Published online 5 October 2015 *Nucleic Acids Research, 2016, Vol. 44, No. 3* ***e29***
doi: 10.1093/nar/gkv1013

NCLscan: accurate identification of non-co-linear transcripts (fusion, *trans*-splicing and circular RNA) with a good balance between sensitivity and precision

Trees-Juen Chuang*, Chan-Shuo Wu, Chia-Ying Chen, Li-Yuan Hung, Tai-Wei Chiang and Min-Yu Yang

Division of Physical and Computational Genomics, Genomics Research Center, Academia Sinica, Taipei 11529, Taiwan

Received May 20, 2015; Revised September 23, 2015; Accepted September 24, 2015

ABSTRACT

Analysis of RNA-seq data often detects numerous 'non-co-linear' (NCL) transcripts, which comprised sequence segments that are topologically inconsistent with their corresponding DNA sequences in the reference genome. However, detection of NCL transcripts involves two major challenges: removal of false positives arising from alignment artifacts and discrimination between different types of NCL transcripts (*trans*-spliced, circular or fusion transcripts). Here, we developed a new NCL-transcript-detecting method ('NCLscan'), which utilized a stepwise alignment strategy to almost completely eliminate false calls (>98% precision) without sacrificing true positives, enabling NCLscan outperform 18 other publicly-available tools (including fusion- and circular-RNA-detecting tools) in terms of sensitivity and precision, regardless of the generation strategy of simulated dataset, type of intragenic or intergenic NCL event, read depth of coverage, read length or expression level of NCL transcript. With the high accuracy, NCLscan was applied to distinguishing between *trans*-spliced, circular and fusion transcripts on the basis of poly(A)- and nonpoly(A)-selected RNA-seq data. We showed that circular RNAs were expressed more ubiquitously, more abundantly and less cell type-specifically than *trans*-spliced and fusion transcripts. Our study thus describes a robust pipeline for the discovery of NCL transcripts, and sheds light on the fundamental biology of these noncanonical RNA events in human transcriptome.

INTRODUCTION

Technological advances in transcriptome sequencing have enabled biologists to better understand transcriptomes in a global manner. Comparative transcriptomics through the lens of high-throughput RNA sequencing (RNA-seq) has discovered a considerable number of 'non-co-linear' (NCL) transcripts, which are composed of sequence segments that are topologically inconsistent with the corresponding DNA sequences in the reference genome. An NCL transcript may consist of sequence segments either from a single gene but ordered inconsistently with the reference genome (i.e., intragenic NCL transcripts) or from two or more non-adjacent genes with distant regions of DNA (i.e., intergenic NCL transcripts), by means of genetic rearrangements or post-transcriptional events. Genetic rearrangement events may generate fusion transcripts/genes at the DNA level, while post-transcriptionally NCL transcripts (PtNCL transcripts) are produced during post-transcriptional RNA processing (e.g., *trans*-splicing or *cis*-backsplicing). Fusion and *trans*-spliced transcripts are usually called 'chimeric transcripts' or 'chimeras' because both of them comprise of sequence segments from different genes or precursor mRNAs (pre-mRNAs) (1). Figure 1 summarizes the categorization of NCL transcripts in transcriptome.

Fusion genes have been reported to be associated with malignant hematological disorders and sarcomas (2–5). The most prominent example is the *BCR-ABL1* fusion gene, which is a key factor of adult acute lymphoblastic leukemia cases and an effective biomarker for chronic myeloid leukemia (6–9). Other examples, such as *ETV6-NTRK3* in breast carcinoma (10), *BCAS4-BCAS3* in breast cancer (11), *TMPRSS2-ERG/ETS* in prostate cancer (12,13), *EML4-ALK* in lung cancer (14), *MYB-NFIB* in head and neck tumors (15) and *VTI1A-TCF7L2* in colorectal cancer (16), emphasize the critical importance of fusion genes in cancer detection and diagnosis.

PtNCL transcripts, as mentioned above, can be generated through *trans*-splicing or *cis*-backsplicing. *Trans*-splicing can take place between separate pre-mRNAs either of a single gene (i.e., intragenic *trans*-splicing) or among different genes (i.e., intergenic *trans*-splicing) (17,18). Two promi-

*To whom correspondence should be addressed. Tel: +886 2 27871244; Fax: +886 2 27899923; Email: trees@gate.sinica.edu.tw

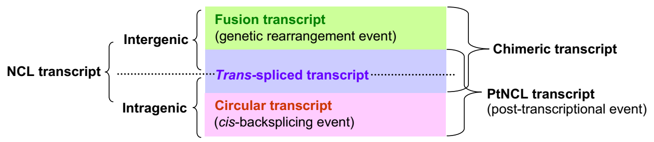
NCL transcript
Intergenic
Intragenic
Fusion transcript
(genetic rearrangement event)
Trans-spliced transcript
Circular transcript
(*cis*-backsplicing event)
Chimeric transcript
PtNCL transcript
(post-transcriptional event)

Figure 1. Categorization of NCL transcripts.

nent examples of intergenic *trans*-spliced RNAs are *JAZF1-SUZ12* and *SLC45A3-ELK4*, which have been reported to be associated with anti-apoptotic function (1,19,20) and prostate cancer (1,21), respectively. In addition, ts*RMST*, which is produced via intragenic *trans*-splicing, plays a role in pluripotency maintenance of human embryonic stem cells, as evidenced by our recent discovery (22). On the other hand, *cis*-backsplicing occurs within a single pre-mRNA, leading to the formation of circular RNA (circRNA) (23,24). CircRNAs have been observed in diverse species (25–28), some of them are evolutionarily conserved between species (26,28–32). *CDR1as/ciRS-7* and circRNA of *Sry*, two representative examples of circRNA in mammals, were demonstrated to function as microRNA sponges (29,33). In addition, circRNAs may play regulatory roles during aging of the central nervous system (31) or cell proliferation (34), and may store, sort or localize RNA-binding proteins (29,35). These findings suggest that circular RNA is an ancient, essential and fine-tuned member in the roster of functional transcripts.

Computational strategies for the identification of NCL transcripts have been proposed to facilitate RNA-seq-based studies (36–42) and have enabled the discovery of thousands of NCL transcript candidates in diverse species (43–53). However, false positives arising from sequencing or alignment errors appear to be unavoidable (48,54,55). Discrepancies among the NCL transcript candidates identified by different strategies imply that a large proportion of candidates may be spurious (40,56–58). Similarities among paralogous genes or repetitive sequences often lead to ambiguities during short-read mapping, which are often misinterpreted as NCL events (59). It remains a major challenge to effectively eliminate such false calls and detect genuine ones without losing sensitivity. Also, most previous studies focus on identifying either intergenic or intragenic NCL events. There thus remains a need for a robust pipeline capable of identifying both intergenic and intragenic NCL transcripts with high sensitivity and precision from RNA-seq data.

To address these issues, we developed a new method, NCLscan, which achieves the goal through a series of knowledge-based processes and integrates different mapping algorithms to enable the stepwise elimination of spurious events without losing sensitivity. NCLscan was found to be superior to 18 other currently available tools in terms of sensitivity and precision, when used to detect intragenic or intergenic NCL events in several benchmark datasets, including real and simulated datasets with different strategy choices for generating simulated dataset, depths of coverage, read lengths and expression levels of NCL transcripts. With the high accuracy, NCLscan was applied to distinguishing between different types of NCL transcripts (i.e., *trans*-spliced, circular or fusion transcripts) and then investigating these three types of NCL transcripts in terms of the prevalence, expression level and expression breadth by comparing the composition and populations of NCL transcripts between RNA samples with different treatments (i.e., poly(A)- and nonpoly(A)-selected RNA-seq data) from diverse human cell types. This comparative analysis revealed some endogenous features of NCL transcripts, which allowed us to distinguish between circular RNAs, *trans*-splicing events and fusion transcripts. NCLscan promises to facilitate the comprehensive characterization of various types of NCL transcripts on a transcriptome-wide scale.

MATERIALS AND METHODS

Data retrieval and availability

The human genomic sequences (hg19/GRCh37 assembly) and annotated transcripts were downloaded from the GENCODE project (version 19) at http://www.gencodegenes.org/releases/19.html. Datasets A and C (see Table 1) were downloaded from the website of the FusionMap group at http://www.arrayserver.com/wiki/index.php?title=FusionMap (60) and from the NCBI Sequence Read Archive at http://www.ncbi.nlm.nih.gov/sra under the accession number SRP003186 (61), respectively. Dataset B (negative dataset; Table 1) was provided from the author of (57). The links and the used parameters of the tools examined in this study were listed in Supplemental Table S1. The simulated datasets generated in this study are available from our ftp site at ftp://treeslab1.genomics.sinica.edu.tw/NCLscan. The poly(A)-/nonpoly(A)-selected RNA-seq (see Table 2) and SOLiD-based DNA-PET (1, 10 and 20 kb libraries) data were downloaded from the ENCODE consortium (62) at https://www.encodeproject.org/. NCLscan was implemented under the Bio-Linux (Ubuntu 14.04) or Mac operating systems on a 64-bit machine with ≥10 GB RAM. The NCLscan program, document and test dataset are publicly accessible from GitHub at https://github.com/TreesLab/NCLscan or our FTP site at ftp://treeslab1.genomics.sinica.edu.tw/NCLscan.

The NCLscan pipeline

As shown in Figure 2A, NCLscan first aligned RNA-seq reads against the human reference genome (GRCh37) and annotated transcripts (GENCODE version 19) using BWA (63) with default parameters. The unmapped reads were aligned against the genome and transcripts using Novoalign (Novocraft Technologies) with the parameters: -r A 1 -n 30. Both ends of the unmapped paired-end reads were then joined end to end, to form concatenated sequences. These concatenated sequences were aligned against the reference genome using BLAT (64) with default parameters. The concatenated sequences with two split sequence segments that were linearly inconsistent with the reference genome were retained. The pipeline did not consider read-through gene fusions. To this end, concatenated sequences in which the distance between two split segments on the same strand of the same chromosome was <2 Mbp were discarded. In addition, a concatenated sequence satisfied one of the two criteria was not considered: (i) the concatenated sequence mapped to an unplaced (undetermined)

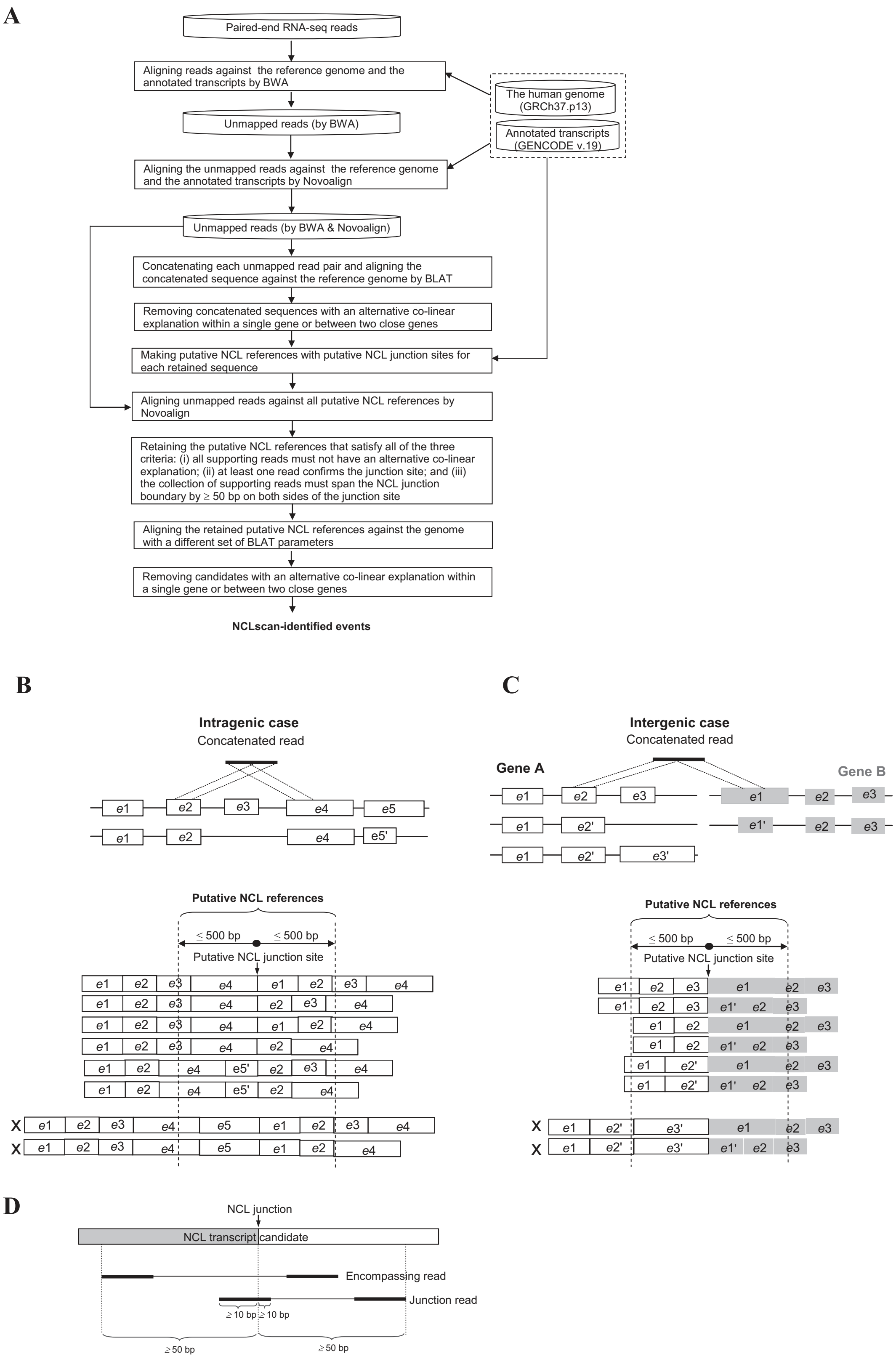
A
Paired-end RNA-seq reads
Aligning reads against the reference genome and the annotated transcripts by BWA
The human genome (GRCh37.p13)
Annotated transcripts (GENCODE v.19)
Unmapped reads (by BWA)
Aligning the unmapped reads against the reference genome and the annotated transcripts by Novoalign
Unmapped reads (by BWA & Novoalign)
Concatenating each unmapped read pair and aligning the concatenated sequence against the reference genome by BLAT
Removing concatenated sequences with an alternative co-linear explanation within a single gene or between two close genes
Making putative NCL references with putative NCL junction sites for each retained sequence
Aligning unmapped reads against all putative NCL references by Novoalign
Retaining the putative NCL references that satisfy all of the three criteria: (i) all supporting reads must not have an alternative co-linear explanation; (ii) at least one read confirms the junction site; and (iii) the collection of supporting reads must span the NCL junction boundary by ≥ 50 bp on both sides of the junction site
Aligning the retained putative NCL references against the genome with a different set of BLAT parameters
Removing candidates with an alternative co-linear explanation within a single gene or between two close genes
NCLscan-identified events
B
Intragenic case
Concatenated read
e1 e2 e3 e4 e5
e1 e2 e4 e5'
Putative NCL references
≤ 500 bp ≤ 500 bp
Putative NCL junction site
X
X
C
Intergenic case
Concatenated read
Gene A
Gene B
e1 e2 e3 e1 e2 e3
e1 e2' e1' e2 e3
e1 e2' e3'
Putative NCL references
≤ 500 bp ≤ 500 bp
Putative NCL junction site
X
X
D
NCL junction
NCL transcript candidate
Encompassing read
Junction read
≥ 10 bp ≥ 10 bp
≥ 50 bp
≥ 50 bp

Figure 2. Identification of NCL transcripts. (**A**) Flowchart depicting the NCLscan pipeline. (**B** and **C**) Schematic illustrations of possible 'putative NCL references' with putative NCL junction sites (based on BLAT alignment output and GENCODE annotation): (B) intragenic case and (C) intergenic case. The characters 'e' and 'X' represent 'exon' and the putative NCL reference that is not considered, respectively. (**D**) Schematic illustration of a retained putative NCL reference, which satisfies all of the criteria listed in (A).

contig; and (ii) the donor (or acceptor) side of the concatenated sequence mapped to more than one positions with similar BLAT mapping scores (the score difference between matches <5). Subsequently, all possible 'putative NCL references' with putative NCL junction sites were made according to the information of BLAT alignment output and GENCODE annotation. In a retained sequence, the NCL junction site had to locate at the splicing junction of an annotated exon and the corresponding putative NCL reference had to be no longer than the read length with insert sizes (e.g., 1000 bp in this study; Figure 2B and C). Reads that could not be mapped to the genome or annotated transcripts were aligned against these putative NCL references using Novoalign with the following parameters: -r A 1 -g 99 -x 99. There are two types of read that match to a putative NCL reference: encompassing reads and junction reads. A putative NCL reference was retained if it satisfied all of the following criteria: (i) all reads that match to the putative NCL reference must not have an alternative co-linear explanation within a single gene or between two close genes; (ii) at least one junction read supports the putative NCL junction of the putative NCL reference, which must span the NCL junction boundary by ≥10 bp on both sides of the junction site; and (iii) the collection of supporting reads (including encompassing and junction reads) must span the NCL junction boundary by ≥50 bp (default value; the span range is a user-assignable parameter) on both sides of the junction site (Figure 2D). Of note, the default value was set because the read lengths of most currently-available RNA-seq reads were >50 bp. Finally, since different BLAT parameters may generate different alignment results, the retained NCL references were aligned against the reference genome using BLAT with a different set of parameters (-titleSize = 9 -stepSize = 9 -repMatch = 32768). Only the NCL transcript candidates that did not contain alternative co-linear explanations were reported in the final output.

Generation of simulated datasets

Simulated NCL transcripts were selected from the annotated genes (GENCODE version 19). The genes would be considered to be the source of simulated NCL transcripts only when they satisfied both of the following criteria: (i) they must not be in the blacklist provided by FusionMap (e.g., pseudogenes, mitochondrial or ribosomal genes) (60); (ii) for intergenic NCL events, the gene pairs must not belong to the same gene families; and (iii) the distance between paired genes on the same strand of the same chromosome must not be <2 Mbp. The junction site in each simulated NCL transcript was random and located at the boundary of an annotated exon. In addition, the length of simulated NCL transcripts should be ≥500 bp and the upstream and downstream sequences should be longer than 100 bp (59). After the simulation process, we obtained 100 intragenic NCL transcripts and 100 intergenic NCL transcripts. The co-linear transcripts were obtained from the annotated protein-coding genes which met two criteria: (i) possessed a status of 'known'; and (ii) ≥300 bp in length. Based on the co-linear transcripts and the 200 simulated intragenic/intergenic NCL transcripts, we used Mason v.0.1.2 (65) to generate paired-end RNA-seq reads (INS = 170 bp, SD = 20) with different depths of coverage (5-, 10-, 20- and 50-fold) and different read lengths (2 × 50, 2 × 100 and 2 × 150 bp). In addition, to evaluate the performance and accuracy of NCL transcript identification tools under different expression levels, we also generated paired-end RNA-seq reads from the 200 intragenic/intergenic NCL transcripts with 5-, 10-, 20-, 50-, 100-, 150- and 200-fold expression levels, and then mixed these simulated reads with a RNA-seq dataset (used as background data) generated from the co-linear transcripts. Of note, the Mason parameters were the same as those used in the Bowtie2 paper (66).

Accuracy measurement of NCL-transcript identification tools

Sensitivity (S_n), precision (S_p) and $F1$ values are defined as follows:

$$S_n = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

$$S_p = \frac{\text{TP}}{\text{TP} + \text{FP}} \text{ and}$$

$$F1 = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}}.$$

TP (true positive), FP (false positive) and FN (false negative) represent the number of correctly identified NCL events, the number of incorrectly identified NCL events and the number of missing NCL events, respectively. S_n, S_p, and $F1$ values all range from 0 to 1; higher value indicates higher level of accuracy. In consideration of mapping criteria of each tool, for each simulated or validated fusion transcript, a true positive hit was recorded when the distance between the correct NCL junction site and junction site identified by the tested tool was <10 bp (59).

Detection of genomic structure variations

To examine whether certain identified circRNAs were potentially derived from SVs, we used SVDetect (67), which can identify SVs from paired-end/mate-pair NGS data generated by the SOLiD platform, to detect SVs based on the K562 DNA-PET data from the ENCODE project. If an identified circRNA overlapped any breakpoint regions ±1000 bp of a detected SV, such an intragenic event was regarded as a potential SV-derived NCL event. Of the 8915 circRNA candidates detected in K562 cells, 594 were identified as potential SV-derived NCL events (Supplemental Table S2).

RESULTS

Identification of NCL transcripts

NCLscan identifies NCL transcripts from paired-end RNA-seq data. The overall schematic of NCLscan is depicted in Figure 2A. Firstly, co-linearly matched reads are eliminated by mapping the RNA-seq reads against the reference genome and well-known transcripts (i.e., annotated coding and non-coding transcripts in GENCODE) using a stepwise alignment strategy, which integrates two read-mapping algorithms ('Materials and Methods' section; Figure 2A). Next, the two ends of each unmapped read are concatenated to generate a continuous sequence. Each concatenated sequence is then aligned against the reference genome using BLAT (64). Only concatenated sequences that contain NCL segments are retained after this step ('Materials and Methods' section). For each retained sequence, we make 'putative NCL references' with putative NCL junction sites (Figure 2B and C) on the basis of the corresponding BLAT alignment result and GENCODE annotation. Of note, we only consider candidates in which splice junctions agree to well-known junction sites, because such candidates are more reliable than those with junction sites not matching exon boundaries (22,32,43,52). All possible combinations of transcript isoforms are considered while making the putative NCL references. To avoid the putative NCL references inferred from reads with an abnormal inner size, we limit the putative NCL references to 1000 bp in length (Figure 2B and C). Subsequently, reads that cannot be mapped to the genome or transcriptome (i.e., unmapped reads in Figure 2A) are aligned against the putative NCL references. There are two types of reads that can be mapped to a putative NCL reference: 'encompassing read', which connects two parental transcript segments but doesn't support the NCL junction site, and 'junction read', which overlaps the junction site. A putative NCL reference will be retained if it satisfies all of the following criteria: (i) all reads that match to the putative NCL reference must not have an alternative co-linear explanation, (ii) at least one junction read supports the putative NCL reference and (iii) the NCL junction boundary must be supported by read evidence (including encompassing and junction reads) spanning across 50 bp on both sides of the junction site (Figure 2D). Of note, the third rule is used to eliminate skew mapping between reads and the corresponding putative NCL reference. Since the use of BLAT-alignments with different sets of parameters could be more effective at detecting possible co-linear explanations of an expressed sequence than single operation with default parameters (32), the retained references are aligned against the reference genome using a different set of BLAT parameters ('Materials and Methods' section) to further remove potential false calls. Only the candidates that meet all the requirements are considered as high-confidence NCL transcripts, which are used in the following analyses.

Sensitivity and precision of NCLscan for intergenic NCL transcript detection

We first assess the accuracy of NCLscan at detecting intergenic NCL transcripts based on two simulated datasets (i.e., Datasets A (60) and B (57,58)), which were generated by different methods. Dataset A, which is a semi-artificial dataset generated by the FusionMap group (60), contains a set of artificial 75-bp paired-end reads generated according to 50 simulated fusion transcripts and a background set of real RNA-seq reads that is not expected to harbor any fusion events and has been applied to evaluation of fusion-identification tools (57–60). Dataset B is an artificial dataset provided by Carrara *et al.* (57,58), in which 100-bp paired-end reads were built by BEERS (68) based on well-annotated co-linear transcripts. Dataset B is regarded as a negative dataset, as it does not contain any fusion events. Thus, it is particularly useful for assessing the number of false calls reported by fusion-identification tools (57,58). With respect to the size of each dataset, Dataset A is a small collection of 0.057 million reads (Table 1), which can be used to efficiently evaluate the accuracy of fusion-detecting tools. In contrast, Dataset B, which contains 70 million reads, is much closer to real datasets (Table 1).

We used these two simulated datasets to compare NCLscan with 15 publicly-available tools, including BreakFusion (69), ChimeraScan (70), CRAC (71), deFuse (38), FusionAnalyser (41), FusionFinder (37), FusionHunter (72), FusionMap (60), FusionQ (73), MapSplice2 (74), STAR (75), segemehl (76), SOAPfuse (59), TopHat-Fusion (36) and TopHat-Fusion-post (36). All tools for evaluation were used with default parameters or the parameters suggested by the authors (Supplemental Table S1). We then evaluated the performance of each tool based on sensitivity (S_n), precision (S_p) and $F1$ score ('Materials and Methods' section). For Dataset A, NCLscan exhibited the greatest precision and $F1$ score, and the third highest sensitivity ($S_n = 0.86$) among 16 tools (Figure 3A). It is worth noting that NCLscan did not identify any false positives ($S_p = 1$); furthermore, although STAR and TopHat-Fusion had better sensitivity than NCLscan, their advances in sensitivity were subject to the trade-off in precision ($S_p = 0.79$ for STAR; $S_p = 0.2$ for TopHat-Fusion) (Figure 3A). Moreover, although FusionHunter and ChimeraScan also exhibited 100% precision ($S_p = 1$) on Dataset A, they were not satisfying with respect to sensitivity (both $S_n < 0.4$) (Figure 3A). Overall, NCLscan exhibited the highest $F1$ score (Figure 3A), indicating its superiority in terms of sensitivity and precision.

Table 1. RNA-seq datasets for evaluating the accuracy of fusion-identification tools

Dataset	Dataset type	Read type	Number of reads		Reference
A	Simulated data	75-bp paired-end	57 209		(60)
B	Simulated data	100-bp paired-end	70 000 000		(57,58)
C[a]	Real data	50-bp paired-end	8 412 431	(MCF-7)	(61)
			6 800 166	(KPL-4)	
			13 515 132	(BT-474(1))	
			7 915 382	(BT-474(2))	
			9 048 352	(SK-BR-3(1))	
			9 097 152	(SK-BR-3(2))	

[a] The dataset includes RNA-seq reads from four cell lines: MCF-7, KPL-4, BT-474 and SK-BR-3.

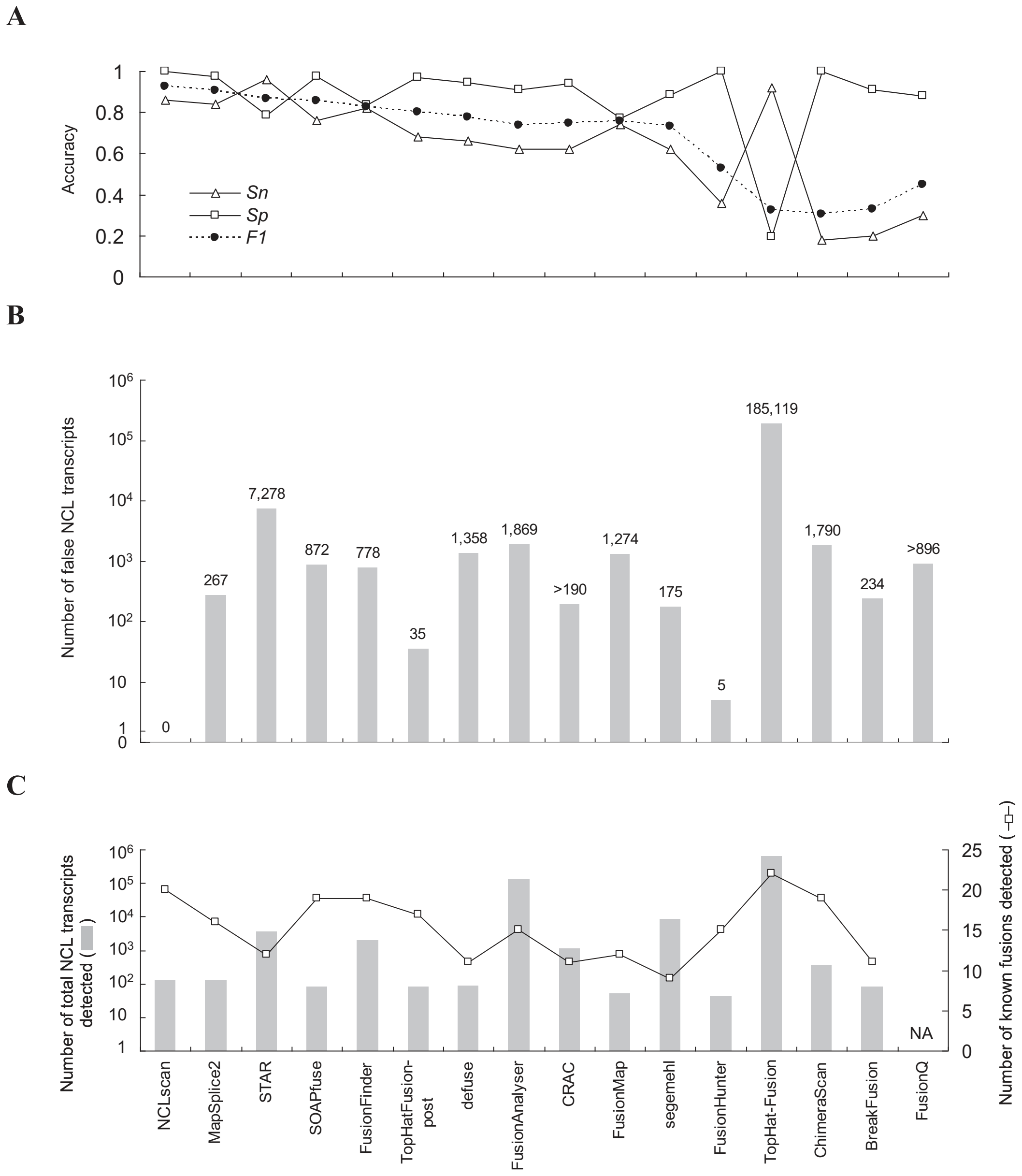

Figure 3. Comparison of the accuracy of 16 tools for detecting intergenic NCL transcripts based on (**A**) Dataset A, (**B**) Dataset B and (**C**) Dataset C. Datasets A and B are simulated datasets; Dataset C is a real dataset. The results for CRAC and FusionQ on Dataset B represent only one half and one fifth of the RNA-seq reads, respectively, for the reason that the two tools could not finish the calculation of the whole reads of Dataset B within 1 week. For the same reason, the result of FusionQ on Dataset C is not available. NA, not available.

For Dataset B, while most tools detected an enormous number of false NCL events (e.g., TopHat-Fusion detected >180 000 events; Figure 3B), NCLscan didn't report any false calls. Since Dataset B is a negative dataset, any NCL events detected in this dataset must be false positives. The result revealed that the existing fusion-detecting tools generally have severe problems with false positives, particularly when detecting fusion transcripts in a large dataset. In contrast, NCLscan still kept 100% precision on a large dataset, indicating the robustness of NCLscan for detecting NCL events with high accuracy. It should be noted that CRAC and FusionQ could not finish analyzing Dataset B after more than 1 week had elapsed, and so the results shown for the two tools were for only part of this dataset; even so, hundreds of false fusions were delivered by the two tools (Figure 3B).

Furthermore, we applied the 16 tools to a real RNA-seq dataset (Dataset C). This dataset contains 50-bp paired-end reads from four breast cancer cell lines (Table 1), in which 22 non-read-through fusion transcripts (e.g., the fusion events with a distance of ≤2 Mbp between paired genes on the same strand of the same chromosome were not considered) were experimentally validated by both RT-PCR and array comparative genomic hybridization (aCGH) (Supplemental Table S3) (61). Since the authenticity of the observed NCL events were often hampered by *in vitro* artifacts (22,32,45,77,78), the identified fusion events that have passed multiple experimental validations were more likely to be genuine than those did not. We found that NCLscan achieved high sensitivity, which was able to detect 20 out of the 22 fusion events (Figure 3C). *SKA2-MYO19* was not reported by NCLscan because of the ambiguous alignments with an alternative co-linear explanation or mapping to an unplaced (undetermined) contig. *LAMP1-MCF2L* was missed due to the skew mapping between reads and the corresponding putative NCL reference (Supplemental Figure S1A). Of note, if the smaller size of the span range (e.g., 30 bp) was set, *LAMP1-MCF2L* could be detected by NCLscan but the number of the identified NCL candidates would be increased, increasing the risk of accepting false positives (Supplemental Figure S1B). Although TopHat-Fusion detected all 22 known fusions, it should be noted that TopHat-Fusion reported as many as 630 812 candidates. In view of the high false-positive rates of TopHat-Fusion when applied to the simulated datasets (Datasets A and B; Figure 3A and B), users of this tool may need to make extra efforts to reduce potential false calls. NCLscan was also applied to another read RNA-seq data from the prostate cancer cell line VCaP and successfully detected all the three non-read-through fusion transcripts that were validated by both qRT-PCR and fluorescence *in situ* hybridization (FISH) (48) and collected in the ChiTARS database (79) (Supplemental Tables S3 and S4), further supporting the high level of sensitivity of NCLscan. On the other hand, some tools capable of detecting intergenic NCL transcripts also delivered a great number of intragenic NCL events from Dataset C (i.e., NCLscan, STAR, MapSplice2, CRAC, segemehl and TopHat-Fusion) (Supplemental Table S5). This suggests that there remains a considerable number of uncharacterized intragenic NCL transcripts (or circRNAs) awaiting further discovery within cancer samples.

Sensitivity and precision of NCLscan for simultaneously detecting intergenic and intragenic NCL transcripts

We emphasize that NCLscan is also capable of detecting intragenic NCL transcripts. To further evaluate the sensitivity and precision of NCLscan for simultaneously detecting intergenic and intragenic NCL transcripts, we applied NCLscan and eight other tools (STAR, MapSplice2, CRAC, segemehl, TopHat-Fusion, find_circ (29), CIRCexplorer (80) and CIRI (81)) to simulated RNA-seq datasets. Of note, find_circ, CIRCexplorer and CIRI were designed for detection of intragenic NCL transcripts (circRNAs) only. We used short-read simulator Mason (65) to generate artificial paired-end RNA-seq reads from the mix of the following transcripts: (i) 100 simulated intergenic NCL transcripts, (ii) 100 simulated intragenic NCL transcripts and (iii) well-annotated co-linear transcripts, and simulated a variety of data conditions of different read depths (5-, 10-, 20- and 50-fold) and of different read lengths (2 × 50, 2 × 100 and 2 × 150 bp) ('Materials and Methods' section). The identification results of these tools were reported in Supplemental Table S6. As expected, S_n values increased with increasing depth or length of reads, regardless of whether intragenic or intergenic NCL transcripts were detected (Figure 4A). In terms of sensitivity with shorter reads (e.g., 2 × 50 bp), NCLscan and TopHat-fusion were effective at detecting intragenic events, whereas MapSplice2 appeared to yield good sensitivity for detecting intergenic ones. For longer reads (e.g., 2 × 150 bp), NCLscan, STAR, TopHat-fusion, segemehl and the three circRNA-detecting tools (i.e., find_circ, CIRCexplorer and CIRI) represented comparable sensitivity for detecting intragenic events; MapSplice2 was more sensitive for detecting intergenic events than for detecting intragenic ones, whereas segemehl exhibited the reverse trend (Figure 4A). As for S_p, NCLscan reported zero false calls on all tested datasets under different simulated conditions (Supplemental Figure S2), yielding the highest precision (all $S_p = 1$) among the tested tools (Figure 4A). Obviously, NCLscan was much more precise than the tool (i.e., CIRCexplorer) that exhibited the second best precision (all $S_p < 0.72$; Figure 4A and Supplemental Figure S3A). With the exception of NCLscan, the S_p values appeared to be affected by read depth and length (Figure 4A), in which the number of false positives markedly increased with increasing depth of reads and marginally decreased with increasing length of reads (Supplemental Figure S2). In contrast to NCLscan with no false positives, all the other tested tools identified a considerable number of false positives (Supplemental Figure S2), which severely reduced the S_p values and then $F1$ scores under varied simulated conditions (Figure 4A). For example, although TopHat-fusion generally had higher S_n values than the other tools, it yielded the lowest S_p, resulting in a relatively poor balance between sensitivity and precision (all $F1$ scores < 0.1; Figure 4A). In general, NCLscan had comparable S_n values and the highest S_p values without compromising the false positive rate, exhibiting the best balance between sensitivity and precision on all tested datasets.

We further explored the effect of expression level on the sensitivity and precision of NCL transcript identification. We utilized Mason to generate paired-end reads from the aforementioned 100 simulated intragenic and 100 simulated intergenic NCL transcripts with different expression levels (5- to 200-fold), and then mixed these datasets with the same background dataset generated from co-linear transcripts ('Materials and Methods' section). The identification results of the examined tools were listed in Supplemental Table S6. For all examined tools, sensitivity generally increased with increasing expression level of NCL transcript; however, this tendency weakened when NCL transcript expression level reached 100-fold or greater (Figure 4B and Supplemental Figure S3B). In general, TopHat-fusion exhibited the highest S_n values, but it still possessed the lowest S_p and $F1$ values for all tested datasets. With the exception of TopHat-fusion, for intragenic events, NCLscan was more sensitive at lower expression levels, whereas NCLscan, STAR, segemehl and the three circRNA-detecting tools yielded comparable sensitivity at higher expression levels (Figure 4B). For intergenic event, while NCLscan, STAR, MapSplice2 and segemehl exhibited the comparable sensitivity for detecting events at lower expression levels, NCLscan, STAR and MapSplice2 appeared to be effective for detecting ones at higher expression levels (Figure 4B). These results indicated that our method exhibited compar-

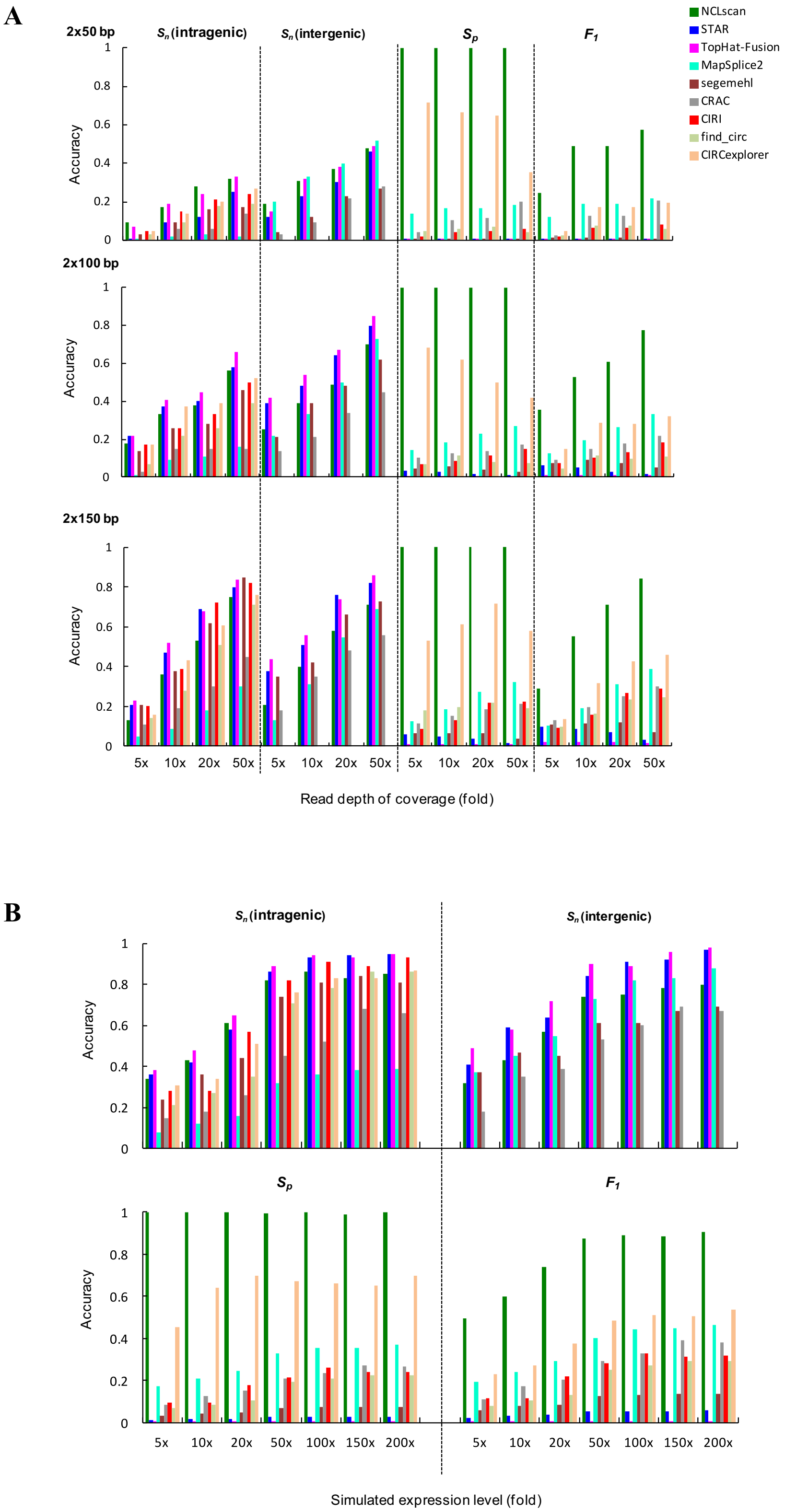

Figure 4. Evaluation of the sensitivity, precision and *F*1 score of NCLscan for simultaneous detection of intergenic and intragenic NCL transcripts based on simulated datasets of (**A**) different depths and lengths of reads, and (**B**) different expression levels (5- to 200-fold) of NCL transcripts.

ative sensitivity at all simulated expression levels. Importantly, while the other tools suffered poor precision, we emphasized that NCLscan maintained the highest S_p values and $F1$ scores on the tested datasets, regardless of the expression level of NCL transcripts (Figure 4B and Supplemental Figure S3B).

Regarding the NCLscan identification on the 19 simulated datasets (see Supplemental Table S6), the true events were not reported by NCLscan because of the following reasons (which were not mutually exclusive): (i) a low level of read depth or NCL transcript expression (ii) skew mapping (see also Supplemental Figure S1A); and (iii) ambiguous alignments with very close mapping scores between matches (see 'Materials and Methods' section; an example was given in Supplemental Table S7). On the other hand, NCLscan yielded three false positives on two of these simulated datasets (Supplemental Table S6). The reason was that the simulator (i.e., Mason) had randomly added a reasonable proportion of mismatches into the generated reads for simulating real reads (65). It was possible that a generated read matched to a false position with a higher mapping score than all the other possible alignments ('Materials and Methods' section), thus yielding false positives (Supplemental Table S7). We emphasized that NCLscan yielded only three false positives on these 19 simulated datasets, supporting its high level of precision.

Taken together, our results reveal that NCL transcript-identification tools are often susceptible to false-positive detections. In contrast, NCLscan, while retaining comparable sensitivity, significantly outperforms the other tools in terms of precision (all P-values $< 10^{-15}$ by the two-tailed Fisher's exact test), regardless of read depth, read length or NCL transcript expression level. NCLscan thus achieves the highest $F1$ scores on all simulated datasets, indicating its superior performance with a good balance between sensitivity and precision.

Discrimination between circRNAs, *trans*-spliced RNAs and fusion transcripts

We next applied NCLscan to NCL transcript identification using six real datasets, each of which contained both poly(A)- and nonpoly(A)-selected RNA-seq data (Table 2), from the ENCODE project (82). These datasets comprised of multiple types of cultured cell lines (including primary, immortalized and stem cells; Table 2). The poly(A)- and nonpoly(A)-selected RNA samples were generated by the same laboratory (82). The nonpoly(A)-selected samples were derived from ribo-and poly(A)-depleted RNA. It has been shown that most *trans*-spliced RNA products are polyadenylated, but circRNAs are not (25,28,32,78,80). Therefore, it is possible to classify the detected NCL transcript candidates into the following four groups based on their presence status in poly(A)- and nonpoly(A)-selected data (see also Supplemental Table S8).

Group I: intragenic events detected in poly(A)-selected data, which were expected to be intragenic *trans*-spliced RNAs.

Group II: intergenic events detected in poly(A)-selected data, which were expected to be intergenic *trans*-spliced RNAs or fusion transcripts.

Group III: intragenic events detected in nonpoly(A)-selected data, which were expected to be circRNAs.

Group IV: intergenic events detected in nonpoly(A)-selected data, which were expected to be fusion transcripts.

As shown in Figure 5A, we observed that NCL transcripts in Group III (predicted circRNA products) were more common than those in the other groups; this trend was independent of cell type and read depth of RNA-seq data. This suggests that circRNAs might be the predominant type of NCL RNA products. However, one might consider that some of the events in Group III could in fact be *trans*-spliced rather than circRNAs, due to the incomplete depletion of poly(A)-tailed RNAs. To address this, we calculated the ratio of the number of intergenic events to that of intragenic ones in each sample. Based on the facts that events in Groups II and IV were definitely not circRNAs and that both Groups I and II were highly probable to be *trans*-splicing events, we speculated that if *trans*-splicing events contributed highly to the candidates detected in the nonpoly(A)-selected samples, the quantity ratio of intragenic to intergenic events in nonpoly(A)-selected sample (i.e., Group III versus Group IV) should be comparable to the ratio in poly(A)-selected sample (i.e., Group I versus Group II). However, this analysis showed that the quantity ratio of intragenic-to-intergenic events was much higher in nonpoly(A)-selected samples than in poly(A)-selected ones (all P-values $< 10^{-15}$ by the two-tailed Fisher's exact test;

Table 2. Poly(A)- and nonpoly(A)-selected RNA-seq data used in this study

Sample	Description	Biosample type	Sex (life stage)	Poly(A)-selected data[a]	Nonpoly(A)-selected data[a]
H1	Human embryonic stem cell	stem cell	Male (embryonic)	SRR307911, SRR307912	SRR307923, SRR307924
K562	Chronic myelogenous leukemia	immortalized cell	Female (adult)	SRR315336, SRR315337	SRR307930, SRR307931
BJ	Skin fibroblast	immortalized cell	Male (newborn)	SRR307903, SRR307904	SRR317065
HepG2	hepatocellular carcinoma	immortalized cell	Male (child)	SRR307926, SRR307927	SRR307913, SRR307914
HUVEC	Umbilical vein endothelial cell	primary cell	Male (newborn)	SRR307905, SRR307906	SRR317067
NHEK	Epidermal keratinocytes	primary cell	Female (unknown)	SRR315327	SRR315321, SRR315322

[a] The RNA-seq data were downloaded from the ENCODE project (62).

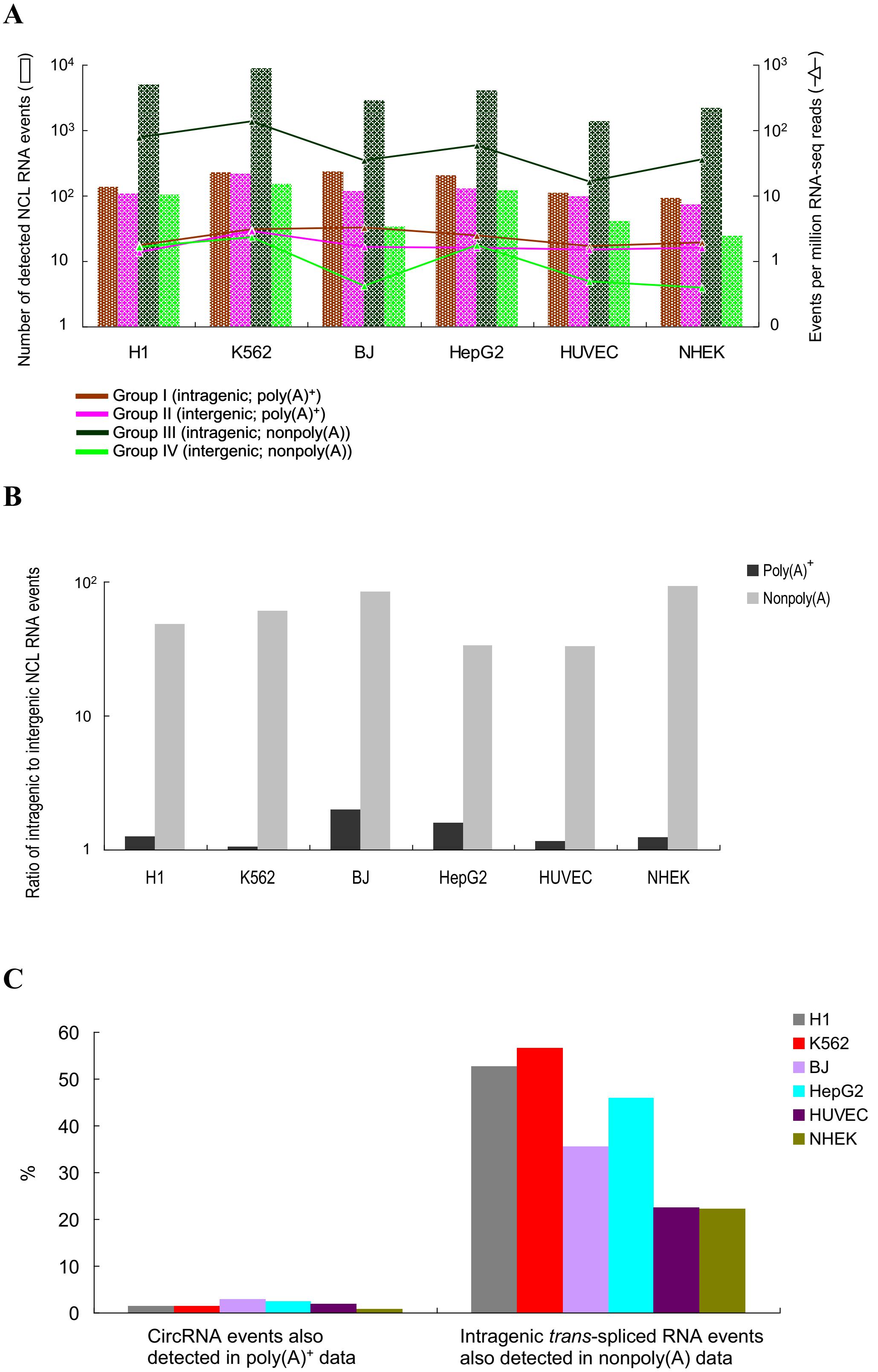

Figure 5. Distinctions between *trans*-splicing events, circular RNAs and fusion transcripts based on poly(A)- and nonpoly(A)-selected RNA-seq data. (**A**) Numbers of the identified Groups I to IV NCL events (see the text) in the six cultured cell lines (Table 2). (**B**) Quantity ratio of intragenic to intergenic NCL RNA events. (**C**) Percentages of circular RNA candidates (Group III) observed in poly(A)-selected samples and those of intragenic *trans*-spliced transcript candidates (Group I) observed in nonpoly(A)-selected samples.

Figure 5B). In addition, we found that only a small fraction (0.9–2.9%) of events in Group III were detected in poly(A)-selected sample (Figure 5C), in agreement with a previous report (28). These results suggested that most events in Group III should belong to circRNAs.

Intriguingly, in contrast to the low percentage of events in Group III detected in poly(A)-selected sample, a considerable fraction (22.2–56.6%) of events in Group I were also detected in nonpoly(A)-selected sample (Figure 5C). Since *trans*-spliced transcripts and circRNAs have been shown to be able to share the same junctions (32), the events detected in both Groups I and III may be attributed to both *trans*-splicing and circRNA products. Our results thus suggest that a considerable proportion of intragenic *trans*-spliced RNAs (i.e., Group I) may share the same NCL junctions with circRNAs (i.e., Group III).

Moreover, genomic structural variations (e.g., insertions, duplications and translocations) can also create scrambled exons (54,83,84), which may contribute false positives in the identified circRNA candidates. To address this possibility, we detected potential structure variations (SVs) by analyzing DNA paired-end tags (DNA-PETs) from K562 cells (62) using SVDetect (67). We examined the circRNA candidates (i.e., the Group III candidates; Supplemental Table S4), which were detected from the same sample (K562 cells), and found that only 6.7% of them (594 out of 8915 events) might be formed by the detected SVs (Supplemental Table S2; 'Materials and Methods' section). This suggested that most identified intragenic events in nonpoly(A)-selected samples might not be subject to the consequences of SVs.

We next examined the cell-type specificities and expression levels of the four groups of NCL transcripts. Generally, the majority (>70%) of NCL transcripts were detected in only one cell type (Figure 6A), and a considerable percentage of them were supported by a small number of reads (Figure 6B and Supplemental Figure S4). Comparisons of NCL transcripts in the four groups revealed two observations: (i) intragenic NCL transcripts (Groups I and III) tended to present in more cell types and supported by more reads than intergenic ones (Groups II and IV); and (ii) for intragenic NCL events, circRNAs (Group III) tended to exhibit less cell type-specificity (Figure 6A) and higher expression level (Figure 6B and Supplemental Figure S4) than intragenic *trans*-spliced transcripts (Group I). Of particular importance, more than a quarter of circRNAs (27.8%; 4468 out of 16 063 events) could be detected in multiple cell types (Figure 6A). Our results suggest that circRNAs exhibit a greater expression breadth and a higher expression level than other groups of NCL transcripts, also reflecting that the number of the detected circRNA events is much greater than those of the detected *trans*-splicing and fusion transcript events (Figure 5A). Moreover, 17.7% (2849 out of 16 063 events) of the circRNAs identified by NCLscan were also detected in different cell types by Guo *et al.* (28), and over half (51%, 2527 out of 4975 events) of the detected circRNAs in H1 human embryonic stem cells were also detected from H9 human embryonic stem cells by Zhang *et al.* (80) (Supplemental Table S6). These observations further suggest that certain circRNAs are broadly expressed in a variety of human cells, also reflecting previous reports that circRNAs are abundant in transcriptomes (25,26).

DISCUSSION

In this study, we developed a new pipeline, NCLscan, which is rather advantageous in the identification of both intragenic and intergenic NCL transcripts from paired-end RNA-seq data. We showed that the accuracy performance of NCLscan was superior over 18 currently available tools, including 11 fusion-detecting tools, 3 circRNA-detecting tools and 5 tools that are capable of detecting both intragenic and intergenic NCL transcripts, in terms of sensitivity and precision (Figures 3 and 4). Interestingly, by evaluating accuracy of these tools, we observed striking differences in accuracy between different versions or parameter settings of the same tool. For example, for Dataset A, TopHat-Fusion was advantageous in sensitivity ($S_n = 0.92$), but exhibited poor precision ($S_p = 0.2$) as trade-off. In contrast, TopHat-Fusion-post, an updated version of TopHat-Fusion for detecting intergenic NCL events, achieved excellent precision ($S_p = 0.97$), but it was worse in terms of sensitivity ($S_n = 0.68$) (Figure 3A). Another example is CIRCexplorer, which is a TopHat-Fusion-based tool for detecting intragenic NCL events (80). Our results revealed that CIRCexplorer was more precise but less sensitive than TopHat-Fusion (Figure 4). A similar scenario was observed for the updated version of segemehl (v. 0.2.0) and its previous release (v. 0.1.9) (Supplemental Table S9). In addition, many more false calls were reported in Dataset B by FusionAnalyzer without filters than by the same tool with filters (1869 versus 59); moreover, the use of its filters would even rule out all candidates in Dataset A (Supplemental Table S9). These results suggested that certain NCL RNA-identification strategies achieved better precision by sacrificing sensitivity, emphasizing the difficulty in reaching a balance between sensitivity and precision. It was noteworthy that although some tools, such as MapSplice2, SOAPfuse, defuse, FusionAnalyser, CRAC, ChimeraScan and BreakFusion, exhibited good precision ($S_p > 0.9$) on a small dataset (e.g., Dataset A; Figure 3A), they reported hundreds to thousands of false calls on a larger one (e.g., Dataset B; Figure 3B). This reveals that it is more challenging to remove false positives from a large dataset than from a smaller one. It seems that the larger the dataset is, the more ambiguous alignments there exist, resulting in a larger number of false calls. Here, we confirmed that NCLscan consistently achieved both the highest S_p and $F1$ values on a variety of simulated datasets, regardless of strategy choice for generating simulated dataset (i.e., Datasets A and B and the simulated datasets generated in this study), read depth, read length, or NCL transcript expression level (Figures 3A, B and 4). It is important to note that NCLscan achieved near 100% precision on all the simulated datasets examined (all $S_p > 0.98$). We also applied NCLscan to two real RNA-seq datasets (Supplemental Table S3), for which it demonstrated a high level of sensitivity (Figure 3C and Supplemental Table S4). These results indicated that NCLscan can effectively minimize false discovery rate, while maintain a good balance between sensitivity and precision.

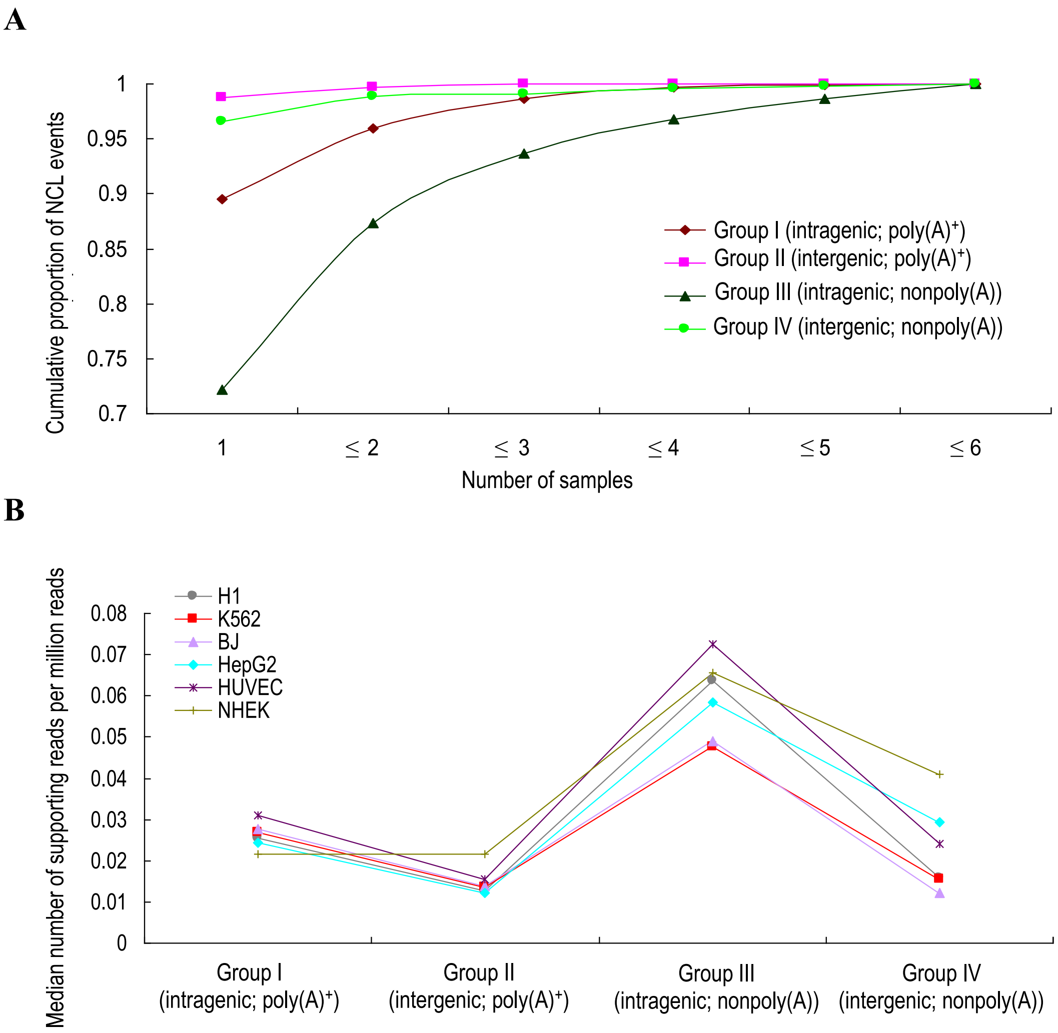

Figure 6. Comparison of (**A**) cell-type specificity and (**B**) expression level of the four groups of NCL transcripts.

Of note, since several studies have suggested that *in vitro* artifacts are the source of most detected sense-antisense fusion candidates (22,77,85), the current version of NCLscan does not consider this type of chimeric RNAs. However, a few sense-antisense fusions have been confirmed to be associated with the response to chemotherapy in cancer patients (86,87). Some sense-antisense fusion candidates were also provided in a prominent chimera database (i.e., ChiTARS (79)). It is worthwhile to add the capability of detecting sense-antisense chimeras to NCLscan in the future.

We further analyzed poly(A)- and nonpoly(A)-selected RNA-seq data from the ENCODE project to categorize the NCL transcripts into four groups: intragenic and intergenic events in poly(A)-selected samples, and intragenic and intergenic events in nonpoly(A)-selected samples; which enables us to distinguish between *trans*-splicing events, circRNAs and fusion transcripts. Our results revealed that the number of circRNAs was >10× larger than the numbers of other groups of NCL transcripts (Figure 5A). This trend was generally observed in diverse cell types, suggesting that circRNAs predominated in NCL RNAs. Recent studies have demonstrated that exon circularization is closely associated with *Alu*-based complementary sequences in flanking introns (26,80). Thus, widespread *Alu* elements in human introns may make a considerable contribution to the biogenesis of circRNAs (80), accounting for the wide expression of such RNAs in transcriptomes. In addition, we found that a considerable proportion (22.2–55.8%) of intragenic *trans*-splicing events were also observed in nonpoly(A)-selected samples (Figure 5C). By performing RNase R treatment and RT-PCR/qRT-PCR experiments in human embryonic stem cells (ESCs, line H1), we showed that the selected intragenic *trans*-splicing events were also resistant to RNase R degradation (Supplemental Note). These results suggest that certain intragenic *trans*-spliced RNAs may share the same NCL junction with circRNAs, although we cannot completely eliminate the possibility that poly(A)-selected RNA-seq data may still remain a few circRNA products (25). The above result is consistent with our previous study that an observed PtNCL splicing junction may result from *trans*-splicing RNA, circRNA or both (32).

Moreover, comparison of these four groups of NCL transcripts further revealed two interesting findings. First, intragenic events (Groups I and III) appeared to be less cell type-specific and expressed at a higher level than intergenic ones (Groups II and IV) (Figure 6A and B). Intragenic events may be more abundant than intergenic ones because intragenic splicing within the same gene tends to cause a higher local concentration of transcripts than intergenic splicing between different genes. On the basis of comparison of different reverse transcription products (22,32,77,78) in human H1 ESCs, our experimental validations also exhibited that the tested Groups I and III events had a higher proportion of authentic NCL transcripts than the tested Groups II and IV ones (Supplemental Note), in accord with the previous observations that intergenic events were less common than intragenic ones (22,32,45). Second, circRNAs (Group III) tended to be expressed more ubiquitously, less cell type-specifically and more abundantly than other types of NCL transcript (Figures 5A, 6A and B). This implies that circRNAs may play more diverse cellular roles than expected. For example, some circRNAs have been demonstrated to act as microRNA sponges or play a role in other non-catalytic cellular functions (29,33,88), and as such, these circRNAs are expressed at substantial levels within cells.

Interestingly, in addition to *BCR-ABL1*, a classic example of gene fusion, we also detected a known fusion event *IMMP2L-DOCK4* in K562 cells (Supplemental Table S4). These two intergenic events were detected in both poly(A)- and nonpoly(A)-selected samples, supporting the hypothesis that fusion transcripts can simultaneously contribute to both Groups II and IV. The *IMMP2L-DOCK4* fusion event was previously shown to be associated with autism and dyslexia (89,90), and might play an important role in neurite differentiation (91). This is the first time to detect the *IMMP2L-DOCK4* fusion event in K562 cells, providing a hint of future studies to investigate the role of this fusion event in chronic myeloid leukemia.

In conclusion, we presented a high accurate method, NCLscan, for detecting intragenic and intergenic NCL transcripts. We showed that NCLscan achieved a better accuracy in terms of sensitivity and precision than 18 other tools, minimizing long and costly experimental validations. Applying NCLscan to different types of RNA-seq data, such as poly(A)- and nonpoly(A)-selected RNA-seq data, was found to be a feasible way of distinguishing between *trans*-splicing events, circRNAs and fusion transcripts. We suggest that circRNAs exhibit a higher level of prevalence, expression level and expression breadth than other types of NCL transcripts. With different biological and mechanical roles, NCL transcripts provide an alternative means of increasing transcriptome complexity. Our study thus provides an efficient methodology for the design of future experimental studies to functionally probe different types of NCL event in transcriptome. Accumulating evidence has indicated the biological significance of NCL events (1,22,92–99), and as such, these largely uncharted classes of transcripts should not be overlooked in biomedical studies, especially in those seeking to develop therapeutics.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Chuang-Chieh Lin for operating FusionFinder, MapSplice2 and segemehl, and Shuo-Huang Chen for operating STAR. The authors are especially grateful to Dr. Raffaele A. Calogero for providing Dataset B (negative dataset), Dr. Steve Hoffmann (segemehl group) and Dr. Nicolas Philippe (CRAC group) for helpful suggestions on operating software, Dr. Heng Li for helpful comments on generating simulated datasets and Prof. Hung-Chih Kuo and his group members for providing human H1 ESC samples and experimental facilities.

FUNDING

Genomics Research Center, Academia Sinica, Taiwan (internal funding); Ministry of Science and Technology (MOST), Taiwan [MOST 103-2628-B-001-001-MY4, MOST 103-2911-I-001-507]. Funding for open access charge: Genomics Research Center, Academia Sinica, Taiwan; Ministry of Science and Technology (MOST), Taiwan [MOST 103-2628-B-001-001-MY4, MOST 103-2911-I-001-507].

Conflict of interest statement. None declared.

REFERENCES

1. Gingeras,T.R. (2009) Implications of chimaeric non-co-linear transcripts. *Nature*, **461**, 206–211.
2. Shtivelman,E., Lifshitz,B., Gale,R.P. and Canaani,E. (1985) Fused transcript of abl and bcr genes in chronic myelogenous leukaemia. *Nature*, **315**, 550–554.
3. Mitelman,F., Johansson,B. and Mertens,F. (2004) Fusion genes and rearranged genes as a linear function of chromosome aberrations in cancer. *Nat. Genet.*, **36**, 331–334.
4. Mitelman,F., Johansson,B. and Mertens,F. (2007) The impact of translocations and gene fusions on cancer causation. *Nat. Rev. Cancer*, **7**, 233–245.
5. Frohling,S. and Dohner,H. (2008) Chromosomal abnormalities in cancer. *N. Engl. J. Med.*, **359**, 722–734.
6. Westbrook,C.A., Hooberman,A.L., Spino,C., Dodge,R.K., Larson,R.A., Davey,F., Wurster-Hill,D.H., Sobol,R.E., Schiffer,C. and Bloomfield,C.D. (1992) Clinical significance of the BCR-ABL fusion gene in adult acute lymphoblastic leukemia: a Cancer and Leukemia Group B Study (8762). *Blood*, **80**, 2983–2990.
7. O'Brien,S.G., Guilhot,F., Larson,R.A., Gathmann,I., Baccarani,M., Cervantes,F., Cornelissen,J.J., Fischer,T., Hochhaus,A., Hughes,T. *et al.* (2003) Imatinib compared with interferon and low-dose cytarabine for newly diagnosed chronic-phase chronic myeloid leukemia. *N. Engl. J. Med.*, **348**, 994–1004.
8. Druker,B.J., Guilhot,F., O'Brien,S.G., Gathmann,I., Kantarjian,H., Gattermann,N., Deininger,M.W., Silver,R.T., Goldman,J.M., Stone,R.M. *et al.* (2006) Five-year follow-up of patients receiving imatinib for chronic myeloid leukemia. *N. Engl. J. Med.*, **355**, 2408–2417.
9. Tkachuk,D.C., Westbrook,C.A., Andreeff,M., Donlon,T.A., Cleary,M.L., Suryanarayan,K., Homge,M., Redner,A., Gray,J. and Pinkel,D. (1990) Detection of bcr-abl fusion in chronic myelogeneous leukemia by in situ hybridization. *Science*, **250**, 559–562.
10. Tognon,C., Knezevich,S.R., Huntsman,D., Roskelley,C.D., Melnyk,N., Mathers,J.A., Becker,L., Carneiro,F., MacPherson,N., Horsman,D. *et al.* (2002) Expression of the ETV6-NTRK3 gene fusion as a primary event in human secretory breast carcinoma. *Cancer Cell*, **2**, 367–376.

11. Barlund,M., Monni,O., Weaver,J.D., Kauraniemi,P., Sauter,G., Heiskanen,M., Kallioniemi,O.P. and Kallioniemi,A. (2002) Cloning of BCAS3 (17q23) and BCAS4 (20q13) genes that undergo amplification, overexpression, and fusion in breast cancer. *Genes Chromosomes Cancer*, **35**, 311–317.
12. Tomlins,S.A., Rhodes,D.R., Perner,S., Dhanasekaran,S.M., Mehra,R., Sun,X.W., Varambally,S., Cao,X., Tchinda,J., Kuefer,R. *et al.* (2005) Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science*, **310**, 644–648.
13. Kumar-Sinha,C., Tomlins,S.A. and Chinnaiyan,A.M. (2008) Recurrent gene fusions in prostate cancer. *Nat. Rev. Cancer*, **8**, 497–511.
14. Soda,M., Choi,Y.L., Enomoto,M., Takada,S., Yamashita,Y., Ishikawa,S., Fujiwara,S., Watanabe,H., Kurashina,K., Hatanaka,H. *et al.* (2007) Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. *Nature*, **448**, 561–566.
15. Persson,M., Andren,Y., Mark,J., Horlings,H.M., Persson,F. and Stenman,G. (2009) Recurrent fusion of MYB and NFIB transcription factor genes in carcinomas of the breast and head and neck. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 18740–18744.
16. Bass,A.J., Lawrence,M.S., Brace,L.E., Ramos,A.H., Drier,Y., Cibulskis,K., Sougnez,C., Voet,D., Saksena,G., Sivachenko,A. *et al.* (2011) Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 fusion. *Nat. Genet.*, **43**, 964–968.
17. Konarska,M.M., Padgett,R.A. and Sharp,P.A. (1985) Trans splicing of mRNA precursors in vitro. *Cell*, **42**, 165–171.
18. Solnick,D. (1985) Trans splicing of mRNA precursors. *Cell*, **42**, 157–164.
19. Li,H., Wang,J., Mor,G. and Sklar,J. (2008) A neoplastic gene fusion mimics trans-splicing of RNAs in normal human cells. *Science*, **321**, 1357–1361.
20. Schoenfelder,S., Clay,I. and Fraser,P. (2010) The transcriptional interactome: gene expression in 3D. *Curr. Opin. Genet. Dev.*, **20**, 127–133.
21. Rickman,D.S., Pflueger,D., Moss,B., VanDoren,V.E., Chen,C.X., de la Taille,A., Kuefer,R., Tewari,A.K., Setlur,S.R., Demichelis,F. *et al.* (2009) SLC45A3-ELK4 is a novel and frequent erythroblast transformation-specific fusion transcript in prostate cancer. *Cancer Res.*, **69**, 2734–2738.
22. Wu,C.S., Yu,C.Y., Chuang,C.Y., Hsiao,M., Kao,C.F., Kuo,H.C. and Chuang,T.J. (2014) Integrative transcriptome sequencing identifies trans-splicing events with important roles in human embryonic stem cell pluripotency. *Genome Res.*, **24**, 25–36.
23. Hsu,M.T. and Coca-Prados,M. (1979) Electron microscopic evidence for the circular form of RNA in the cytoplasm of eukaryotic cells. *Nature*, **280**, 339–340.
24. Nigro,J.M., Cho,K.R., Fearon,E.R., Kern,S.E., Ruppert,J.M., Oliner,J.D., Kinzler,K.W. and Vogelstein,B. (1991) Scrambled exons. *Cell*, **64**, 607–613.
25. Salzman,J., Gawad,C., Wang,P.L., Lacayo,N. and Brown,P.O. (2012) Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types. *PLoS One*, **7**, e30733.
26. Jeck,W.R., Sorrentino,J.A., Wang,K., Slevin,M.K., Burd,C.E., Liu,J., Marzluff,W.F. and Sharpless,N.E. (2013) Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA*, **19**, 141–157.
27. Wang,P.L., Bao,Y., Yee,M.C., Barrett,S.P., Hogan,G.J., Olsen,M.N., Dinneny,J.R., Brown,P.O. and Salzman,J. (2014) Circular RNA is expressed across the eukaryotic tree of life. *PLoS One*, **9**, e90859.
28. Guo,J.U., Agarwal,V., Guo,H. and Bartel,D.P. (2014) Expanded identification and characterization of mammalian circular RNAs. *Genome Biol.*, **15**, 409.
29. Memczak,S., Jens,M., Elefsinioti,A., Torti,F., Krueger,J., Rybak,A., Maier,L., Mackowiak,S.D., Gregersen,L.H., Munschauer,M. *et al.* (2013) Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature*, **495**, 333–338.
30. Ivanov,A., Memczak,S., Wyler,E., Torti,F., Porath,H.T., Orejuela,M.R., Piechotta,M., Levanon,E.Y., Landthaler,M., Dieterich,C. *et al.* (2015) Analysis of intron sequences reveals hallmarks of circular RNA biogenesis in animals. *Cell Rep.*, **10**, 170–177.
31. Westholm,J.O., Miura,P., Olson,S., Shenker,S., Joseph,B., Sanfilippo,P., Celniker,S.E., Graveley,B.R. and Lai,E.C. (2014) Genome-wide analysis of drosophila circular RNAs reveals their structural and sequence properties and age-dependent neural accumulation. *Cell Rep.*, **9**, 1966–1980.
32. Yu,C.Y., Liu,H.J., Hung,L.Y., Kuo,H.C. and Chuang,T.J. (2014) Is an observed non-co-linear RNA product spliced in trans, in cis or just in vitro? *Nucleic Acids Res.*, **42**, 9410–9423.
33. Hansen,T.B., Jensen,T.I., Clausen,B.H., Bramsen,J.B., Finsen,B., Damgaard,C.K. and Kjems,J. (2013) Natural RNA circles function as efficient microRNA sponges. *Nature*, **495**, 384–388.
34. Bachmayr-Heyda,A., Reiner,A.T., Auer,K., Sukhbaatar,N., Aust,S., Bachleitner-Hofmann,T., Mesteri,I., Grunt,T.W., Zeillinger,R. and Pils,D. (2015) Correlation of circular RNA abundance with proliferation–exemplified with colorectal and ovarian cancer, idiopathic lung fibrosis, and normal human tissues. *Sci. Rep.*, **5**, 8057.
35. Hentze,M.W. and Preiss,T. (2013) Circular RNAs: splicing's enigma variations. *EMBO J.*, **32**, 923–925.
36. Kim,D. and Salzberg,S.L. (2011) TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol.*, **12**, R72.
37. Francis,R.W., Thompson-Wicking,K., Carter,K.W., Anderson,D., Kees,U.R. and Beesley,A.H. (2012) FusionFinder: a software tool to identify expressed gene fusion candidates from RNA-Seq data. *PLoS One*, **7**, e39987.
38. McPherson,A., Hormozdiari,F., Zayed,A., Giuliany,R., Ha,G., Sun,M.G., Griffith,M., Heravi Moussavi,A., Senz,J., Melnyk,N. *et al.* (2011) deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. *PLoS Comput. Biol.*, **7**, e1001138.
39. Sakarya,O., Breu,H., Radovich,M., Chen,Y., Wang,Y.N., Barbacioru,C., Utiramerur,S., Whitley,P.P., Brockman,J.P., Vatta,P. *et al.* (2012) RNA-Seq mapping and detection of gene fusions with a suffix array algorithm. *PLoS Comput. Biol.*, **8**, e1002464.
40. Abate,F., Acquaviva,A., Paciello,G., Foti,C., Ficarra,E., Ferrarini,A., Delledonne,M., Iacobucci,I., Soverini,S., Martinelli,G. *et al.* (2012) Bellerophontes: an RNA-Seq data analysis framework for chimeric transcripts discovery based on accurate fusion model. *Bioinformatics*, **28**, 2114–2121.
41. Piazza,R., Pirola,A., Spinelli,R., Valletta,S., Redaelli,S., Magistroni,V. and Gambacorti-Passerini,C. (2012) FusionAnalyser: a new graphical, event-driven tool for fusion rearrangements discovery. *Nucleic Acids Res.*, **40**, e123.
42. Hoffmann,S., Otto,C., Doose,G., Tanzer,A., Langenberger,D., Christ,S., Kunz,M., Holdt,L., Teupser,D., Hackermueller,J. *et al.* (2014) A multi-split mapping algorithm for circular RNA, splicing, trans-splicing, and fusion detection. *Genome Biol.*, **15**, R34.
43. Kim,P., Yoon,S., Kim,N., Lee,S., Ko,M., Lee,H., Kang,H. and Kim,J. (2010) ChimerDB 2.0–a knowledgebase for fusion genes updated. *Nucleic Acids Res.*, **38**, D81–D85.
44. Ha,K.C., Lalonde,E., Li,L., Cavallone,L., Natrajan,R., Lambros,M.B., Mitsopoulos,C., Hakas,J., Kozarewa,I., Fenwick,K. *et al.* (2011) Identification of gene fusion transcripts by transcriptome sequencing in BRCA1-mutated breast cancers and cell lines. *BMC Med. Genomics*, **4**, 75.
45. McManus,C.J., Duff,M.O., Eipper-Mains,J. and Graveley,B.R. (2010) Global analysis of trans-splicing in Drosophila. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 12975–12979.
46. Zhang,G., Guo,G., Hu,X., Zhang,Y., Li,Q., Li,R., Zhuang,R., Lu,Z., He,Z., Fang,X. *et al.* (2010) Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. *Genome Res.*, **20**, 646–654.
47. Zhao,Q., Caballero,O.L., Levy,S., Stevenson,B.J., Iseli,C., de Souza,S.J., Galante,P.A., Busam,D., Leversha,M.A., Chadalavada,K. *et al.* (2009) Transcriptome-guided characterization of genomic rearrangements in a breast cancer cell line. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 1886–1891.
48. Maher,C.A., Kumar-Sinha,C., Cao,X., Kalyana-Sundaram,S., Han,B., Jing,X., Sam,L., Barrette,T., Palanisamy,N. and Chinnaiyan,A.M. (2009) Transcriptome sequencing to detect gene fusions in cancer. *Nature*, **458**, 97–101.
49. Maher,C.A., Palanisamy,N., Brenner,J.C., Cao,X., Kalyana-Sundaram,S., Luo,S., Khrebtukova,I., Barrette,T.R., Grasso,C., Yu,J. *et al.* (2009) Chimeric transcript discovery by paired-end transcriptome sequencing. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 12353–12358.
50. Ma,L., Yang,S., Zhao,W., Tang,Z., Zhang,T. and Li,K. (2012) Identification and analysis of pig chimeric mRNAs using RNA sequencing data. *BMC Genomics*, **13**, 429.

51. Inaki,K., Hillmer,A.M., Ukil,L., Yao,F., Woo,X.Y., Vardy,L.A., Zawack,K.F., Lee,C.W., Ariyaratne,P.N., Chan,Y.S. *et al.* (2011) Transcriptional consequences of genomic structural aberrations in breast cancer. *Genome Res.*, **21**, 676–687.
52. Al-Balool,H.H., Weber,D., Liu,Y., Wade,M., Guleria,K., Nam,P.L., Clayton,J., Rowe,W., Coxhead,J., Irving,J. *et al.* (2011) Post-transcriptional exon shuffling events in humans can be evolutionarily conserved and abundant. *Genome Res.*, **21**, 1788–1799.
53. Glazar,P., Papavasileiou,P. and Rajewsky,N. (2014) circBase: a database for circular RNAs. *RNA*, **20**, 1666–1670.
54. Shao,X., Shepelev,V. and Fedorov,A. (2006) Bioinformatic analysis of exon repetition, exon scrambling and trans-splicing in humans. *Bioinformatics*, **22**, 692–698.
55. Ozsolak,F. and Milos,P.M. (2011) RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet.*, **12**, 87–98.
56. Nacu,S., Yuan,W., Kan,Z., Bhatt,D., Rivers,C.S., Stinson,J., Peters,B.A., Modrusan,Z., Jung,K., Seshagiri,S. *et al.* (2011) Deep RNA sequencing analysis of readthrough gene fusions in human prostate adenocarcinoma and reference samples. *BMC Med. Genomics*, **4**, 11.
57. Carrara,M., Beccuti,M., Cavallo,F., Donatelli,S., Lazzarato,F., Cordero,F. and Calogero,R.A. (2013) State of art fusion-finder algorithms are suitable to detect transcription-induced chimeras in normal tissues? *BMC Bioinformatics*, **14**(Suppl. 7), S2.
58. Carrara,M., Beccuti,M., Lazzarato,F., Cavallo,F., Cordero,F., Donatelli,S. and Calogero,R.A. (2013) State-of-the-art fusion-finder algorithms sensitivity and specificity. *Biomed. Res. Int.*, 340620.
59. Jia,W., Qiu,K., He,M., Song,P., Zhou,Q., Zhou,F., Yu,Y., Zhu,D., Nickerson,M.L., Wan,S. *et al.* (2013) SOAPfuse: an algorithm for identifying fusion transcripts from paired-end RNA-Seq data. *Genome Biol.*, **14**, R12.
60. Ge,H., Liu,K., Juan,T., Fang,F., Newman,M. and Hoeck,W. (2011) FusionMap: detecting fusion genes from next-generation sequencing data at base-pair resolution. *Bioinformatics*, **27**, 1922–1928.
61. Edgren,H., Murumagi,A., Kangaspeska,S., Nicorici,D., Hongisto,V., Kleivi,K., Rye,I.H., Nyberg,S., Wolf,M., Borresen-Dale,A.L. *et al.* (2011) Identification of fusion genes in breast cancer by paired-end RNA-sequencing. *Genome Biol.*, **12**, R6.
62. Bernstein,B.E., Birney,E., Dunham,I., Green,E.D., Gunter,C. and Snyder,M. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
63. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
64. Kent,W.J. (2002) BLAT–the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
65. Holtgrewe,M., Emde,A.K., Weese,D. and Reinert,K. (2011) A novel and well-defined benchmarking method for second generation read mapping. *BMC Bioinformatics*, **12**, 210.
66. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
67. Zeitouni,B., Boeva,V., Janoueix-Lerosey,I., Loeillet,S., Legoix-ne,P., Nicolas,A., Delattre,O. and Barillot,E. (2010) SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. *Bioinformatics*, **26**, 1895–1896.
68. Grant,G.R., Farkas,M.H., Pizarro,A.D., Lahens,N.F., Schug,J., Brunk,B.P., Stoeckert,C.J., Hogenesch,J.B. and Pierce,E.A. (2011) Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). *Bioinformatics*, **27**, 2518–2528.
69. Chen,K., Wallis,J.W., Kandoth,C., Kalicki-Veizer,J.M., Mungall,K.L., Mungall,A.J., Jones,S.J., Marra,M.A., Ley,T.J., Mardis,E.R. *et al.* (2012) BreakFusion: targeted assembly-based identification of gene fusions in whole transcriptome paired-end sequencing data. *Bioinformatics*, **28**, 1923–1924.
70. Iyer,M.K., Chinnaiyan,A.M. and Maher,C.A. (2011) ChimeraScan: a tool for identifying chimeric transcription in sequencing data. *Bioinformatics*, **27**, 2903–2904.
71. Philippe,N., Salson,M., Commes,T. and Rivals,E. (2013) CRAC: an integrated approach to the analysis of RNA-seq reads. *Genome Biol.*, **14**, R30.
72. Li,Y., Chien,J., Smith,D.I. and Ma,J. (2011) FusionHunter: identifying fusion transcripts in cancer using paired-end RNA-seq. *Bioinformatics*, **27**, 1708–1710.
73. Liu,C., Ma,J., Chang,C.J. and Zhou,X. (2013) FusionQ: a novel approach for gene fusion detection and quantification from paired-end RNA-Seq. *BMC Bioinformatics*, **14**, 193.
74. Wang,K., Singh,D., Zeng,Z., Coleman,S.J., Huang,Y., Savich,G.L., He,X., Mieczkowski,P., Grimm,S.A., Perou,C.M. *et al.* (2010) MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.*, **38**, e178.
75. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
76. Hoffmann,S., Otto,C., Doose,G., Tanzer,A., Langenberger,D., Christ,S., Kunz,M., Holdt,L.M., Teupser,D., Hackermuller,J. *et al.* (2014) A multi-split mapping algorithm for circular RNA, splicing, trans-splicing and fusion detection. *Genome Biol.*, **15**, R34.
77. Houseley,J. and Tollervey,D. (2010) Apparent non-canonical trans-splicing is generated by reverse transcriptase in vitro. *PLoS One*, **5**, e12271.
78. Chen,I., Chen,C.Y. and Chuang,T.J. (2015) Biogenesis, identification, and function of exonic circular RNAs. *Wiley Interdiscip. Rev. RNA*, **6**, 563–579.
79. Frenkel-Morgenstern,M., Gorohovski,A., Vucenovic,D., Maestre,L. and Valencia,A. (2015) ChiTaRS 2.1–an improved database of the chimeric transcripts and RNA-seq data with novel sense-antisense chimeric RNA transcripts. *Nucleic Acids Res.*, **43**, D68–D75.
80. Zhang,X.O., Wang,H.B., Zhang,Y., Lu,X., Chen,L.L. and Yang,L. (2014) Complementary sequence-mediated exon circularization. *Cell*, **159**, 134–147.
81. Gao,Y., Wang,J. and Zhao,F. (2015) CIRI: an efficient and unbiased algorithm for de novo circular RNA identification. *Genome Biol.*, **16**, 4.
82. Djebali,S., Davis,C.A., Merkel,A., Dobin,A., Lassmann,T., Mortazavi,A., Tanzer,A., Lagarde,J., Lin,W., Schlesinger,F. *et al.* (2012) Landscape of transcription in human cells. *Nature*, **489**, 101–108.
83. Patthy,L. (1999) Genome evolution and the evolution of exon-shuffling–a review. *Gene*, **238**, 103–114.
84. Zhang,F., Gu,W., Hurles,M.E. and Lupski,J.R. (2009) Copy number variation in human health, disease, and evolution. *Annu. Rev. Genomics Hum. Genet.*, **10**, 451–481.
85. Kapranov,P., Drenkow,J., Cheng,J., Long,J., Helt,G., Dike,S. and Gingeras,T.R. (2005) Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. *Genome Res.*, **15**, 987–997.
86. Murga Penas,E.M., Cools,J., Algenstaedt,P., Hinz,K., Seeger,D., Schafhausen,P., Schilling,G., Marynen,P., Hossfeld,D.K. and Dierlamm,J. (2003) A novel cryptic translocation t(12;17)(p13;p12-p13) in a secondary acute myeloid leukemia results in a fusion of the ETV6 gene and the antisense strand of the PER1 gene. *Genes Chromosomes Cancer*, **37**, 79–83.
87. Tang,M., Foo,J., Gonen,M., Guilhot,J., Mahon,F.X. and Michor,F. (2012) Selection pressure exerted by imatinib therapy leads to disparate outcomes of imatinib discontinuation trials. *Haematologica*, **97**, 1553–1561.
88. Bahn,J.H., Zhang,Q., Li,F., Chan,T.M., Lin,X., Kim,Y., Wong,D.T. and Xiao,X. (2015) The Landscape of MicroRNA, Piwi-Interacting RNA, and Circular RNA in Human Saliva. *Clin. Chem.* **61**,221–230.
89. Pagnamenta,A.T., Bacchelli,E., de Jonge,M.V., Mirza,G., Scerri,T.S., Minopoli,F., Chiocchetti,A., Ludwig,K.U., Hoffmann,P., Paracchini,S. *et al.* (2010) Characterization of a family with rare deletions in CNTNAP5 and DOCK4 suggests novel risk loci for autism and dyslexia. *Biol. Psychiatry*, **68**, 320–328.
90. Maestrini,E., Pagnamenta,A.T., Lamb,J.A., Bacchelli,E., Sykes,N.H., Sousa,I., Toma,C., Barnby,G., Butler,H., Winchester,L. *et al.* (2010) High-density SNP association study and copy number variation analysis of the AUTS1 and AUTS5 loci implicate the IMMP2L-DOCK4 gene region in autism susceptibility. *Mol. Psychiatry*, **15**, 954–968.
91. Xiao,Y., Peng,Y., Wan,J., Tang,G., Chen,Y., Tang,J., Ye,W.C., Ip,N.Y. and Shi,L. (2013) The atypical guanine nucleotide exchange factor Dock4 regulates neurite differentiation through modulation of Rac1 GTPase and actin dynamics. *J. Biol. Chem.*, **288**, 20034–20045.
92. Parra,G., Reymond,A., Dabbouseh,N., Dermitzakis,E.T., Castelo,R., Thomson,T.M., Antonarakis,S.E. and Guigo,R. (2006) Tandem

chimerism as a means to increase protein complexity in the human genome. *Genome Res.*, **16**, 37–44.
93. Akiva,P., Toporik,A., Edelheit,S., Peretz,Y., Diber,A., Shemesh,R., Novik,A. and Sorek,R. (2006) Transcription-mediated gene fusion in the human genome. *Genome Res.*, **16**, 30–36.
94. Frenkel-Morgenstern,M. and Valencia,A. (2012) Novel domain combinations in proteins encoded by chimeric transcripts. *Bioinformatics*, **28**, i67–i74.
95. Frenkel-Morgenstern,M., Lacroix,V., Ezkurdia,I., Levin,Y., Gabashvili,A., Prilusky,J., Del Pozo,A., Tress,M., Johnson,R., Guigo,R. *et al.* (2012) Chimeras taking shape: potential functions of proteins encoded by chimeric RNA transcripts. *Genome Res.*, **22**, 1231–1242.
96. Zhou,J., Liao,J., Zheng,X. and Shen,H. (2012) Chimeric RNAs as potential biomarkers for tumor diagnosis. *BMB Rep.*, **45**, 133–140.
97. Conn,S.J., Pillman,K.A., Toubia,J., Conn,V.M., Salmanidis,M., Phillips,C.A., Roslan,S., Schreiber,A.W., Gregory,P.A. and Goodall,G.J. (2015) The RNA binding protein quaking regulates formation of circRNAs. *Cell*, **160**, 1125–1134.
98. Lasda,E. and Parker,R. (2014) Circular RNAs: diversity of form and function. *RNA*, **20**, 1829–1842.
99. Jeck,W.R. and Sharpless,N.E. (2014) Detecting and characterizing circular RNAs. *Nat. Biotechnol.*, **32**, 453–461.