BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 13 2013, pages 1600–1606
doi:10.1093/bioinformatics/btt185

Sequence analysis

Advance Access publication April 24, 2013

Janus—a comprehensive tool investigating the two faces of transcription

Matthias Barann[1,*], Daniela Esser[1], Ulrich C Klostermeier[1], Tuuli Lappalainen[2], Anne Luzius[1], Jan W. P. Kuiper[1], Ole Ammerpohl[3], Inga Vater[3], Reiner Siebert[3], Vyacheslav Amstislavskiy[4], Ralf Sudbrak[4], Hans Lehrach[4,5], Stefan Schreiber[1,6] and Philip Rosenstiel[1,*]

[1]Institute for Clinical Molecular Biology (ICMB), Christian-Albrechts-University, 24105 Kiel, Germany, [2]Department of Genetic Medicine and Development, University of Geneva Medical School, Switzerland, [3]Institute for Human Genetics, Christian-Albrechts-University, 24105 Kiel, Germany, [4]Max Planck Institute for Molecular Genetics and [5]Dahlem Centre for Genome Research and Medical Systems Biology, 14195 Berlin, Germany and [6]Department of General Internal Medicine, Christian-Albrechts-University, 24105 Kiel, Germany

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: Protocols to generate strand-specific transcriptomes with next-generation sequencing platforms have been used by the scientific community roughly since 2008. Strand-specific reads allow for detection of antisense events and a higher resolution of expression profiles enabling extension of current transcript annotations. However, applications making use of this strandedness information are still scarce.

Results: Here we present a tool (*Janus*), which focuses on the identification of transcriptional active regions in antisense orientation to known and novel transcribed elements of the genome. *Janus* can compare the antisense events of multiple samples and assigns scores to identify mutual expression of either transcript in a sense/antisense pair, which could hint to regulatory mechanisms. *Janus* is able to make use of single-nucleotide variant (SNV) and methylation data, if available, and reports the sense to antisense ratio of regions in the vicinity of the identified genetic and epigenetic variation. *Janus* interrogates positions of heterozygous SNVs to identify strand-specific allelic imbalance.

Availability: *Janus* is written in C/C++ and freely available at http://www.ikmb.uni-kiel.de/janus/janus.html under terms of GNU General Public License, for both, Linux and Windows 64×. Although the binaries will work without additional downloads, the software depends on bamtools (https://github.com/pezmaster31/bamtools) for compilation. A detailed tutorial section is included in the first section of the supplemental material and included as brief readme.txt in the tutorial archive.

Contact: m.barann@mucosa.de or p.rosenstiel@mucosa.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on July 16, 2012; revised on April 16, 2013; accepted on April 17, 2013

1 INTRODUCTION

The ancient Greek historian Plutarch writes about Janus: 'For this Janus, whether in remote antiquity he were a demigod or a king, was certainly a great lover of civil and social unity, and one who reclaimed men from brutal and savage living; for which reason they figure him with two faces, to represent the two states and conditions out of the one of which he brought mankind, to lead them into the other.' (Plutarchus, 75 A.C.E.). We decided to call our tool *Janus*, as it observes transcription of the double-stranded DNA, which often happens in opposite directions at a given locus. We term the transcription occurring on the strand opposite to the template strand 'antisense' transcription. Antisense transcripts may have a regulatory effect on the expression of the sense transcript, which has been reported for several sense/antisense (S/AS) pairs before, for example, in the X-chromosomal inactivation involving *XIST* and *TSIX* (Ng *et al*., 2007). Another example is given by Morris *et al*., who reported that expression of the p21 antisense transcript mediates methylation of the p21 sense promoter by recruitment of epigenetic regulatory complexes and showed that this effect was reversible by small interfering RNA (siRNA)-induced knockdown of the antisense transcript (Morris *et al*., 2008). Published data about antisense transcription are sometimes contradictory, underlining the high complexity of the matter. In 2008, He *et al*. published an article in which they investigated S/AS patterns in different human cell lines using the Illumina GA, detecting a high abundance of antisense tags within exons (He *et al*., 2008). In contrast, in an article published by Klevebring *et al*. in 2010, who used the SOLiD system, only few antisense tags were detected in the coding regions of genes (Klevebring *et al*., 2010). However, both articles agree on an abundance of antisense transcription occurring in promoter and terminator regions. To the best of our knowledge, no publicly available tool has been released to the scientific community to easily identify S/AS pairs using next-generation sequencing data. To address this, we developed *Janus*. It allows identification of S/AS pairs on the genome-wide level, using common input formats of sequencing data from strand-specific experiments, and detects differences between multiple samples. It is capable of including methylation data in the results and investigates S/AS ratios at single-nucleotide variant (SNV) positions. *Janus* was developed to work under Linux and Windows, but is restricted to 64-bit environments owing to the memory requirements.

*To whom correspondence should be addressed.

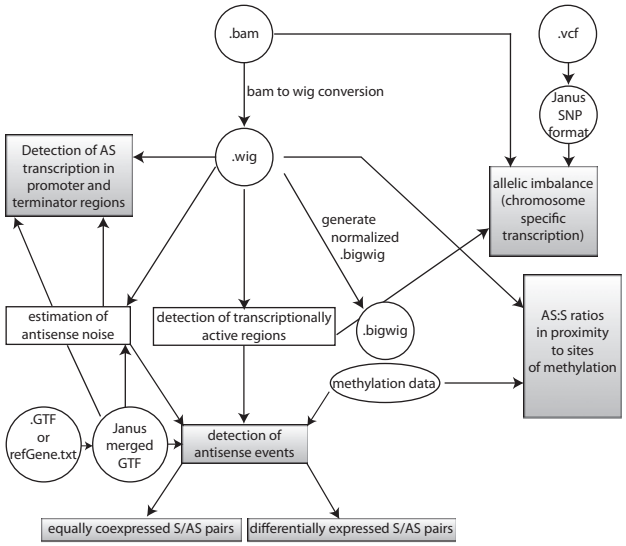

Fig. 1. *Janus* workflow. Input files are marked by circles and final analysis output in boxes shaded in gray. *Janus* generates wig files from a BAM file, which is the primary input. The wig files can be converted into bigwigs and are used to estimate the level of antisense noise and to detect transcriptionally active regions. To determine the location of an antisense event relative to other transcripts, *Janus* depends on a gene annotation file, which can be either a refGene.txt file or a GTF file. Finally, *Janus* detects antisense events above the estimated background noise and annotates them using the gene annotation file. Optionally, *Janus* can incorporate methylation data. Multiple lists of S/AS pairs can be compared with *Janus* to detect differentially expressed and equally co-expressed S/AS pairs. Additionally, Janus includes the functionality to determine AS:S ratios in proximity of methylation sites, quantify the amount of antisense transcription in promoter and terminator regions and the ability to detect strand-specific allelic imbalance

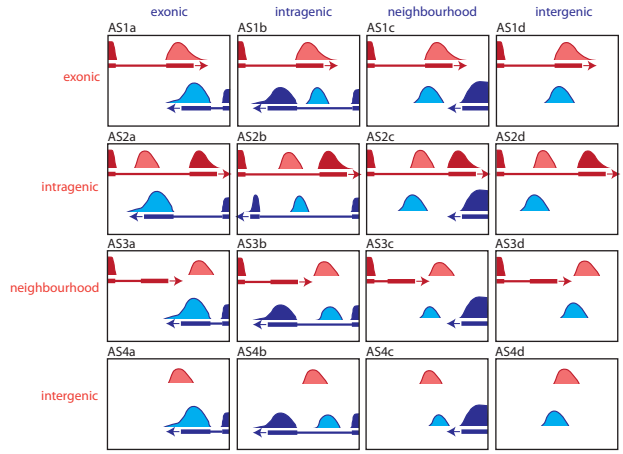

Fig. 2. Antisense classes. Coverage is shown for the first strand (red) and second strand (blue). Regions showing coverage on both strands (S/AS pairs) are shown in lighter colors

2 METHODS

The workflow of *Janus* is shown in Figure 1. *Janus* expects a BAM file as input, which is currently the most commonly used output format for mapped reads. Reads from the BAM file are used to generate one wig file per chromosome and strand. For further analysis, a transcript annotation file in gene transfer format (GTF) is required. *Janus* is designed to convert an existing GTF file or the refGene.txt, which can be obtained by the table browser of the UCSC homepage (http://genome.ucsc.edu/cgi-bin/hgTables), into a merged GTF file that contains one entry per gene symbol, including the merged meta-exons of all transcript isoforms of the same gene. Even if the annotation file is already in the GTF format, we recommend running the built-in converter of *Janus* to improve the results. Other GTF files, i.e. those that were not modified by *Janus*, may be used, but might generate multiple similar results for different isoforms of the same gene.

Using the reference annotation, the level of 'antisense-noise' is calculated (Supplementary Fig. S1). We define this noise as transcription on the opposite strand of known exons, which is a result of imperfect library generation and mismapping of reads. Often, transcription extends beyond the 3′-end of annotated transcripts. As this would bias the noise estimation in case of two adjacent transcripts in tail-to-tail orientation, we exclude exons within 1 kb of other exons on the opposite strand. The mean and standard deviation of the noise are calculated and reported in a detailed tab-delimited file for every investigated meta-exon. Mean and standard deviation are reported in a stats file that also contains the total coverage of the sample. The wig files are used to identify transcriptional active regions (TARs) (Supplementary Fig. S2). Three criteria must be met for a TAR to be regarded as an antisense event. First, the TAR must have a minimum length (default = 51 bp) and minimum average coverage per base (default = 5). The coverage criterion uses unnormalized values, i.e. the raw mapped coverage for the identification of TARs. Second, on the strand opposite to the identified TAR, transcription must occur within a given distance (default = 10 kb). Third, each TAR requires an antisense- to sense-transcription ratio above the estimated noise level. Events that do not differ by more than 1.94 standard deviations ($P \leq 0.05$) from the mean noise level are discarded. An antisense event is regarded as known if it is overlapping an exon on the same strand and novel otherwise. Each antisense event is classified depending on its location in relation to the sense event (Fig. 2).

Janus includes the functionality to compare antisense events of multiple samples to identify differences in gene expression that might affect only one strand or both strands equally. Supplementary Figure S3A–C illustrates possible types of AS events, which show equally strong expression as the sense transcript and Supplementary Figure S3D an S/AS pair with mutually exclusive expression of either the sense or antisense transcript. The workflow to identify potential self-regulatory S/AS pairs is shown in Figure 3. First, all overlapping antisense events are merged for all samples, and expression values are calculated for each strand and sample. The expression values are normalized by event length and total genomic coverage before scoring. Two quality scores are incorporated in the scoring algorithm: the mean value of covered length to full event length ($score_1$) and the number of samples that have a minimum coverage of 1 for the whole event ($score_4$). For identification of differential expressed S/AS pairs, two scores based on the AS:S ratio are used: the mean of the difference of the AS:S ratio from an equal distribution (1:1) and the distance between the minimum and maximum AS:S ratio. For the final score, the four sub-scores are weighted differently. The two scores representing the quality of the AS event ($score_1$ and $score_4$) are weighted double, while the two scores concerning the AS/S ratios ($score_2$ and $score_3$) are weighted triple, reflecting their higher importance for the detection of differential expression. Additionally, a flag is generated that marks the more extreme cases, in which one sample shows an AS:S ratio below 0.2 and another above 0.8 representing empirical borders determined from our training dataset.

The chosen scoring algorithm was not designed to preferentially identify S/AS events exhibiting the same expression level on both strands, as

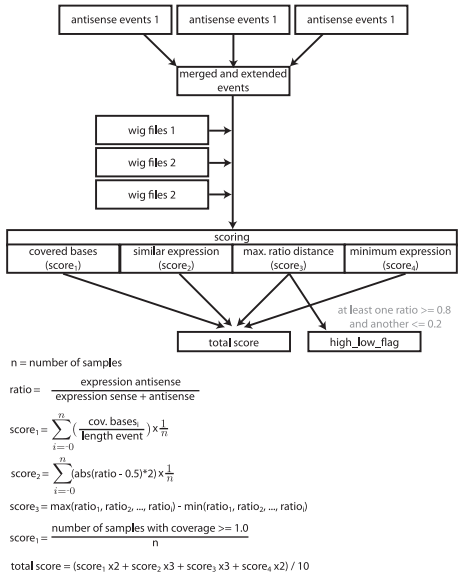

Fig. 3. Scoring algorithm for S/AS pairs. First the lists of antisense events detected in different samples are merged and overlapping events fused into single events. S/AS ratios are calculated and expression values are normalized for total genomic coverage and event length. Antisense events are given scores for mutually exclusive expression. All scores, including the total score, are between 0 and 1. $Score_1$ represents the fraction of the event that was covered. $Score_2$ (0–1) marks the distance of the event's S/AS ratio to an equal expression on both strands. $Score_3$ calculates the maximum S/AS-ratio-distance between the sample with the lowest S/AS ratio to the sample with the highest S/AS ratio. Additionally, a flag (either TRUE or FALSE) is generated, which tells the user if a sample with a very low and another with a very high S/AS ratio was detected. $Score_4$ represents the fraction of samples that achieved a normalized coverage of at least 1. The total score weights all four scores in slight favor of the expression ratios

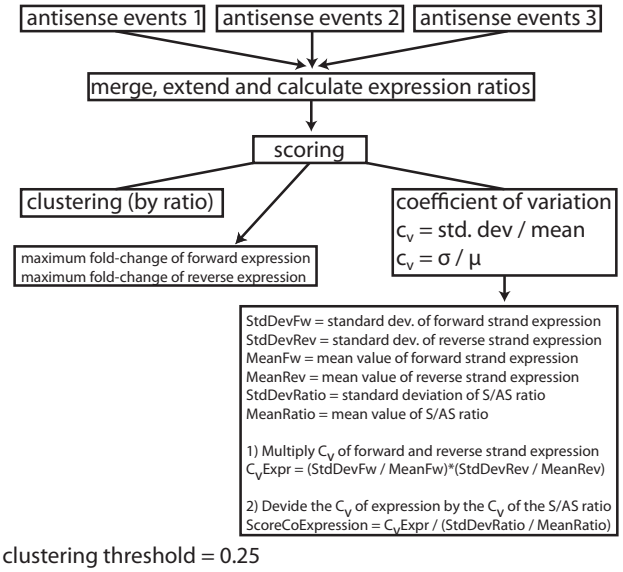

Fig. 4. Calculations helping to identify S/AS pairs that show an equal expression (co-expressed S/AS pairs). The co-expression score is a value for the dispersion of the coverage values in relation to the dispersion of the ratio. The value gets better/higher when the dispersion of coverage is high, while the dispersion of the ratio is low

this will not commonly be the case in regulatory events. However, it is possible to filter for them using additional parameters, which are also calculated during the scoring step for differential expressed S/AS pairs (Fig. 4). Samples are grouped by their AS:S ratio for a given S/AS pair. For an equally co-expressed S/AS pair, only one group should be present. Also, the similarity score, used in the calculation of differential S/AS pairs, is supposed to be low (0 indicates an AS:S expression ratio of 0.5) and the maximum difference of the highest and lowest AS:S ratio should be very low as well. For identification of equally co-expressed S/AS pairs, the fold change of overall expression of the S/AS transcripts between samples needs to be high. *Janus* generates a score for the co-expression by calculating the coefficients of variation for the forward and reverse strand expression and AS:S ratio. The score grows with higher expression dissimilarities between samples while maintaining a similar AS:S expression ratio.

Methylation data can be incorporated in the results (list of antisense events) by specifying a tab-delimited file with methylation data (see documentation for file formats). With these data, *Janus* can generate an additional file for the AS:S ratio in 100 bp, 500 bp, 2.5 kb and 10 kb range of the given site of methylation (Supplementary Fig. S4).

S/AS pairs that might derive exclusively from either homolog, i.e. one transcript is transcribed from the maternal chromosome while the antisense transcript is transcribed from the paternal chromosome, can be investigated based on a list of SNVs or without prior SNV information. Variant Call Format (VCF) files can be processed, which results in tab-delimited files that are then used by Janus (see documentation for file formats). For all SNV positions located within detected S/AS events, *Janus* counts the number of reads for each allele, which might be evenly distributed on both strands or prefer one strand for the sense and antisense transcript each (Supplementary Fig. S5), and calculates a *P*-value based on Fisher's exact test for a 2×2 contingency table (first, second strand and allele A, B). Only SNVs are considered for this analysis. Read counts for all SNV loci that show at least one allele, without requiring reads on both strands, will be reported and a warning will be displayed at the end of each entry if the 'inferred' (=dominant allele per strand) alleles differ from the alleles specified in the SNV file. This type of analysis can be used to identify genes where one haplotype is preferentially expressed, possibly in a strand-specific manner, which can be a sign of regulatory variation in *cis*. However, *Janus* was not specifically designed to detect monoallelic expression. Without a specified SNV list, *Janus* will investigate all positions covered by TARs and search for allele differences between the two strands. By this method, only loci that are covered by forward and reverse reads showing two distinct alleles will be reported and Fisher's exact test will be applied using the dominant allele from each strand.

3 RESULTS

Simulated data have been used to assess the performance of *Janus* (see Supplementary Material for generation of simulated data). Using a coverage threshold of 3 per base for the detection of antisense events, ~84.59% of simulated AS events were detected with exact start and end position (Supplementary Fig. S6). An additional 3.26% of simulated AS events were detected with exact start or end position and 12.15% did not match any of the simulated ends exactly. However, these 12.15% include AS events that are within simulated AS events but do not fit the simulated start and end exactly. Also, some AS events were generated with less than three coverage per base, and thus will not be

detected. Therefore, the 85% given above is a conservative estimation.

Additionally, we estimated the effect of different parameters on the detection of AS events by *Janus* (Supplementary Fig. S7). The most important filter in the simulated dataset is the coverage per base threshold. Using a higher coverage threshold effectively reduces the number of false-positive antisense events. However, it also removes lowly covered true-positive events. Using a length cutoff and using the noise filter had only limited consequences on the detection of simulated antisense events. This is partially due to the minimum length requirement during simulation that is at least two times the read length for sense transcripts and 300 bp for the antisense transcripts. Also, besides mismapped reads, no additional noise was added by the simulation. Therefore, the effect of these parameters will be bigger in real datasets.

Janus performance to identify S/AS transcript pairs with allelic imbalance was also assessed using simulated data (see Supplementary Material for data generation). Data were only simulated for chromosome 1, which resulted in 59 transcript pairs that shared overlapping regions (Supplementary Table S1). In these regions, SNVs were introduced with varying numbers of alternative allele reads, resulting in 11 transcript pairs that showed a different major allele ['valid' allelic imbalance events (AIE)]. *Janus* identified a total of 46 AIE including 8 'valid' simulated AIE. The remaining 3 valid AIE were generated with low numbers of alternative alleles (2, 1 and 1 for the AS transcript, respectively), which were not detected owing to mapping problems. Twenty-two of the AIE detected by *Janus* showed a *P*-value below 0.05, including six of the 'valid' AIE. Of the remaining 16 AIE, 15 can be traced back to mapping problems close to splice junctions (Supplementary Fig. S8), where reads containing only one to five bases of the neighboring exons are mapped partially into the intron. The remaining false-positive AIE with *P*-value below 0.05 is caused by other types of mismapping. Although *Janus* identifies most of the simulated AIE, it is important to consider the results with care, as mismapping (especially close to splice-junctions) can easily lead to a false-positive AIE detection.

SOLiD RNAseq data from real experiments (see next section for more details about this data) was mapped using different tools to assess the effect of different mappers on the analysis of *Janus*. The Bioscope software of Applied Biosystems performed best when comparing the number of mapped reads of all mappers used (Bioscope, BWA, Bowtie/TopHat) (Supplementary Fig. S9a). The number of identified antisense events is highly dependent on the number of mapped sequences, and thus the data mapped with Bioscope yields more antisense events (Supplementary Fig. S9b). The run time of *Janus* shows an almost linear increase with higher read numbers (Supplementary Fig. S9c), and the overlap of detected antisense events from the different mappings shows that there is almost complete overlap between different Bioscope mappings, but far less overlap with the other two mappers (Supplementary Fig. S9d). Therefore, the Bioscope software seems to work best for the investigated color-space libraries.

Janus has been applied to transcriptome sequence data generated from transformed lymphoblastoid cell lines (LCL) of five individuals included in the 1000 Genomes project (2010), which have been sequenced on the SOLiD system and on the Illumina

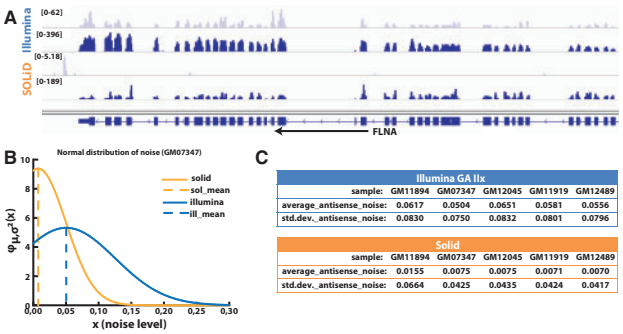

Fig. 5. (**A**) Expression shown for the same transcript with Illumina data (top) and solid data (bottom). First strand expression is shown in light blue and second strand expression in dark blue. The transcript (*FLNA*) is located on the second strand. Other than the Illumina data, SOLiD data show almost no expression on the first strand. The Illumina coverage on the first strand shows the same exon–intron pattern as the expression of the second (sense) strand. (**B**) Plotted mean and standard deviation for individual GM07347, the solid data show a smaller mean and standard deviation than the Illumina data. (**C**) Mean and standard deviation of noise level calculation for SOLiD and Illumina GA IIx data in numbers for all five LCL. In the SOLID data, ~0.7–1.6% of coverage in exonic elements considered for noise calculation is located on the antisense strand, whereas this contributes to 5.0–6.5% of coverage in the Illumina Data

GA IIx [EBI accession numbers: (SOLID) ERR012184, ERR012185, ERR012186, ERR012187, ERR012188; (Illumina) ERR011450, ERR011451, ERR011457, ERR011458, ERR011459]. The reads were mapped against the human hg19 reference, and the UCSC knownGenes annotation for hg19 was used. SNV calls from the 1000 Genomes project (pilot release from 23 November 2010) were used for the allele-specific expression analysis. *Janus* was also applied to SOLiD data from three classical Hodgkin Lymphoma (cHL) cell lines [L-1236, KM-H2 and U-HO1 (Mader *et al*., 2007)], for which also methylation data based on Infinium HumanMethylation27 BeadChips (Illumina, San Diego, USA) was available (Ammerpohl *et al*., 2012). These data were mapped against the human hg18 reference, and the corresponding gene annotation (knownGenes) from the UCSC was used (see Supplementary Material for mapping parameters). Also stranded data for two different mouse cell types (small intestine, colon) was investigated (GEO accession number GSE21746) that has been mapped against the mm9 reference. Duplicates in BAM files were marked using PICARD tools.

The two technologies [SOLiD WTAK (Klostermeier *et al*., 2011) and Illumina GA IIx (Parkhomchuk *et al*., 2009)] displayed different patterns. The Illumina data show a high abundance of antisense tags in exonic regions that follow the intron–exon structure of the sense transcript. We estimated a failure rate of ~5–6% for Illumina GA IIx strand-specific data and <1% failure rate for SOLiD WTAK ('Whole Transcriptome Analysis Kit') strand-specific data (Fig. 5).

To demonstrate the biological impact of Janus, we focused the following analysis on the SOLiD data, as it exhibits less background noise. For each sample, a large amount of antisense events was detected [~10 000 events for 30–40 million mapped

reads (~44–60 million total reads), Supplementary Table S2]. In a similar fashion to Figure 2, one example for each S/AS class is presented in Supplementary Figure S10. Many unannotated transcripts were detected by *Janus*, one such example is given in Supplementary Figure S11. Antisense transcription occurring in promoter and terminator regions, here defined as 1 kb before and after the transcript start and end site, was investigated. Of 32 110 (hg18, cHL data) and 66 065 (hg19, LCL data) annotated transcripts, an average of 760 promoter and 980 terminator regions (hg18) and 2100 promoter and 2400 terminator regions (hg19) showed antisense transcription with a minimum length of 150 bp and a mean coverage per base of at least five (Supplementary Table S3). A comparison of identified antisense events in promoter and terminator regions for different coverage limits is shown in Supplementary Figure S12. Considering all occurring transcription, terminator and promoter regions show a high abundance of sense tags. By filtering for promoter and terminator regions, which show an antisense event of a certain size (here 150 bp), the amount of sense tags is dramatically reduced and the amount of antisense tags increases, which is not surprising. However, by further filtering for minimum antisense coverage, the amount of antisense tags increases stronger in the terminator regions relative to the promoter region.

For both sets of human samples (1000 Genomes individuals, cHL cell lines), we were able to identify S/AS pairs that show mutually exclusive expression of either transcript. In the LCL, only one lucid pair (*SEPP1*/*CCDC152*, Supplementary Table S4 and Supplementary Fig. S13) could be identified. In contrast, despite being also derived from B-cells, the cHL cell lines showed multiple such events that could indicate that AS-mediated gene silencing plays a role in the pathogenesis of these tumors (Supplementary Table S5 and Supplementary Figs S14–S16). Supplementary Figure S17 shows tissue-specific examples of differentially expressed antisense transcripts in the investigated mouse libraries.

In the 1000 Genomes samples, several S/AS pairs that show a positively correlated expression pattern could be detected by the default parameters of the program (Supplementary Table S6 and Supplementary Figs S18–S24). Three examples for S/AS pairs that correlate in expression in the cHL dataset are shown in Supplementary Figures S25–S27. Supplementary Figure S28 illustrates an example of a tissue-specific co-expressed S/AS pair in the investigated mouse libraries. One probably false-positive example of a positive correlated S/AS pair is shown in Supplementary Figure S29, the pair is occurring in a repeat region and has roughly the same size and shape on both strands, and thus it is unlikely to be real expression on both strands. The repeat region shows strong expression in the L-1236 dataset but not in the other samples. Interestingly, it has been shown that derepression of a terminal repeat can lead to the activation of a proto-oncogene in human lymphoma (Lamprecht *et al*., 2010).

One case with a slight correlation of CpG methylation degree and differential expression of S/AS pairs was detected in the cHL data (Supplementary Fig. S16). The mean methylation beta values at two distinct locations near the ZSCAN16 promoter range from 0.02 to 0.15 for L-1236 and U-HO1 and 0.45 to 0.54 in KM-H2 (Supplementary Table S7). Expression of a short antisense transcript occurs only in KM-H2, while the sense transcript is only expressed in the other two cell lines.

Allele-specific expression data of known SNVs did not show any S/AS pairs with both transcripts expressed from different homologs exclusively. However, we identified some cases of allelic imbalance between the transcripts on the first and second strand (Supplementary Figs S30 and S31). The first example shows expression of *NCAPH2* and *SCO2*. While most reads corresponding to *NCAPH2* expression show the G allele and only few show the T allele, the opposite was found for reads corresponding to *SCO2* expression. The second example shows expression of *FAN1* and *MTMR10*. Reads corresponding to *FAN1* expression show both SNV alleles (T/C), while the reads corresponding to *MTMR10* expression show the C allele only. *Janus* also identified several SNVs (in the approach without a priori knowledge of SNV positions) where both alleles are exclusively expressed on either strand. Supplementary Figure S32 shows one such example. The region transcribed in antisense direction is overlapping a VEGA annotated pseudogene, overlaps partially a non-repeat masked duplication and the identified SNVs are not included in the well characterized SNV list of the 1000 Genomes project, and thus this example represents likely a false positive.

4 DISCUSSION

Evidence is growing that antisense transcription is not random and unspecific transcription of the genome, but an essential and complex component of the even more complex transcription machinery. Here we present a tool, *Janus*, to characterize strand-specific transcription from high-throughput sequencing data, and applied it to several real datasets. First, we used *Janus* to detect differences between the two investigated next-generation technologies. The data derived using the Illumina GA IIx shows a noisy background that is probably caused by imperfect library generation (i.e. not all reads preserved the correct strand). This gives a good explanation for the high antisense tag density in exonic regions observed by He *et al*., who used the Illumina GA, which could not be confirmed by Klevebring *et al*., who used the SOLiD system (He *et al*., 2008; Klevebring *et al*., 2010). It can thus be assumed that some of the differences between the two studies result from the obvious technical bias (background noise) of the two sequencing protocols. This underlines the importance to assess the 'noise' level, i.e. the error rate of the library preparation, before identifying antisense events. Knowing the higher level of background noise of the Illumina protocol, we used only the SOLiD data for further analysis. *Janus* performs well in detecting antisense events, and many of the identified events are still not annotated. However, it must be kept in mind that neighboring antisense events reported by *Janus* are often derived from the same transcript and do not represent independent results. Currently, *Janus* does not make use of pairing information or split reads, which could be used to link multiple AS events and improve the result list. Previous publications mention a high abundance of antisense tags both in the promoter and terminator regions of genes (He *et al*., 2008; Katayama *et al*., 2005; Layer and Weil, 2009). We used *Janus* to investigate the amount of sense and antisense transcription occurring in 1 kb up- and downstream annotated transcripts. Interestingly, when

considering antisense events with low coverage in the promoter and terminator regions, the amount of antisense tags in promoter regions is higher than in terminator regions, while this shifts toward more antisense tags in terminator regions when considering only higher covered antisense events. This could be caused by a higher mismapping in promoter regions, maybe due to conserved motifs of transcription factor binding sites and TATA-like sequences, which generates a higher noise level. Another explanation is that there is more low-level antisense transcription occurring in promoter regions and stronger antisense transcription in terminator regions. Also, it is possible that S/AS pairs in head-to-head conformation show lower expression than S/AS pairs in tail-to-tail conformation.

Among the huge amount of detected S/AS pairs, there are some pairs that show mutually exclusive expression of one strand in some samples, and expression of the other strand in others. The low number of these S/AS pairs suggests that they are rare events in the population and/or affect only a small proportion of transcripts. It remains unclear if there is an inhibitory effect of one transcript on its counterpart or if other mechanisms such as epigenetic modifications are involved, as has been shown for the highly imprinted *gnas* locus in mice (Holmes *et al*., 2003). We could identify several S/AS pairs that are co-expressed and show an equal expression rate of both transcripts, but differ in expression between samples. It could be argued this is caused by different library performance (i.e. this region was just unlucky in some samples) or a coverage bias might persist despite normalizing for the total coverage per sample. However, as shown in Supplementary Figure S18, *GLB1* shows a similar expression rate in all samples, while *TRIM71* and the unknown antisense event show a positively correlated expression, which differs between samples. It is unlikely that the observed co-expression is due to coverage effects. Katayama *et al*. (2005) suggested that long non-coding RNAs might be involved in S/AS pairs, reasoned by a higher number of S/AS pairs detected by random primed Cap-Analysis-Gene-Expression (CAGE) compared with oligo-dT primed CAGE. It could be speculated that the identified AS event upstream of *TRIM71* might present such a non-coding RNA, as no exon–intron structure is visible. Interestingly, in the same article, Katayama *et al*. reported a frequent concordant regulation of S/AS pairs. Another study by Watanabe *et al*. (2010) provides further evidence of a high occurrence of positively correlated S/AS pairs. The reason for this remains unclear, but it could be due to the fact that transcription of one strand makes the other strand more accessible as well. Linked to this, palindromic sequences in regulatory regions, such as promoter regions, might be used by both strands. In this case, the observed transcript could be nonsense transcription, which ends by the first possible encountered transcription end signal. This might explain why antisense transcripts in promoter regions are often described as short and long non-coding RNAs. However, it is not unlikely that these antisense transcripts have some function, as they do not occur in all promoter or terminator regions of every expressed gene. Multiple studies showed that non-coding RNAs might alter chromatin conformation or lead to a local change of methylation if they are located within promoter regions (Guttman and Rinn, 2012; Imamura *et al*., 2004; Murrell *et al*., 2004).

Using the data available to us, we could not link sense and antisense transcription to methylation of CpG islands, except weakly for one case, which might be coincidental. It is possible that the limited number of CpG islands interrogated by the chip (27 K) prevented identification of CpG methylation conditional antisense expression. A larger dataset or data based on histone modification might lead to new results, as a link to histone methylation has been shown before, i.e. in the case of the p21 antisense transcript. By investigating known SNV positions (generated by the 1000 Genomes Consortium), we were able to identify some SNVs showing signs of strand-specific allelic imbalance. *Janus* also identifies S/AS-specific allelic imbalance for positions that are not known to be SNV positions. Yet, most of these positions can be traced back to regions that are difficult to map to (i.e. human leukocyte antigen), repeat regions, pseudogenes and genes with multiple copies, and thus these results have to be considered with care.

While we have used independent, but highly similar, samples (lymphoma cell lines and lymphoblastoid cells and two different intestinal tissues) under steady-state conditions for development of the tool, it must be emphasized that detection of context-dependent regulation of S/AS transcription will require an appropriate number of biological replicates to filter away biological noise of a dynamic system.

With *Janus* we provide a useful tool for researchers to identify S/AS pairs within transcriptomes based on data of next-generation sequencing. While it remains difficult to identify differences between samples, *Janus* identifies S/AS pairs efficiently and may help to identify interesting S/AS pairs in disease-relevant data.

ACKNOWLEDGEMENTS

The authors gratefully appreciate the technical assistance of Lena Bossen, Melanie Friskovec, Dorina Oelsner, Kerstin Runde and Thomas Giger for processing the 1000 Genomes RNA samples and SOLiD sequencing data.

Funding: EU FP7 framework program gEUVADIS (Project no. 261123), the BMBF Network 'Systematic Genomics of chronic inflammation' GP9/GP10 and SP2-3 in the DEEP project, the DFG Clusters of excellence Inflammation at Interfaces and Future Ocean, an internal grant from the Medizinausschuss SH and the Max Planck Society; Hans-Dietrich Bruhn foundation (to M.B.); BMBF grant 01GS08201 (1000 Genomes Project) from the German program of medical genome research (NGFN plus) (to H.L.).

Conflict of Interest: none declared.

REFERENCES

Ammerpohl,O. *et al*. (2012) Array-based DNA methylation analysis in classical Hodgkin lymphoma reveals new insights into the mechanisms underlying silencing of B cell-specific genes. *Leukemia*, **26**, 185–188.

Guttman,M. and Rinn,J.L. (2012) Modular regulatory principles of large non-coding RNAs. *Nature*, **482**, 339–346.

He,Y. *et al*. (2008) The antisense transcriptomes of human cells. *Science*, **322**, 1855–1857.

Holmes,R. *et al*. (2003) A comprehensive transcript map of the mouse Gnas imprinted complex. *Genome Res*., **13**, 1410–1415.

Imamura,T. *et al.* (2004) Non-coding RNA directed DNA demethylation of Sphk1 CpG island. *Biochem. Biophys. Res. Commun.*, **322**, 593–600.

Katayama,S. *et al.* (2005) Antisense transcription in the mammalian transcriptome. *Science*, **309**, 1564–1566.

Klevebring,D. *et al.* (2010) In-depth transcriptome analysis reveals novel TARs and prevalent antisense transcription in human cell lines. *PLoS One*, **5**, e9762.

Klostermeier,U.C. *et al.* (2011) A tissue-specific landscape of sense/antisense transcription in the mouse intestine. *BMC Genomics*, **12**, 305.

Lamprecht,B. *et al.* (2010) Derepression of an endogenous long terminal repeat activates the CSF1R proto-oncogene in human lymphoma. *Nat. Med.*, **16**, 571–579.

Layer,J.H. and Weil,P.A. (2009) Ubiquitous antisense transcription in eukaryotes: novel regulatory mechanism or byproduct of opportunistic RNA polymerase? *F1000 Biol. Rep.*, **1**, 33.

Mader,A. *et al.* (2007) U-HO1, a new cell line derived from a primary refractory classical Hodgkin lymphoma. *Cytogenet. Genome Res.*, **119**, 204–210.

Morris,K.V. *et al.* (2008) Bidirectional transcription directs both transcriptional gene activation and suppression in human cells. *PLoS Genet.*, **4**, e1000258.

Murrell,A. *et al.* (2004) Interaction between differentially methylated regions partitions the imprinted genes Igf2 and H19 into parent-specific chromatin loops. *Nat. Genet.*, **36**, 889–893.

Ng,K. *et al.* (2007) Xist and the order of silencing. *EMBO Rep.*, **8**, 34–39.

Parkhomchuk,D. *et al.* (2009) Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res.*, **37**, e123.

Plutarchus,L.M. (1973) *Plutarchi vitae parallelae*. Vol.3 Fasc. 2, 2nd edn [Konrat,Z. and Hans,G. (eds)]. Teubner, Leipzig.

Watanabe,Y. *et al.* (2010) Genome-wide analysis of expression modes and DNA methylation status at sense-antisense transcript loci in mouse. *Genomics*, **96**, 333–341.

Genomes Project Consortium. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.