2000 Illumina® from 16 normal tissues mixtures from 11 adult individuals of widespread ages ([19;86]) from Experiment E-MTAB-513 of Illumina bodyMap2 transcriptome (see details at http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-513; this collection is accessible at http://trace.ddbj.nig.ac.jp/DRASearch/experiment?acc=ERX011226.

The tools, versions and parameters used for the comparison in all analyses are given in Table 4 of Additional File 2.

2 Predicting splice junctions on real RNA-seq data

Four programs, CRAC, TopHat, GSNAP, and MapSplice were launched to predict splice junctions on a data set of 75 million stranded 100 nt reads (ERR30856). Splice junctions were then confronted to Human RefSeq transcripts to determine whether positions found coincide with start/end of known RefSeq exons. Found junctions were partitioned into *known*, *new* and *other* junctions (see the main manuscript for a definition). We determined the intersections between the set of predicted junctions for any combination of tools. The agreement, *i.e.* the size of these intersections, are displayed in the form of Venn diagrams. These plots were obtained using Venny at `http://bioinfogp.cnb.csic.es/tools/venny/index.html`.

Figures 1 and 2 show the agreement between the predictions of each tool respectively on novel junctions, and on multi-exon RefSeq transcript for which at least one known or novel splice junction was detected.

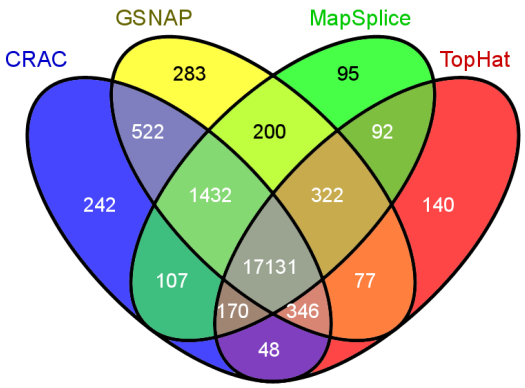


Figure 1: Venn diagram showing the agreement among tools on known junctions using known RefSeq transcripts on the ERR030856 Human dataset.

2.1 Identifying reads covering small exons

Thanks to its *k*-mer profiling approach, CRAC can detect reads that covers multiple adjacent splice junctions in the same transcript, and therefore includes entirely some small exons. CRAC identifies several breaks in the location profile of such reads and determines the donor and acceptor genomic positions of each junction. An example of read that covers two adjacent junctions

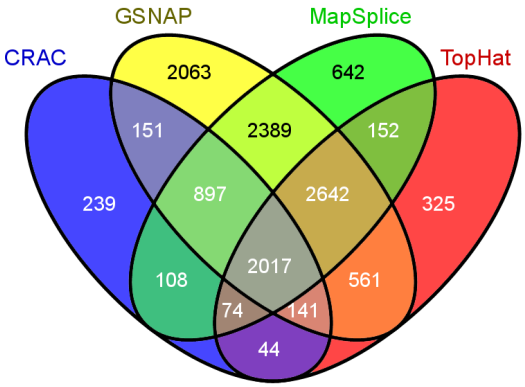


Figure 2: Venn diagram showing the agreement among tools on new splice junctions using known RefSeq exons on the ERR030856 Human dataset.

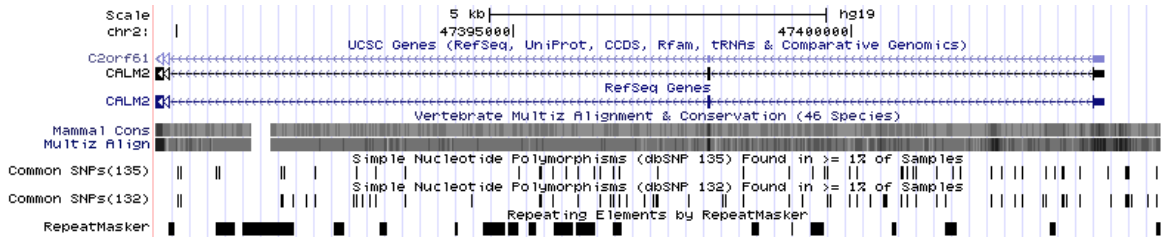


Figure 3: A read spanning three exons and two splice junctions of human Calmodulin 2 (CALM2) gene. This graphical view was obtained from the UCSC genome browser.

and incorporates a 29 nt exon of the Human calmodulin 2 gene (CALM2) is illustrated in Figure3 as viewed in the UCSC genome browser at http://genome.ucsc.edu/cgi-bin/hgTracks?org=human.

2.2 Agreement on splice junctions found by CRAC, TopHat, GSNAP, and MapSplice on the ERR030856 library

We predicted splice junctions on the ERR030856 library with each of CRAC, GSNAP, MapSplice, and TopHat (see Results in the main manuscript). First, we investigated the agreement between these four tools on *Known Junctions* (KJ) in the tables 3 and 2. Table 3 gives the number of junctions reported by each tool, as well as percentages of junctions in the intersection of all four tools, or among the three tools that perform best on this dataset (CRAC, GSNAP, MapSplice). As commented in the manuscript, we observed a large agreement among them. For more details, we also computed the numbers and percentages of KJ that are specific to each tool,

	CRAC	GSNAP	MapSplice	TopHat
Total	142,000	144,180	140,876	116,687
Shared %	97.41	97.53	99.42	98.83
Shared C-G-M %	89	87	89	NA
Shared all %	72	70	72	87

Table 2: Agreement on Known Junctions (KJ) predicted on library ERR030856 by four tools. *Total*: number of reported KJ. *Shared %*: percentage of KJ shared with at least one other tool. *Shared C-G-M*: percentage of KJ shared by CRAC, GSNAP, MapSplice. NA: not applicable. *Shared all*: percentage of KJ shared with all three other tools. For each tool, $> 97\%$ of the KJ it finds are also predicted by one other program. The agreement on well annotated junctions is larger among CRAC, GSNAP, MapSplice, than with TopHat; this is explained by the fact that TopHat finds $\simeq 25,000$ splice junctions less than the other tools.

or in the intersection of any combination of tools; see Table 2.

Finally, we computed the percentage of known junctions found by CRAC that are also reported by the other tools. We then focused on i/ reads covering entirely small exons and ii/ KJ with a large intron reported by CRAC. We computed for each category, how many items the other tools were able to report. Results are displayed in Table 4, where we also calculated the probability that a given tool found that many reads/junctions or less. The probability is computed assuming a binomial distribution and therefore assuming that the category considered represents a random sample of known junctions.

2.3 Further investigations on junctions

If the four tools show a good agreement on known junctions, it is less the case with new junctions and other junctions. Regarding other junctions, we cannot rely on RefSeq annotations to infer canonical junctions that would easily be comparable among the four tools.

To circumvent those problems, we performed another experiment that should give more insights on the predictions made by the four tools. We used the predictions made by the four tools to extract a genomic sequence of 25 nt upstream and 25 nt downstream of the junction. The 50 nt sequence is then Blasted against both the human mRNA refseq [2] and the human ESTs [3]. Blastn was launched using the following options `-F F -W 15 -a3 -G 5 -E 3 -e 0.001 -w -1 -B 1`. For obvious reasons, there are much more hits on the ESTs than on mRNA RefSeq. Therefore in the following we only report hits on ESTs. Good hits, with low E-values ($\leq 10^{-15}$), witness the fact that a predicted junction is found with high confidence, (almost) exactly on existing ESTs. Good hits should be taken as additional evidence rather than as a guarantee of the existence of this junction. On the other hand, in hits with high E-values ($\geq 10^{-10}$), only one half of

[2] Recovered using `homo sapiens[organism] AND mrna [Filter] AND refseq [Filter]` on http://www.ncbi.nlm.nih.gov/nuccore.

[3] Recovered from http://www.ncbi.nlm.nih.gov/nucest/?term=homosapiens[organism] and filtered out identical sequences resulting in 8,469,118 distinct sequences.

	Known Junctions only found by				Intersection of the junctions found by										
	CRAC	GSNAP	MapSplice	TopHat	C-G	C-M	C-T	G-M	G-T	M-T	C-G-M	C-G-T	C-M-T	G-M-T	All
KJ #	3,683	3,565	815	1,370	2,500	2,019	775	2,418	951	852	25,170	3,137	3,163	4,886	101,553
CRAC %	2.59				1.76	1.42	0.55				17.73	2.21	2.23		71.52
GSNAP %		2.47			1.73			1.68	0.66		17.46	2.18		3.39	70.43
MapSplice %			0.58			1.43		1.72		0.6	17.87		2.25	3.47	72.09
TopHat %				1.17			0.66		0.82	0.73		2.69	2.71	4.19	87.03

Table 3: Agreement on Known Junctions (KJ) predicted on library ERR030856 by four tools: detailed figures for any combination of tools. *KJ #*: number of KJ found specifically by a tool or a combination of tools. *Tool %*: percentage of the corresponding combination (in column) over the total found by the tool on that line. Empty columns are combination not including that tool. A combination of tools is denoted by the initials of the corresponding programs: for instance, the combination C-G-T corresponds to junctions found by CRAC, GSNAP, and TopHat.

	CRAC	GSNAP	MapSplice	TopHat
Agreement with CRAC %	100	93	93	76
Reads covering two KJ	9,817	8,338	9,167	7,496
Probability		9.61×10^{-178}	0.972	0.374
Reads covering three KJ	89	34	78	52
Probability		2.36×10^{-41}	5.09×10^{-2}	1.20×10^{-4}
KJ with intron $\geq$ 100 Knt	752	695	589 [1]	470
Probability		0.212	2.06×10^{-3}	6.46×10^{-18}

Table 4: Finding read covering multiple Known splice Junctions (KJ) and KJ with large introns. Ratio of KJ found by CRAC and also reported by the other tool. In the prediction of CRAC, we consider first the reads that cover two or three KJ (such reads include entirely one or more exons), and then KJ with large introns. Among the reads, respectively KJ, found by CRAC, we computed how much are also reported by the tool in that column, as well as the probability that it finds that many reads or less, according to its global agreement with CRAC. The probability says if the tool does at least as good at finding such reads/junctions as one would expect given its agreement with CRAC. For most of the category, GSNAP, MapSplice, and TopHat find less reads/junctions than CRAC. However, *e.g.* MapSplice and TopHat find about as much reads covering 2 exons as expected "by chance" ($p > 0.05$), while GSNAP finds significantly less than expected. All tools find less than expected reads covering three junctions, while MapSplice, and TopHat find less KJ with large introns than expected.

MapSplice, due to the default parameters, was not able to report junctions with an intron $\geq$ 200 knt. In the probability calculation we therefore removed 96 junctions reported by CRAC, that have such a large intron.

	CRAC	GSNAP	MapSlice	TopHat	All but CRAC
Aligned	115	704	258	131	1 056
Percentage aligned	48 %	34 %	40 %	40 %	40 %

Table 5: Absolute and relative numbers of new junctions only predicted by CRAC, GSNAP, MapSplice or TopHat that were aligned to human ESTs with an E-value $\leq 10^{-15}$ or junctions that where predicted by all tools but CRAC.

	CRAC	GSNAP	MapSlice	TopHat
Aligned	11 395	15 975	13 907	11 579
Percentage aligned	69 %	47 %	50 %	44 %

Table 6: Absolute and relative numbers of other junctions predicted by CRAC, GSNAP, MapSplice or TopHat that were aligned to human ESTs with an E-value $\leq 10^{-15}$.

the junction has been aligned. Such hits demonstrate that the predicted junction was not seen in the whole collection of human ESTs, and are therefore likely to be false positives.

2.3.1 Blasting specific new junctions

Since there exists a discrepancy among the predictions of new junctions, we started by blasting them. More specifically, we focus on junctions that are detected by only one tool. Since the intersection between GSNAP, MapSplice and TopHat is the largest one, we also take into account junctions from that set.

CRAC yields less new junctions that are specific to it compared to GSNAP or MapSplice, but, as can be seen in Table 5, CRAC is more accurate than concurrent methods. Predictions made by the other tools are slightly less reliable than CRAC's. On the other hand, CRAC delivers less predicted junctions of that specific category than the other tools. For reasons explaining that, see section 2.4.

2.3.2 Blasting other junctions

We also reproduced the experiment on the sets of other junctions of each tool. We also focus on high quality hits having an E-value lower than or equal to 10^{-15}. The results are presented in Table 6. We observe that GSNAP and MapSplice have the highest number of high quality alignments, while CRAC has the highest proportion.

2.3.3 Blasting all junctions

Since the separation between known, new and other junctions is somehow arbitrary, and is relative to RefSeq, it is also interesting to consider all junctions predicted by a tool altogether to assess each tool's performance. As a summary we made two plots, in Figure 4. We notice that GSNAP predicts more high quality hits (159,702), followed by MapSplice (152,957), followed

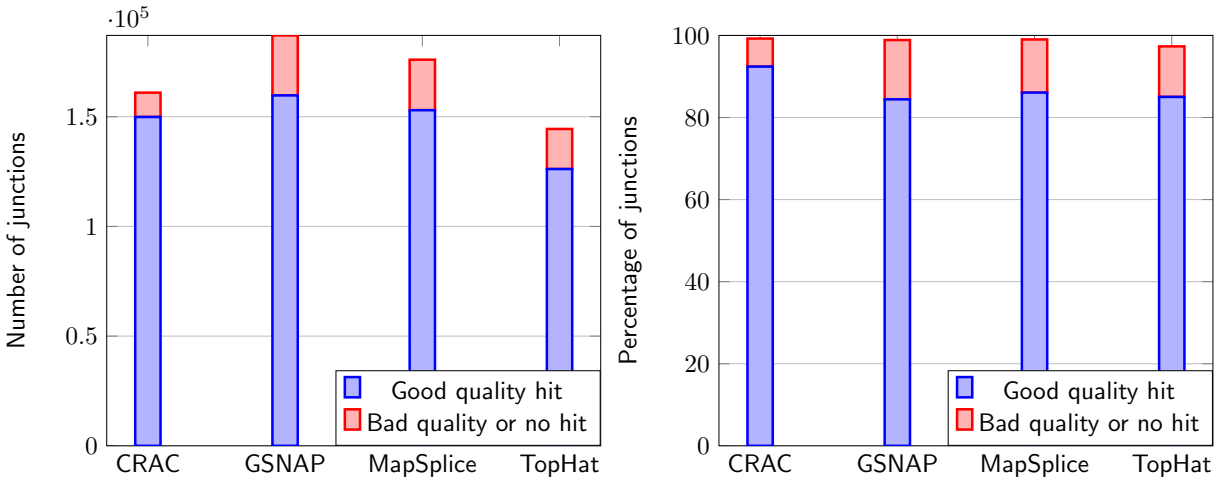


Figure 4: Aligning all the junctions from CRAC, GSNAP, MapSplice, TopHat to human ESTs using BlastN. Hits with E-values $\leq 10^{-15}$ are considered as good quality hits whereas hits with E-values $\geq 10^{-10}$ or junctions that were not aligned are considered as bad quality or no hit. Figures are provided first with absolute numbers (number of distinct junctions) and then as a percentage of the total number of distinct junctions.

by CRAC (149,872) and TopHat (126,143). CRAC is characterised by a low proportion of bad quality hits (6.8 %) versus 14 % for GSNAP, 13 % for MapSplice and 12 % for TopHat.

2.4 Investigating new junctions unmapped by CRAC

To understand why CRAC had its worst performances with the new junctions, we analyse a random sample drawn from the junctions predicted by the three other tools together. Twenty-one junctions are sampled out of 2,642, and the corresponding read where they appear are considered for a manual analysis. Of these junctions, nineteen are weakly expressed alternative transcripts. Meaning that these specific junctions are rare but the involved exons also participate in other junctions, that are much more expressed. Therefore CRAC identifies a variation in the support profile (the exons are well expressed, but the junction is poorly expressed) and considers that it may consist of a sequencing error. However CRAC is aware that this kind of error is unusual for a sequencing error. That is why CRAC classifies sixteen of these cases as an "undetermined error" and gives more clue by stating that it is probably a splicing event (the positions of the event are also given).

2.5 Testing junction prediction on negative controls.

We report in the Results section of the MS, the output of CRAC on a set of negative controls splice junctions obtained by associating true RefSeq exons. The command line used for running CRAC is:

```
crac -i GRCh37 -r random-refseq-junction-reads-100k.fa -k 22 -m 76
```

```
--splice random-refseq-junction-reads-100k-GRCh37-22.splice
--nb-threads 2 -n 15 --max-splice-length 300000 --max-duplication 5
 --min-percent-duplication-loc 0.5 --min-loc-repetition 2
```

The collection of reads used as negative controls is available at: http://crac.gforge.inria.fr/

3 Predicting chimeric RNAs on four breast cancer libraries.

3.1 Parameters for CRAC

To test CRAC on real data regarding the identification of chimeric RNA (chRNA), we compared its results to the findings of Edgren *et al.* [1] and of TopHat-fusion on four breast cancer RNA-seq libraries. These were published in Edgren [1] and also analysed in TopHat-fusion [2]. Contrarily to the other data we used, either simulated or real, these RNA-seq libraries contain shorter reads: 50 nt. Hence, we needed to adapt CRAC's parameters to take this shorther length into account. We alter two parameters:

- the number of adjacent k-mers that must consistently indicate the same unique location in a read was decreased from 15% to 10% of the read length, that is from 7 to 5 (`-min-percent-single-loc 0.10`)
- the number of k-mers adjacent to each side of the break border whose location is checked for concordance was lowered to 2 instead of 10 (`-max-extension-length 2`). This parameter is used during the break fusion procedure to determine whether we face a colinear (*i.e.*, normal) rather than a chimeric splice junction.

We used $k = 22$, as for the other analyses to avoid an increase in false locations; all other parameters were left by default or as for the other analyzes (see Table 4 of Additional File 2).

We used stringent criteria for predicting chRNA, which is done by setting the following parameters:

```
chimera_break >= k-1-(5)
min_support_in_chimera >= 2
max_extension_for_find_single_loc =5 for each border break
```

3.2 Filtering for normal splice junctions with GSNAP

We filtered the chRNA predicted by both CRAC and TopHat-fusion using GSNAP to avoid those that could have a continuous or colinear splice alignment with slightly less identities. Such an alignment represents an alternative to the detected chimeric alignment. Thus, we consider such candidates to be less robust. For this filtering, we set the parameters that enable GSNAP to detect splice junctions in individual reads, *i.e.* the `--novelsplicing (or -N)` flag. All other options were set to default.

3.3 Rerunning TopHat-fusion

In the article, we report several recurrent chRNAs detected by CRAC but not found by TopHat-fusion. We sought to understand the reasons of this difference, especially if TopHat-fusion detects these chimeric junctions based on alignment criteria, but then filter them out based on biological knowledge. As TopHat-fusion reports first the set of reads that generates the initial hits (in file `accepted_hits.sam`) before its internal filtration step, it is possible to answer this question. For this sake, we ran TopHat-fusion on the four libraries as described in their article [2], and searched all detected chRNAs in its intermediate file.

Parameters of TopHat-fusion: `--fusion-anchor-length 20`

3.4 Running times for the breast cancer libraries

Table 7 gives the running times of CRAC and TopHat-fusion to analyze each of the four breast cancer libraries of 50 nt reads. CRAC is between 5 and 10 times faster than TopHat-fusion.

Breast cancer libraries [1]	BT-474	KPL-4	MCF-7	SK-BR-3
CRAC	1h50m	41m	54m	1h05m
TopHat-fusion	11h58m	3h28m	4h22m	11h12m

Table 7: CPU time for CRAC and TopHat-fusion to process with 4 threads the Breast cancer libraries BT-474, KPL-4, MCF-7 and SK-BR-3 from [1].

3.5 Distribution of candidate chimeric RNA found by CRAC

CRAC predicted 455 candidate chRNAs that are partitioned in five classes, as explained in Section 6 of Additional File 2. Class 2 candidates represent only two percents of the total, thereby showing that, although arbitrary, the threshold used to distinguish between splice inside one gene or across distinct genes, works reasonably for Human data. Annotations show that some of these cases are indeed normal splice junctions inside a known gene.

Class	Nb	Total	Proportion
1	118	455	0.26
2	10	455	0.02
3	109	455	0.24
4	127	455	0.28
5	91	455	0.20

3.6 Case candidate "chimeric" RNA with internal repeat located inside LONP1 gene

This candidate chRNA is identified in class 5: it appears as an inversion because of an internal repeat. We use the term "chimeric" simply because such reads cannot be explained with sim-

ple colinear alignments. It means "non colinear" and makes no assumption about underlying mechanisms.

Figure 5 shows the analysis of one of the reads that gave rise to this prediction. Neither can it be mapped continuously on the genome, nor did GSNAP find a continuous alignment for it. Instead, it is mapped as a chimeric read with a small scale inversion on chromosome 19 minus strand in two parts depicted in blue and yellow. The k-mer location profile exhibited a break after the blue part, and the first located k-mer after the break is at the start of the yellow part. The blue part ends at position $5,692,012$, while the yellow part starts at position $5,691,992$, *i.e.* slightly before. Hence, CRAC classifies it as a chimera with inversion. Both parts overlap on the chromosome 19, which implies that the read contains a sequence repeated twice $TCA \ldots AGA$ (shown in boldface below). This chimeric alignment is confirmed by BLAT (below), which finds exactly the same junction point.

This duplication could be due to a known variant. We thus searched for possible known variants in this chromosomal region in eight distinct Human genomes on Ensembl, but find none [3]. However, we observed this chimeric junction, but also found the same junction without the duplication in other libraries. Both variants are found in public EST libraries in equal proportion and at non negligible expression levels. Moreover, we found the variant with duplication also in five private (healthy and tumoral) libraries, but neither in ERR030856, nor in a K562, while the variant without duplication is present in three private libraries and in K562. These evidences raise the possibility that this LONP1 unannotated junction may not just be due to transcriptomic noise, may be regulated, and thus functional. It is striking that such a type of read (class 5) is found in high proportion among the chimeric RNA candidates, suggesting that this LONP1 variant is not an isolated case. Larger investigations over more libraries are needed to confirm or infirm our assumptions.

References

[1] Edgren, H., Murumagi, A., Kangaspeska, S., Nicorici, D., Hongisto, V., Kleivi, K., Rye, I. H., Nyberg, S., Wolf, M., Borresen-Dale, A., and Kallioniemi, O. Identification of fusion genes in breast cancer by paired-end rna-sequencing. *Genome Biol.* **12**(1), R6 (2011). 1, 9, 10

[2] Kim, D. and Salzberg, S. L. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol.* **12**(8), R72 (2011). 9, 10

[3] Flicek, P., Aken, B. L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., Down, T., Dyer, S. C., Eyre, T., Fitzgerald, S., Fernandez-Banet, J., Graf, S., Haider, S., Hammond, M., Holland, R., Howe, K. L., Howe, K., Johnson, N., Jenkinson, A., Kahari, A., Keefe, D., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Megy, K., Meidl, P., Overduin, B., Parker, A., Pritchard, B., Prlic, A., Rice, S., Rios, D., Schuster, M., Sealy, I., Slater, G., Smedley, D., Spudich, G., Trevanion, S., Vilella, A. J., Vogel, J., White, S., Wood, M., Birney, E., Cox, T., Curwen, V., Durbin, R., Fernandez-Suarez, X. M., Herrero, J., Hubbard, T. J. P., Kasprzyk, A., Proctor, G., Smith, J., Ureta-Vidal, A., and Searle, S. Ensembl 2008. *Nucleic Acids Res.* **36**(S1), D707–714 (2008). 11

Recurrent candidate chimeric RNA in LONP1 gene

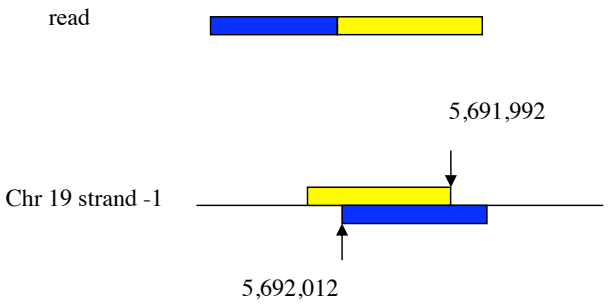
GENOME chr 19 brin -1: 5' towards 3'

Positions: 5,691,988 – 5,692,012 in blue, and 5,691,992 – 5,692,016 in yellow

Read: **TCAGGCCCTGTCTGGGCCAGA***ACTG**GATG***TCAGGCCCTGTCTGGGCCAGA**

repeat sequence inside the read shown in bold

Genomic locations identified by CRAC: black arrows




```
5692017 gatgtcaggccctgtctgggccaga          5691992
5692012     tcaggccctgtctgggccagaactg      5691992
```

BLAT alignment in two pieces

5'end

TCAGGCCCTGTCTGGGCCAGAACTGGATGTCAGGCCCTGTCTGGGCCAGA

```
1    25    25 100.0%    19   -    5691988   5692012     25

acactaccgg gagatcttcg acatcgcctt cccggacgag caggcagagg  5692063
cgctggccgt ggaacggtga cggccacccc gggactgcag gcggcggatg  5692013
TCAGGCCCTG TCTGGGCCAG AACTGagcgc tgtggggagc gcgcccggac  5691963
ctggcagtgg agccaccgag cgagcagctc ggtccagtga cccagatccc  5691913
agggacctca gtcggcttaa tcaga
```

3'end

TCAGGCCCTGTCTGGGCCAGA*ACTG**GATG***TCAGGCCCTGTCTGGGCCAGA**

```
1    25    25 100.0%    19   -    5691992   5692016     25

tggaacacta ccgggagatc ttcgacatcg ccttcccgga cgagcaggca  5692067
gaggcgctgg ccgtggaacg gtgacggcca ccccgggact gcaggcggcg  5692017
GATGTCAGGC CCTGTCTGGG CCAGAactga gcgctgtggg gagcgcgccc  5691967
ggacctggca gtggagccac cgagcgagca gctcggtcca gtgacccaga  5691917
tcccagggac ctcagtcggc ttaat
```

Figure 5: A candidate chimeric RNA involving LONP1 exons.