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

Genome Biology 2013, 14:R30 doi:10.1186/gb-2013-14-3-r30

Nicolas Philippe (nphilippe@lirmm.fr)
Mikaël Salson (Mikael.Salson@lifl.fr)
Thérèse Commes (commes@univ-montp2.fr)
Eric Rivals (rivals@lirmm.fr)

ISSN 1465-6906

Article type Software

Submission date 18 November 2012

Acceptance date 5 March 2013

Publication date 28 March 2013

Article URL http://genomebiology.com/2013/14/3/R30

This peer-reviewed article can be downloaded, printed and distributed freely for any purposes (see copyright notice below).

Articles in Genome Biology are listed in PubMed and archived at PubMed Central.

For information about publishing your research in Genome Biology go to

http://genomebiology.com/authors/instructions/

© 2013 Philippe et al.
This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
CRAC: an integrated approach to the analysis of RNA-seq reads

Nicolas Philippe\textsuperscript{1,2,5,6}, Mikaël Salson\textsuperscript{3,4,6}, Thérèse Commes\textsuperscript{2,5} and Eric Rivals\textsuperscript{*1,5}

\textsuperscript{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. \textsuperscript{2}Institut de Recherche en Biothérapie (IRB), U1040 INSERM, CHRU Montpellier Hôpital Saint-Eloi 80, av. Augustin Fliche 34295 Montpellier Cedex 5, France. \textsuperscript{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. \textsuperscript{4}LITIS EA 4108, Université de Rouen, 1 rue Thomas Becket 76821 Mont-Saint-Aignan Cedex, France. \textsuperscript{5}Institut de Biologie Computationnelle, 95 Rue de la Galéra 34095 Montpellier Cedex 5, France

Email: Nicolas Philippe - nphilippe@lirmm.fr; Mikaël Salson - Mikael.Salson@lifl.fr; Thérèse Commes - commes@univ-montp2.fr; Eric Rivals\textsuperscript{*} - rivals@lirmm.fr;

\textsuperscript{*}Corresponding author

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 to identify functionally important mutations and to explore the transcriptomic diversity of various tissues. RNA sequencing (RNA-seq) provides a genome scale access to the RNA complement of a cell with unprecedented depth, and has therefore proven useful to unravel the complexity of transcriptomes \cite{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.
Moreover, RNA-seq also gives access to those somatic mutations and genetic polymorphisms that are transcribed. If chimeric RNAs may result from the transcription of fused genes put together by chromosomal rearrangements [4], especially in cancer [5], 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 goes cheaper, bioinformatic analyses become more critical and time consuming. They still follow the same paradigm as in the first day 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) pervading erroneous information, (b) lack of integration, and (c) information loss, which induces re-computation at subsequent steps and somehow prevents cross-verification. For instance (b), 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 the 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 to 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/GASSST) [9–13], and both normal (GSNAP/TopHat/MapSplice) [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 all methods sensitivity 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 http://crac.gforge.inria.fr/ or the ATGC bioinformatics platform http://www.atgc-montpellier.fr/crac. 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 to analyze reads when a reference genome is available, although some procedures (e.g., 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, chimeric RNAs) and sequencing errors from a RNA-seq read collection.

CRAC analysis is based on two basic properties (P1-P2).

**P1** For a given genome length, a certain sequence length is sufficient to match in average to a unique genomic position with high probability. This length, denoted $k$, can be computed and optimized [17]. 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 on the reference genome. However, in average over all $k$-mers, the probability of getting a false location (FL) is $\simeq 10^{-4}$, with $k = 22$, for the Human genome size [17].

**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. On the contrary, a biological variation affecting the molecule should be visible in many reads overlapping that position.

CRAC proceeds each read in turn. It considers the $k$-mers starting at any position in the read (i.e., $m - k + 1$ possible $k$-mers). It computes two distinct $k$-mer profiles:

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 (i.e., 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 jointly these two profiles to detect multiple situations and predict in a single analysis sequencing errors, 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 [18], while the support of a $k$-mer is obtained on-the-fly by interrogating a specialized read index, called Gk-arrays [19]. 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 exploiting property P2 for distinguishing sequencing errors from variations, and gaining confidence in predictions. As illustrated below, the location profile delivers a wealth of information on the mapping situation, but detecting the concordance of variations in the two profiles makes the originality of CRAC.
**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 of exposure, we make simplifying assumptions: a/ $k$-mers have no false genomic locations, b/ the read is affected by a single difference (substitution, indel, splice junction), 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 the case of a substitution, which may be of erroneous (sequencing error) or biological origin (SNP, SNV, 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 $[h-k+1, h]$. On the contrary, $k$-mers starting left (resp. 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 positioning the difference in both the read and the genome, but does not distinguish erroneous from biological differences. The support profile will inform us on this matter.

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 homo- or hetero-zygosity 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 in a detection rule. We can apply a similar location/support profiles analysis to predict such events.

**Rule 1** (Figure 1c): Consider a read affected by a single difference (substitution, indels) compared to the genome. Let $j_b < j_a$ (b: before, a: after) be the positions immediately flanking the observed break in the location profile \( (i.e., \) break in the range \([j_b+1, j_a-1]\)). Let $\ell := j_a - j_b$ and $L$ denotes the offset between the genomic locations of the
k-mers starting in \( j_b \) and \( j_a \) \((L := loc(j_a) − loc(j_b))\). If 1) \( \ell = L = k + 1 \) the difference consists in a single substitution at position \( j_a − 1 \) in the read and \( loc(j_a) − 1 \) in the genome; 2) \( \ell = k \) and \( L = k + p \) for some integer \( p \), it 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 \( loc(j_a) − p \) and \( loc(j_a) − 1 \) on the genome; 3) symmetrically, if \( \ell = k + p \) and \( L = k \) for some integer \( p \), the difference is a \( p \) nucleotide insertion with respect to the reference.

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

The observed missing part in the read can be due to a polynucleotidic deletion, or the removal of intronic/intragenic regions by splicing. Without annotations, only the expected length \((\text{i.e., the value of } p)\) 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. Especially, some reads have a break but the genomic locations on 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, c/ different chromosomes. (For chimeric RNAs, CRAC even distinguishes 5 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), excepted that genomic locations are not colinear. Indeed, CRAC checks the break length \( \ell = k \), as well as coherence of adjacent \( k \)-mers left or right of the break. **Coherence** means that, for some (small) integer \( \delta \), \( k \)-mers in the range \([j_b − \delta, j_b]\) (resp. \([j_a, j_a + \delta]\)) 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 (chRNA) (see Discussion).

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

To describe CRAC’s method above, we first assumed simplifying conditions: especially the absence of False
Locations (FL) and of border cases. Some details clarify how the actual procedure handles real conditions.

**Differences with the genome at the read’s extremities (border cases).** Border cases are not processed with a specific procedure by CRAC; rather, the sequencing depth of NGS data saves 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 likely is located in the middle of some 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 (FL) (Figure 1d)** Our criterion to set \( k \) ensures a low average probability of a random \( k \)-mer match on the genome [17], 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 case of a break, CRAC faces two difficulties. First, when a FL happens at an 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, FL may lead to avoid Rule 1, apply Rule 2, and predict a false chimeric read. To overcome 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 as well as explanations on the distinction of dropping and non-dropping supports around a break, on read mapping at multiple locations, on the sub-classification of chimeric reads, and on the simulation protocol.

**Results**

We have evaluated CRAC for the tasks of mapping reads, of predicting candidate SNV, indels, splice junctions or chimeric junctions, and have compared it to other tools. Simulated data are compulsory to compute exact sensitivity and accuracy levels, while real data enable us to confront predictions with biologically validated RNAs. For simulating RNA-seq, we first altered a reference genome by random substitutions, indels, and translocations to derive a mutated genome, then reads are "sequenced in silico" by FluxSimulator [20] using 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: one (hs75) with a typical read length of 75, another (hs200) with reads of 200 nt representing the future to assess different strategies.
Mapping on current (75 nt) and future (200 nt) reads

Mapping, *i.e.*, 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) compute the best *continuous* genome-read alignments up to a certain number of differences [9, 11, 12]. CRAC, GSNAP [14], and Bowtie2 [21] also consider *discontinuous* alignments to search for the locations of reads spanning a splice junction: they can find both continuous and spliced alignments.

Global view of mapping results with 75 nt reads (Table 1a) 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 seek for continuous alignments with a few substitutions and small indels. Although its approach differs, GASSST also targets these (with an advantage on longer indels). Even within this group, sensitivity varies significantly: from 70% for GASSST to 79% for BWA. Such 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.

One gets a clearer view by considering the subsets of reads that carry a SNV, an indel, an error, a splice or a chimeric junction (Figure 2). Strikingly, CRAC is the only tool that achieves similar performances, 94–96% sensitivity, in all categories. For instance with indels, GSNAP yields 65 and 83% sensitivity on insertions and deletions respectively, while that of other tools remain below 30%. BWA/GASSST/Bowtie/SOAP2 output continuous alignments for 9–19% of spliced reads. Although their output locations are considered correct, for they are comprised in one exon, their alignments are not. Such reads are considered as mapped and thus not reanalyzed by tools like TopHat or MapSplice to 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 data set, 36% of the reads cover a splice junction, and 50% carry an error. Compared to 75 nt data, while their precision remains > 99%, BWA/GASSST/Bowtie/Bowtie2/SOAP2/GSNAP, loose in sensitivity (∼−10 points for BWA-SW, GASSST, and GSNAP, −14 points for Bowtie2, −20 points for Bowtie). Only CRAC remains as precise and gains 1.5 points in sensitivity (Table 1a). The detail by category confirms this situation (Figure 2), leaving CRAC above concurrent tools. CRAC’s *k*-mer profiling approach proves to handle accurately reads altered by distinct classes of biological events, and importantly adapts well to longer reads.

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

**Predicting distinct classes 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 classes (SNV, indels, splice and chimeric junctions) from the reads. The question is: if this, e.g. SNV or splice junction, is present and sequenced, can it be predicted and not buried in a multitude of false positives (FP)? Here, sensitivity and precision are relative to the number of events, not to the number of reads covering them. We assessed CRAC’s prediction capacity and compared it to splice junction prediction tools on our simulated datasets.

Figure 3 gives CRAC’s precision and sensitivity for each class of events and for sequencing error detection. On SNV and indels (< 15 nt), CRAC achieves a sensitivity ranging [60,65]% and a precision ranging in [96.5,98.5]% (Figure 3), making it a robust solution for such purposes. Typically, CRAC missed SNVs that are either low covered (42% of them appear in ≤ 2 reads) or in reads carrying several events (66% of missed SNV reads also cover a splice junction). On the splice junction class, CRAC delivers 340 False and 67,372 True Positives (TP).

The global comparison and the effect of read length on sensitivity and precision can be observed in Table 1b. With 75 nt, all splice detection tools achieve a good sensitivity, ranging from 79% for CRAC to 84% 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% FP; for comparison, MapSplice and GSNAP output four times as much FP (2.32 and 2.97%), while TopHat yields twenty times more FP (10.41%). With 200 nt reads, tools based on k-mers matching, i.e. CRAC and MapSplice, improve their sensitivity (6.5 and 5 points resp.), while mapping based approaches (GSNAP and TopHat) loose resp. 12 and 30 points in sensitivity. With long reads, CRAC offers 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.

Comparison on chimeric RNAs shows that CRAC already offers an acceptable balance between sensitivity and precision with 75 nt reads (53 and 93% resp.), while the sensitivities of TopHat-fusion and MapSplice remain below 32% (Table 1c). With 200nt reads, only CRAC is able to predict chimeric splice junctions with acceptable precision, and it improves its sensitivity compared to shorter reads (Table 1c and Additional file 3).

As for mapping, for all classes of events, CRAC prediction performances improve with longer reads (Figure 3).
Predicting distinct classes of biological events on real data

**Splice junction prediction**

To evaluate CRAC’s capacity to detect splice junctions in real RNA-seq data, we compared it to state-of-the-art tools (TopHat, GSNAP, MapSplice) on a data set of 75 million stranded 100 nt reads (ERR030856 - Additional file 4 Table S1). Splice junctions were searched for using each tool and then confronted to Human RefSeq transcripts. Each found junction consists of a pair of genomic positions (i.e., the exons 3’ end and 5’ start) and we consider that it matches a RefSeq junction if their positions were equal within a 3 nt tolerance. Found junctions were partitioned into known, new and other junctions (KJ, NJ, and OJ 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 remaining junctions go into 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 [22]. Novel junctions will provide new alternative splicing candidates, while junctions in other represent totally new candidates RNAs.

For each tool, the distribution of junctions in 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 (KJ) among the tools is shown as a Venn diagram (Figure 4b); see Additional file 4 for the corresponding figures and Venn diagram on novel junctions (NJ). Clearly, MapSplice, GSNAP, CRAC reports 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 less than other tools, and only 1,370 of its junctions are not detected by any of them. For instance, CRAC covers 93% of TopHat KJ. 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 all other tools they account for [25;27]% of detected junctions.

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 junction.

Are the junctions predicted in the 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 all details and the 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 ESTs and yields a very low E-value (\( \leq 10^{-15} \)), while the latter gets a larger one (\( \geq 10^{-10} \)). As expected,
at least 95% of KJ yield 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 GNSAP, 49 and 50% for MapSplice, 51 and 44% for TopHat. The percentages of OJ and NJ confirmed by mRNAs lie >13% for CRAC and <8% for all other tools (excepted for OJ with TopHat: 17%, same as CRAC). If we consider all junctions, 93% of CRAC junctions align entirely to an EST with a good hit. Whatever the category of junctions, CRAC predicts more unreported junctions that are confirmed by mRNAs or ESTs than 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.

Exploiting simultaneously 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 100knt, respectively. All tools find less such junctions - respectively, 695, 589, and 470 for GSNAP, MapSplice, TopHat, but both MapSplice and TopHat find less than expected by chance according to the global agreement between these tools (Additional file 4). CRAC also reveals 69,674 reads that cover exactly two known RefSeq junctions, i.e. that cover three distinct exons and include one of them. An example of a double junction covering a 29 nt exon of CALM2 gene is illustrated in Additional file 4. Moreover, among such 9,817 junctions, GSNAP, MapSplice, TopHat, find respectively 8,338, 9,167, 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 concurrent tools. For instance, GSNAP does not map such reads. An example on TIMM50 gene is shown in Figure 4c. Altogether, these results suggest that numerous new splice junctions, even between known exons, remain to be discovered [23], 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 discriminating true from false candidates, while it offers a good potential to detect 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 done on the basis of detected junctions [24].

**Comparisons on chimeric splice junction prediction**

Edgren et al. used deep RNA-sequencing to study chimeric gene fusions in four breast cancer cell lines (BT-474, KPL-4, MCF-7, SK-BR-3- Additional file 4 Table S1), found 3 known cases and validated 24 novel intergenic fusion candidates (i.e., involving two different genes) [25]. As CRAC, TopHat-fusion can predict both intra- and inter-genic
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 and confronted to 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 remade all predictions using TopHat-fusion with and without TopHat-fusion-post.

The numbers of distinct candidate chimeric junctions (chRNA for short), and of chimeric single reads detected by both tools in each library are given in Table 2. The 50 nt reads, which are well suited for Bowtie-TopHat, represent an unfavorable case 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 on a six tissue library (ERR030856), TopHat-fusion reports about as many chimeric junctions as CRAC, GSNAP, or MapSplice do 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 has a practicable size and authorizes an in-depth, context dependent investigation to spot promising candidates for validation.

In CRAC’s report, intra- and inter-genic chRNAs account for 58 and 42% respectively, and are partitioned in 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 confronted to the set of validated chimeras of Edgren et al. [25], TopHat-fusion and CRAC detect 21 and 20 out of 27, and agrees on 17 of them (Table 3)\(^2\). Among the 7 cases CRAC misses, only one (BCAS4-BCAS3) is a false negative, 4 are unsure, not enough expressed candidates (CPNE1-P13, STARD3-DOK5, WDR67-ZNF04, PPP1R12A-SEPT10), and no read seems to match the junction of the 2 remaining ones (DHX35-ITCH, NFS1-PREX1). As BCAS4-BCAS3 junction includes a substitution near the splice site, the reads carry two events (SNV+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 4 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 unsure candidates (Undetermined file). Three out of 4 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 [25], 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 (21/14303 ≃ 0.01% vs 20/192 ≃ 10.4%; Table 3, 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, certain evidence suggest 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 MAN1A2 gene, which recur in up to 3 out of 4 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 [26]. Our results support the existence of such mRNA exhibiting shuffled exons, as well as cases where the inversion is short, sometimes inducing a repeat within the read (see an example in LONP1 gene: 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, 25] did not report any recurrent chRNAs. However, when we ran TopHat-fusion, it also outputs 23 of these chRNAs among 193, 163 candidates.

For instance, we report a HSPD1-PNPLA4 chRNA in both KPL-4 and SK-BR-3 libraries: PNPLA4 (GS2) is highly expressed in human SW872 liposarcoma cells [27], while HSPD1, the heat shock protein Hsp60, shows a broad antiapoptotic function in cancer [28]. 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 [29], and was recently found to be recurrently mutated in cystic pancreatic lesions related to invasive adenocarcinomas [30], as well as amplified in breast cancers [31]. Moreover, we also found the same CTDSPL2-HNRNPM chimeric RNA in BT-474, MCF-7, and SK-BR-3 libraries. Both genes belong to the heterogeneous nuclear ribonucleoproteins (hnRNPs) family and play 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 [32].
**Discussion**

CRAC is a multi-purpose tool for analyzing RNA-seq data, that is capable in a single run of predicting sequencing errors, small mutations, 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+TopHat/TopHat-fusion, and can be viewed as an effort to simplify NGS analysis. CRAC is neither a mapper, since it uses local coverage information (in the support profile) before computing the genomic position of a read. Contrary to the current paradigm, mapping and post inferences are not disjoint steps in CRAC. Rather, it implements a novel, integrated approach that draws inferences by analyzing simultaneously 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, thereby becoming crucial for efficiency. Indeed, CRAC takes more time than state-of-the-art mappers, but is considerably faster than splice junction prediction tools (e.g. Bowtie+TopHat). The other key to efficiency is the double indexing strategy: a classical FM-index for the genome and the Gk-arrays for the reads [19]. 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 compulsory to assess a method sensitivity, have shown that CRAC is for each type of prediction at least competitive or surpasses concurrent tools in terms of sensitivity, while it generally achieves better precision. Moreover, CRAC’s performances further improve when processing longer reads: e.g. on 200 nt reads, 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 [17], and quite conservative values – \( k = 22 \) for the Human genome – have been used in our experiments. Smaller \( k \) are possible with smaller genomes (like bacterial ones). \( k \) impacts on the number of false genomic locations (FL) that occurs in the profile; a FL indicates a wrong location for a read \( k \)-mer, i.e. different 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, the concordance of locations for the \( k \)-mers around the break (especially in the break verification and fusion procedures - see Methods). 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. FL should impact even more the prediction of chimeras. Globally, results on both simulated and real data, like the improved mapping sensitivity (+15 points compared to Bowtie/BWA/SOAP2), show that CRAC makes accurate predictions with conservative values. $k$ somehow controls a 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. [25]. Of course, the $k$ value induces two limitations: first, the minimal exon size detectable in a read is $\geq k$, second, reads must be long enough (> 40 nt with $k = 20$ for the Human genome). However, NGS progress 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/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 BCAS3-BCAS4 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, e.g. 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 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 a 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, SOAP2), but the mapping sensitivity is rarely questioned. We performed extensive mapping tests on simulated data, which show that sensitivity can truly be improved and that CRAC makes a significant step in this direction. Of course considering discontinuous alignments (as do CRAC and GSNAP) allows mapping many reads covering splice junctions, which BWA/Bowtie/SOAP2 cannot detect. However, the mapping results on categories of reads carrying one mutation, a short indel, or even errors indicate that classical mappers missed between $[15 – 20]$ 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 $[10 – 20]\%$ of reads containing a splice junction. This can impact negatively 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/TopHat-fusion as potentially containing a junction, and they will not find a spliced alignment for it. To contrast, CRAC’s \( k \)-mer profiling approach is flexible, reliable in this respect (Figure 3), and importantly, adapts well to longer reads (e.g., 200 nt). This last point is a key since longer reads will be available soon. They much more likely incorporate not one, but several events - errors, mutations, splice junctions, etc - and are thus harder to map. Whatever the prediction to perform, CRAC always improves its sensitivity with longer reads. This is crucial for detecting multiple exons within single reads, and CRAC exhibits a better ability in this matter 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 [22]. In this matter, our results (Table 1b) demonstrate that \( k \)-mer profiling approaches (MapSplice, CRAC) profit greatly in sensitivity from longer reads, and that CRAC stands up as the tool providing the highest precision whatever the read length. They also indicate that it handles more sensitively difficult cases, like long distance splice, multi-exon reads, or RNA expressed at low level. The analysis of a multi-tissue library shows that CRAC, GSNAP, and MapSplice obtain 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). By contrast, TopHat misses 21% of these known RefSeq junctions. Comparatively, CRAC reports less "novel" or "unknown" junctions than other tools, and tends to be more "conservative", which likely reflects its precision. Altogether, CRAC provides a solution to explore qualitatively the transcriptome of a sample with high sensitivity and precision, and thus gives the primary material to determine all transcript structures, which is indispensable for estimating the expression levels of all RNA isoforms afterwards [3, 24].

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, 25, 26, 33, 34]. Chimeric RNAs (chRNAs) represent the most difficult and error-prone computational prediction 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 FP. 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 much chRNA candidates compared to CRAC (Table 2). 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 candi-
dates, 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 as 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 [25, 34, 35]. Notably, CRAC predicted on four breast cancer libraries 36 recurrent chRNAs that were not reported previously [16, 25], and 35/36 harbor always the same junction point in 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 [29] or HSPD1 [28]. 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 LONP1 gene, which recurs in normal and tumoral libraries. We also reported cases of chRNAs, that 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, displacement of a mobile element) in DNA. Thus, CRAC may also prove useful for genomic analyses.

**List of abbreviations**

BT-474, KPL-4, MCF-7, SK-BR-3: names of four breast cancer libraries; chRNA: chimeric RNA; FL: false location; indel: insertion or deletion; KJ, NJ, and OJ: known, new and other splice junctions (respectively); mRNAs: messenger RNAs; NGS: Next Generation Sequencing; nt: nucleotide; resp.: respectively; RNA: RiboNucleic Acid; RNA-seq: RNA sequencing; SJ: splice junctions; SNP: Single Nucleotide Polymorphism; SNV: Single Nucleotide Variant (a biological variation); THF: TopHat Fusion; TP/FP: True or False Positive.

**Competing interests**

The author(s) 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 softwares’ comparisons and evaluations, as well as the analysis of real data sets. NP and MS performed the experiments. NP and MS prepared the figures, ER wrote the manuscript with contributions of all authors. ER and TC supervised the research. All authors read and approved the final manuscript.

Acknowledgments
The authors thank Alban Mancheron for help in packaging CRAC software, Gwenael Piganeau for critical reading of 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 the 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 http://www.atgc-montpellier.fr/ngs.

Endnotes
1 TopHat-fusion without the extra post-filtering script.
2 If TopHatFusion-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.
3 Only intergenic chRNAs are counted here.
6 These authors contributed equally to this work.
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, 461:206–11.
5. Mitelman F, Johansson B, Mertens F: Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer. Nat. Rev. Cancer 2007, 7.
6. Rabbitts TH: Commonality but Diversity in Cancer Gene Fusions. Cell 2009, 137:391–395.
7. 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.
8. 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.
9. Li H, Durbin R: Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2010, 26:589–595.
10. Li R, Li Y, Kristiansen K, Wang J: SOAP: short oligonucleotide alignment program. Bioinformatics 2008, 24:713–714.
11. 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.
12. Langmead B, Trapnell C, Pertea G, Mortazavi A, Pachter L: Ultrafast and memory-efficient alignment of RNA-Seq reads 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. Philippe N, Salson M, Lecroq T, Leonard M, 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.
18. Ferragina P, Manzini G: Opportunistic data structures with applications. In Proc. of FOCS 2000:390–398.
19. 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.
26. 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.

27. 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.

28. 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.

29. Wood LD, Parsons DW, Jones S, Lin J, Sjoblom T, Leary RJ, Shen D, Boca SM, Barber T, Ptak J, Silliman N, Szabo S, Dezsö Z, Ustyanovsky 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, Ballinger DG, Sparks AB, Hartigan J, Smith DR, Suh E, Papadopoulos N, Buckhaults P, Markowitz SD, Parmigiani G, Kinzler KW, Velculescu VE, Vogelstein B: The Genomic Landscapes of Human Breast and Colorectal Cancers. Science 2007, 318:1108 –1113.

30. 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. Science translational medicine 2011, 3:92ra66.

31. 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 DF, 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, Seshagiri S: Diverse somatic mutation patterns and pathway alterations in human cancers. Nature 2010, 466:869–873.

32. 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 Ribonucleoproteins hnRNP L, hnRNP A1, and hnRNP M. J. Biol. Chem. 2011, 286:16039 –16051.

33. 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.

34. 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.

35. Houseley J, Tollervey D: Apparent Non-Canonical Trans-Splicing Is Generated by Reverse Transcriptase In Vitro. PLoS ONE 2010, 5:e12271.

Description of additional files

File 1 Supplementary Figure 1: Read classification performed by CRAC

File 2 Supplementary Methods: additional description of CRAC’s algorithm, of the simulation of RNA-seq data, of the tools used for comparison.

File 3 Supplementary Results: Results on simulated RNA-seq data.

File 4 Supplementary Results: Results on real RNA-seq data.

File 5 Supplementary Table: Chimeric RNAs predicted on four breast cancer libraries.
**Figures**

**Figure 1 - CRAC’s 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, *e.g.* a 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 (*e.g.*, a mutation); it drops down 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 *e.g.* the correct splice junction boundaries) and avoid detecting a false chimera (Rule 2a) instead of a deletion.

**Figure 2 - Comparison of mapping results by category for 7 tools.**

Mapping by category of events on simulated RNA-seq against the Human genome on datasets with short and long reads (left 42M-75nt; right 48M-200nt) by six different mapping tools: Bowtie, Bowtie2, BWA/BWA-SW, CRAC, GASSST, GSNAP, and SOAP2. We consider 6 categories of reads depending on whether they contain a Single Nucleotide Variant (SNV), an insertion, a deletion, a junction, a sequence error, or a chimeric splice junction (chimera). In each category, the bar reports the percentage of those reads mapped at a unique location by the corresponding tool. The red tip at the top of the bar represents the percentage of incorrectly mapped reads. With 75 nt reads, CRAC departs from all other tools for constantly reaching > 90% of sensitivity and > 95% of precision whatever the category. Most other tools except GSNAP are below 50% sensitivity for mapping reads in categories where spliced alignments are needed (for which they are not intended for) 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/Bowtie2 increases in all categories. Compared to short reads, other tools achieve a better mapping of insertion/deletion containing reads.
Figure 3 - *Sensitivity and precision* of CRAC predictions by classes of events - Single Nucleotide Variant (SNV), insertion, deletion, splice junction, sequence errors, and chimeric splice junctions (chimera) - on Human simulated data.

(A) Absolute numbers of true and false positives reported by CRAC. These figures count the number of distinct, say SNV, 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, proportion of events found by CRAC for the 75 and 200 nt datasets. The blue bar represents the true positives, while the red bar on top of it represents false positives. The height of blue bar gives the sensitivity of CRAC, the relative height of the red part inside the bar gives the precision. Between the two read lengths, for all categories the sensitivity increases with longer reads, while the precision in each category varies little.

Figure 4 - Splice junction detection on Human real RNA-seq by 4 tools: comparison and agreement.

Splice junction detections by CRAC, MapSplice, TopHat, and GSNAP on a Human six tissue RNA-seq library of 75 M 100 nt reads (ERR030856). (4a) Number and percentage of known, new and other splice junctions (SJ) detected by each tool with +/- 3 nt tolerance on ERR030856. (4b) Venn diagram showing the agreement among tools on known RefSeq splice junctions (KJ); Additional file 4 shows the pendant for novel junctions (NJ) and RefSeq transcripts. (4c) A read spanning four exons (2−5) and three splice junctions of human TIMM50 gene; display from the UCSC genome browser. The included exons, number 3 and 4, measure 32 and 22 nucleotides, respectively. So exon 3 has exactly the $k$-mer size used in this experiment.
Table 1: **Global mapping sensitivity and precision of 7 tools.** Comparison of sensitivity and precision of different tools on human simulated RNA-seq (42M-75nt; 48M-200nt) against the Human genome for (1a) mapping, (1b) splice junction prediction, and (1c) chimeric junction prediction. Sensitivity gives the percentage of correctly reported cases over all sequenced cases, while precision gives the percentage of correct cases among all reported cases. Values in bold indicate the maximum of a column, and those in italics the second highest values. For all tasks on current read length, CRAC combines good sensitivity and very good precision. Importantly, CRAC always improve sensitivity with longer reads, and then offers the best performance over all solutions even against tools that are specialized for that task (e.g., BWA for mapping or TopHat for splice prediction). TopHat-fusion could not process 200 nt reads. The pendant results for the drosophila datasets are presented in Additional file 3 and show similar trends.
Table 2: **Chimeric RNA detection in breast cancer libraries.** Numbers of chimeric splice junctions (also termed chimeric RNA, for short chRNA) and spanning reads found by CRAC and TopHat-fusion in four breast cancer libraries (BT-474, KPL-4, MCF-7, and SK-BR-3 [25], Additional file 4). The figures are shown for the raw results (columns (2, 3) for CRAC and (6, 7) for TopHat-fusion), as well as after filtering for possible suboptimal continuous alignments with GSNAP (columns (4, 5) for CRAC and (8, 9) for TopHat-fusion). TopHat-fusion reports $\simeq 200$ times more raw candidates than CRAC, this ratio increases after the filtering. Confrontation with the set of validated chRNAs by Edgren et al. [25] shows that both the sets filtered and unfiltered predictions of CRAC and TopHat-fusion include respectively 20 and 21 true chRNAs and they agree of 17 of them.

| Edgren libraries | CRAC |                       | TopHat-fusion |                       |
|------------------|------|-----------------------|---------------|-----------------------|
|                  | raw  | after GSNAP           | raw           | after GSNAP           |
|                  | # chRNA | # reads  | # chRNA | # reads  | # chRNA | # reads  | # chRNA | # reads  |
| BT-474           | 692  | 9661  | 153    | 460  | 109711 | 349801 | 81327  | 189523  |
| KPL-4            | 407  | 5157  | 60     | 199  | 32412  | 98330  | 23075  | 53165   |
| MCF-7            | 466  | 3475  | 90     | 180  | 42738  | 121544 | 27267  | 57676   |
| SK-BR-3          | 703  | 9354  | 152    | 577  | 86249  | 241219 | 61494  | 130682  |

| Edgren libraries | CRAC |                       | TopHat-fusion |                       |
|------------------|------|-----------------------|---------------|-----------------------|
|                  | raw  | after GSNAP           | raw           | after GSNAP           |
|                  | # chRNA | # reads  | # chRNA | # reads  | # chRNA | # reads  | # chRNA | # reads  |
| BT-474           | 692  | 9661  | 153    | 460  | 109711 | 349801 | 81327  | 189523  |
| KPL-4            | 407  | 5157  | 60     | 199  | 32412  | 98330  | 23075  | 53165   |
| MCF-7            | 466  | 3475  | 90     | 180  | 42738  | 121544 | 27267  | 57676   |
| SK-BR-3          | 703  | 9354  | 152    | 577  | 86249  | 241219 | 61494  | 130682  |
Table 3: CRAC and TopHat-fusion predictions on the set of validated chimeric junctions from four breast cancer libraries. Confrontation with the set of validated chRNAs by Edgren et al. [25] shows that both the sets filtered and unfiltered predictions of CRAC and TopHat-fusion include respectively 20 and 21 true chRNAs and they agree of 17 of them. [25]. The first 20 cases are found by CRAC, and the 7 remaining ones are not predicted by CRAC, but for the last two, we could not detect any read matching the 15 − 20 nucleotides over junction. A short read length penalizes CRAC: indeed, with \( k = 22 \), only the 6 (= 50 − 2 * 22) middle positions of a read allow to locate exactly any event (splice, mutations). 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).

| Library | Fused genes | Chr. | 5’ position | str | 3’ position | str | Avg. support | CRAC | THF |
|---------|-------------|------|-------------|-----|-------------|-----|--------------|------|------|
| BT474   | SNF8 - RPS6KB1 | 17 - 17 | 47021337 | 1   | 57906864 | -1  | 36           | yes  | yes |
| BT474   | CMTM7 - GLB1  | 3 - 3  | 3248329    | -1  | 3305545   | 1   | 2            | yes  | yes |
| BT474   | SKA2 - MYO19  | 17 - 17 | 57232490 | -1  | 34863351  | -1  | 6            | yes  | yes |
| BT474   | ZMYND8 - CEP250 | 20 - 20 | 45852968 | -1  | 34078459  | 1   | 9            | yes  | yes |
| BT474   | VAPB - IKZF3  | 20 - 17 | 56964572 | 1   | 37934021  | -1  | 6            | yes  | yes |
| BT474   | ACACA - STAC2 | 17 - 17 | 35479452 | -1  | 37374427  | -1  | 46           | yes  | yes |
| BT474   | DIDO1 - TTI1  | 20 - 20 | 61569147 | -1  | 36634800  | -1  | 2            | yes  | yes |
| BT474   | RAB22A - MYO9B | 20 - 19 | 56886178 | 1   | 17256205  | 1   | 9            | yes  | yes |
| BT474   | MCF2L - LAMP1  | 13 - 13 | 113718616 | -1  | 113951811 | 1   | 9            | yes  | yes |
| KPL4    | ACACA - STAC2 | 9 - 9  | 139438475 | -1  | 134062675 | 1   | 2            | yes  | yes |
| KPL4    | BSG - NFIX    | 19 - 19 | 580782    | 1   | 13135832  | 1   | 9            | yes  | yes |
| MCF7    | RPS6KB1 - TMEM49 | 17 - 17 | 57992064 | 1   | 57917126  | 1   | 5            | yes  | yes |
| MCF7    | ARFGF2 - SULF2 | 20 - 20 | 47538548 | 1   | 46365686  | -1  | 10           | yes  | yes |
| SKBR3   | PKIA - RARA   | 8 - 17  | 79485042  | -1  | 38465537  | -1  | 7            | yes  | yes |
| SKBR3   | TATDN1 - GSDMB | 8 - 17  | 12551264  | -1  | 38066177  | -1  | 334          | yes  | yes |
| SKBR3   | KCNB1 - CSE1L | 20 - 20 | 47926856  | -1  | 47688990  | -1  | 6            | yes  | no  |
| SKBR3   | CYTH1 - EIF3H | 17 - 8  | 76778283  | -1  | 117768258 | -1  | 11           | yes  | yes |
| SKBR3   | SUMF1 - LRRFIP2 | 3 - 3  | 4418012   | -1  | 37170640  | -1  | 4            | yes  | yes |
| SKBR3   | SETD3 - CCDC85C | 14 - 14 | 99880273  | 1   | 10002353  | 1   | 3            | yes  | no  |
| SKBR3   | PCDH1 - ANKHD1 | 5 - 5  | 141234002 | NA  | 139825559 | -1  | 2            | yes  | yes |
| BT474   | CPNE1 - P13   | 20 - 20 | 34243123  | NA  | 43804501  | NA  | NA           | no   | yes |
| BT474   | STARD3 - DOK5 | 17 - 17 | 37793479  | NA  | 53259992  | NA  | NA           | no   | yes |
| SKBR3   | WDR67 - ZNF704 | 8 - 8  | 124096577 | NA  | 81733851  | NA  | NA           | yes  | yes |
| MCF7    | BCAS4 - BCAS3 | 20 - 17 | 49411707  | NA  | 59445685  | NA  | NA           | no   | yes |
| KPL4    | PPP1R12A - SEPT10 | 12 - 2 | 80211173  | NA  | 110343144 | NA  | NA           | no   | no  |
| SKBR3   | DHX35 - ITCH  | 20 - 20 | unknown   | NA  | unknown   | NA  | NA           | no   | no  |
| SKBR3   | NFS1 - PREX1  | 20 - 20 | unknown   | NA  | unknown   | NA  | NA           | no   | no  |
(a) Analysis of the location profile

(b) Analysis of the support profile

(c) Substitution

(d) Deletion

(e) Insertion

**Figure 1**
## (A)

|                  | SNV  | Ins. | Del. | Splices | Errors    | Chimera |
|------------------|------|------|------|---------|-----------|---------|
| **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     |

|                  | SNV  | Ins. | Del. | Splices | Errors    | Chimera |
|------------------|------|------|------|---------|-----------|---------|
| **True positives** | 18,670 | 1,641 | 1,637 | 67,372  | 10,324,528 | 624     |
| **False positives** | 609    | 57   | 26   | 340     | 37,660    | 41      |

## (B)

![Bar chart](chart_b.png)

## (C)

![Bar chart](chart_c.png)

---

**Figure 3**

### Table

|               | 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 |        |

### Figures

(a) Table showing percentages and counts of different RNA sequences.

(b) Venn diagram illustrating the overlap of transcripts identified by CRAC, MapSplice, and TopHat.

(c) Image showing a scale for genetic sequencing analysis.
Additional files provided with this submission:

Additional file 1: CRAC-sm-fig-classification.pdf, 37K
http://genomebiology.com/imedia/8136900249190848/supp1.pdf
Additional file 2: CRAC-sm.pdf, 487K
http://genomebiology.com/imedia/1047119690931948/supp2.pdf
Additional file 3: CRAC-sm-res-simul.pdf, 144K
http://genomebiology.com/imedia/1105260244919085/supp3.pdf
Additional file 4: CRAC-sm-res-real.pdf, 431K
http://genomebiology.com/imedia/4300356749190859/supp4.pdf
Additional file 5: CRAC-sm-BC-chimeras.ods, 328K
http://genomebiology.com/imedia/162582344931949/supp5.ods
Additional file 6: turing.png, 219K
http://genomebiology.com/imedia/8002384699360114/supp6.png
Additional file 7: crac_1.0.0_amd64.deb, 921K
http://genomebiology.com/imedia/1450401405851246/supp7.deb