



## Data in Brief

## Analysis of paired end Pol II ChIP-seq and short capped RNA-seq in MCF-7 cells

Adam Scheidegger<sup>a</sup>, Adam Burkholder<sup>c</sup>, Ata Abbas<sup>a</sup>, Kris Zarns<sup>b</sup>, Ann Samarakkody<sup>a</sup>, Sergei Nechaev<sup>a,\*</sup><sup>a</sup> Department of Basic Sciences, University of North Dakota School of Medicine, Grand Forks, ND 58202, USA<sup>b</sup> Department of Computer Sciences, University of North Dakota, Grand Forks, ND 58202, USA<sup>c</sup> Center for Integrative Bioinformatics, National Institute of Environmental Health Sciences, Research Triangle Park, NC 27709, USA

## ARTICLE INFO

## Article history:

Received 20 May 2015

Received in revised form 3 June 2015

Accepted 7 June 2015

Available online 23 June 2015

## Keywords:

Pol II pausing

Paired end ChIP-sequencing

Short capped RNA

## ABSTRACT

While a role of promoter-proximal RNA Polymerase II (Pol II) pausing in regulation of eukaryotic gene expression is implied, the mechanisms and dynamics of this process are poorly understood. We performed genome-wide analysis of short capped RNAs (scrNAs) and Pol II chromatin immunoprecipitation sequencing (ChIP-seq) in human breast cancer MCF-7 cells to better understand Pol II pausing (Samarakkody, A., Abbas, A., Scheidegger, A., Warns, J., Nnoli, O., Jokinen, B., Zarns, K., Kubat, B., Dhasarathy, A. and Nechaev, S. (2015) RNA polymerase II pausing can be retained or acquired during activation of genes involved in the epithelial to mesenchymal transition. *Nucleic Acids Res* **43**, 3938–3949). The data are available at the NCBI Gene Expression Omnibus under accession number GSE67041. For both ChIP and scrNA samples, we used paired end sequencing on the Illumina MiSeq instrument. For ChIP-seq, the use of paired end sequencing allowed us to avoid ambiguities in center-read definition. For scrNA seq, this allowed us to identify both the 5'-end and the 3'-end in the same run that represent, respectively, the transcription start sites and the locations of Pol II pausing. The sharpening of Pol II ChIP-seq metagene profiles when aligned against 5'-ends of scrNAs indicates that these RNAs can be used to define the start sites for the majority of mRNA transcription events.

© 2015 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

**Specifications** [standardized info for the reader] Where applicable, please follow the Ontology for Biomedical Investigations: [http://obi-ontology.org/page/Main\\_Page](http://obi-ontology.org/page/Main_Page).

Organism/cell line/tissue	Human/MCF7 cells (ATCC)
Sex	N/A
Sequencer or array type	Illumina MiSeq
Data format	Raw fastq and processed bedgraph files
Experimental factors	N/A
Experimental features	ChIPseq and short-capped RNA analysis of Pol II pausing
Consent	N/A
Sample source location	N/A

## 1. Direct link to deposited data

<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE67041>.

## 2. Experimental design, materials and methods

## 2.1. Sample collection and wet preparation

## 2.1.1. Cell culture

Human breast cancer MCF-7 cells were obtained from the American Type Culture Collection and cultured in DMEM/F-12 medium (Life

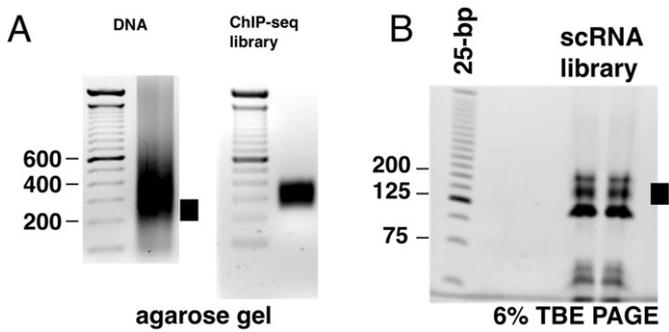
Technologies) supplemented with 10% fetal bovine serum (Gibco) at 37 °C in 5% CO<sub>2</sub> atmosphere.

## 2.1.2. Chromatin immunoprecipitation-sequencing library preparation

Chromatin Immunoprecipitation (ChIP) was performed as described [1], with additional details listed below. Approximately  $5 \times 10^7$  cells ( $3 \times 15$  cm dishes at ~80% confluency) were crosslinked in 1% formaldehyde in serum-free DMEM/F-12 media for 5 min at room temperature in dishes followed by quenching with glycine at 125 mM final concentration. Cells were washed with cold phosphate buffered saline and lysed on ice in 1 ml of 0.5% SDS-RIPA buffer. The lysate was sheared using the Covaris S220 sonicator for 6 min using the high cell program (output = 140, duty factor = 5). The timing of sonication was determined to generate DNA fragments below 500 nt in size as verified by agarose gel electrophoresis of DNA purified from chromatin (Fig. 1A). Chromatin was pre-cleared with 75  $\mu$ l of protein A + G magnetic beads (Life Technologies) for 2 h at 4 °C. The supernatant was supplemented with 5  $\mu$ l of the anti-Pol II antibody (N-20, Santa Cruz sc-899 X) and incubation continued overnight at 4 °C. Fresh protein A + G beads (50  $\mu$ l) were added for 2 h at 4 °C to isolate antibody-bound DNA, followed by washing twice using Low Salt Wash Buffer, High Salt Wash Buffer, LiCl Wash Buffer and TE Buffer. DNA was eluted from beads with two changes of Elution Buffer at 65 °C for 15 min each. Recovered DNA was purified by reverse crosslinking at 65 °C overnight

\* Corresponding author.

E-mail address: [sergei.nechaev@med.und.edu](mailto:sergei.nechaev@med.und.edu) (S. Nechaev).



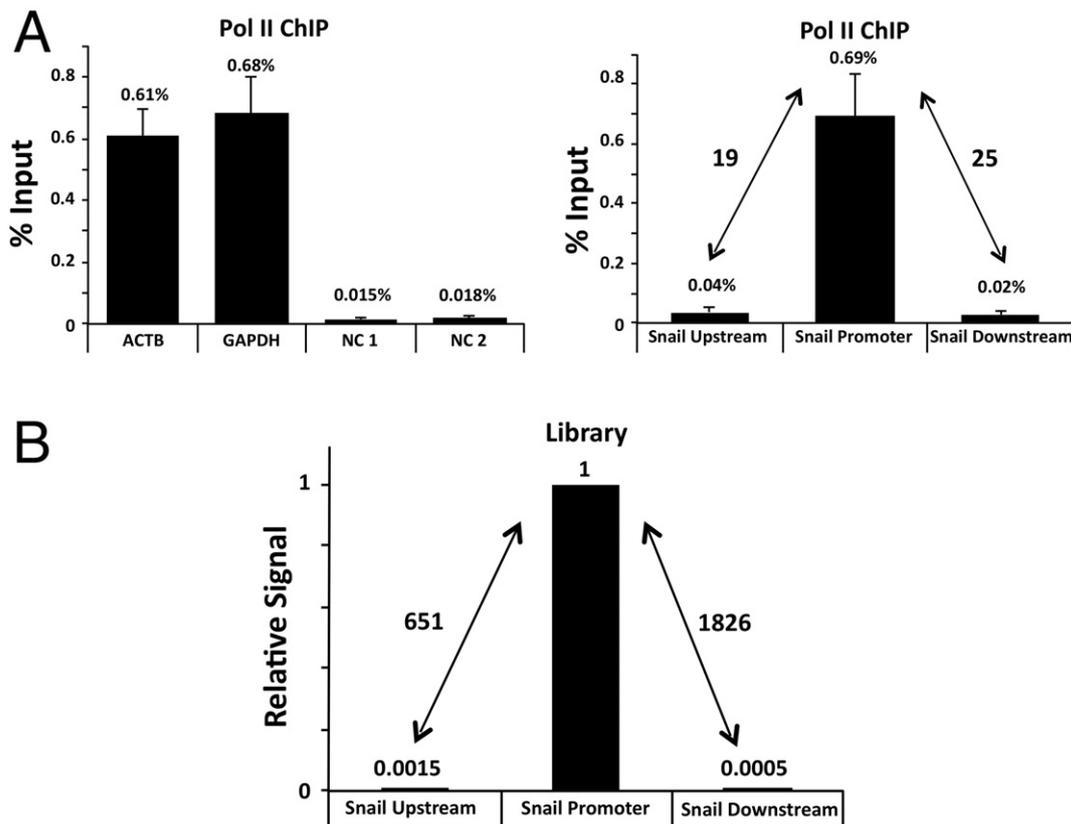
**Fig. 1.** Preparation of ChIP-seq and scRNA-seq libraries. A. Left. Sonicated DNA prepared from chromatin before ChIP. The black square shows the size range of DNA fragments that was extracted from the gel. Right. Example of a ChIP-sequencing library ready for cluster generation. Both images show ethidium bromide stained agarose-TAE gel. B. A scRNA library from MCF-7 cells resolved on a 6% TBE polyacrylamide gel. 25-bp ladder (Life Technologies) was used as a size marker. The band running at 125 bp size corresponds to the linker dimer with no insert and was avoided. The area marked by the black square was extracted from the gel for sequencing.

followed by treatments with RNase, proteinase K, and phenol/chloroform extraction. Purified DNA was concentrated by ethanol precipitation. All buffers are from [http://genome.ucsc.edu/ENCODE/protocols/general/UTA\\_ChIPseq\\_protocol.pdf](http://genome.ucsc.edu/ENCODE/protocols/general/UTA_ChIPseq_protocol.pdf). Approximately 10 ng of precipitated DNA (as determined by Qubit fluorometer, usually all of the available material from the above ChIP after qPCR control) was used for ChIP-sequencing library preparation. About 10% of the ChIP material was used for qPCR with gene specific primers to verify Pol II signal enrichment prior to library preparation (Fig. 2A). Library preparation was

done using the Illumina ChIP-sequencing kit as recommended by the manufacturer, using 14 cycles of PCR amplification. For size selection, a narrow, 275–325 bp band was excised from a 2% agarose gel when indicated per manufacturer's instructions prior to the final PCR step, which corresponds to the expected genomic insert size of approximately 150–200 bp. Quality of ChIP-seq libraries was verified for Pol II enrichment with gene-specific primers and positive and negative Pol II-specific control sets (Active Motif) as well as Snail gene specific primers designed for upstream, promoter, and downstream regions [1] (Fig. 2B), and quantified for cluster generation using the KAPA Library Quantification Kit. Libraries that did not show narrow band distribution or showed lower fold enrichment between positive and negative primer signal than in the original ChIP material (Fig. 2) were discarded. Sequencing was performed on the Illumina MiSeq instrument using V3 150 cycle kit in 75-bp paired end format combining two samples per run.

### 2.1.3. Short capped RNA library preparation

Short-capped RNA (scRNA) libraries were prepared as described [1] using gel-selection of short RNAs, enzyme-based selection of 5'-capped RNA followed by applying the Tru-Seq small RNA kit protocol to the resulting RNA pool. For these scRNA datasets, we did not further optimize size selection of scRNA libraries for human MCF-7 cells compared to the published *Drosophila* procedure [2]. After PCR-amplification for 18 cycles, each library was purified from a 6% nondenaturing TBE gel to remove adapter dimers (Fig. 1B). Before sequencing, each library was validated by ligating 50 ng (approximately 10%) of the gel-purified library into pBlueScript vector using blunt end cloning and blue-white selection, transforming into DH5-alpha *Escherichia coli* and sequencing plasmids from at least 10 individual clones using the T7 promoter primer (EtonBio). At least 2 out of 10 inserts should have



**Fig. 2.** Quality control of Pol II ChIP-seq by qPCR. A. Quantitative PCR on DNA extracted from precipitated chromatin before preparation of ChIP-sequencing libraries. Left panel shows positive (ACTB and GAPDH) and negative (NC1 and NC2) ChIP control primer pairs suitable for Pol II analysis (Active Motif). The right panel shows human Snail ChIP primers specific for the upstream, promoter, and downstream regions [1]. Y-axes show ChIP signal as percent of input DNA. B. Quantitative PCR of libraries prepared from ChIP material in A. Because calculating percent input in libraries is impossible, qPCR signal was normalized against Snail promoter primer pair, which was taken as 1. Numbers next to the arrows indicate fold-difference between indicated values (average of 2 independent biological replicates).

originated from known 5'-ends of genes, as verified by BLAT web interface [3], before a library was sequenced. Libraries that did not show promoter-derived inserts were discarded. Sequencing was performed on the Illumina MiSeq instrument using the RNA LT option in 50-bp paired end format, using V3 150-cycle MiSeq kit.

#### 2.1.4. Sequence alignment and bedgraph generation

ChIP and scRNA sequencing reads were aligned to *hg19* human reference genome using *bowtie* [4] version 1.1.1. Alignment of ChIP reads used a 75-nt seed length with 2 allowable mismatches, and reads with non-unique matches and read pairs mapped over 700-nt apart were discarded (*bowtie* -p16 -l75 -n2 -X700 -q -m1). scRNA reads were mapped separately using R1 and R2.fastq files corresponding, respectively, to 5'- and 3'-ends of scRNAs using a 24-nt seed length after 26-nt trimming from the 3' end of the read, with two allowable mismatches. Reads with non-unique matches were again thrown out (*bowtie* -p16 -q -m1 -l24 -n2 -326).

The center of each *bowtie*-mapped ChIP fragment needed to be translated into bedgraph format for heatmapping and visualization. The *extract\_fragments.pl* script takes a paired-end *bowtie* formatted alignment file as well as a list of chromosome lengths (this list can be automatically pulled from the UCSC genome browser from within the script) to generate a strand-merged bedgraph file of mapped hits with coordinates corresponding to the center of each fragment. For visualization of ChIP-seq tracks, 25nt binned bedgraph files were produced (*extract\_fragments.pl* -o g -min 75 -max 1000 -b25). Analysis of the ChIP-seq data other than visualization at UCSC browser ([https://genome.ucsc.edu/cgi-bin/hgTracks?hgS\\_doOtherUser=submit&hgS\\_otherUserName=ascheidegger&hgS\\_otherUserSessionName=scheidegger\\_gen\\_data](https://genome.ucsc.edu/cgi-bin/hgTracks?hgS_doOtherUser=submit&hgS_otherUserName=ascheidegger&hgS_otherUserSessionName=scheidegger_gen_data)) was done using unbinned bedgraph files (bin size = 1, *extract\_fragments.pl* -o g -min 75 -max 1000 -b1).

scRNA bedgraph files were made separately for each strand (forward and reverse) by returning the end positions of R1 and R2 reads separately, unbinned, generating four bedgraph files per scRNA sample. R1 read files reported 5'-ends of scRNAs and R2 read files 3' ends. To do this we used *bowtie2bedgraph.pl* (default options, -x for 3' ends). For each scRNA sample, four bedgraph files were generated (forward\_R1, reverse\_R1, forward\_R2, reverse\_R2). The resulting bedgraph files were normalized to the number of mapped reads in the alignment files and merged using the *bedgraph\_merge.R* script. The ChIP and scRNA files were normalized separately. Both sets of files were normalized down to the lowest number of reads to minimize the amount of noise.

#### 2.1.5. Reference table generation

The *make\_heatmap* program must be supplied with a reference table containing the genomic coordinates of each gene, the strandedness of that gene, and the coordinate of the feature of interest. For this experiment, the feature of interest is the transcription start site. The first step was the curation of the gene list. The initial list of 50,064 RefSeq genes was downloaded from the UCSC genome browser (GRCh37/hg19). All genes from chrM as well as a variety of non-coding RNAs (MIR, SNORD, SNORA, Y\_RNA, TRNA, rRNA, snoRNA and snRNA) were removed. Transcripts annotated as AF(<numeric>), AJ(<numeric>), AK(<numeric>), AL(<numeric>), AY(<numeric>), and BC(<numeric>), were also removed.

The genes were sorted by Pol II ChIP-seq signal within +/- 500 nucleotides from the annotated start site (based on *cppmatch* counts within the interval, see below), and duplicate gene isoforms with the same name but alternative start sites located farther than 500 nt apart were resolved by keeping the one demonstrating the highest Pol II enrichment. The resultant list of 24,441 genes was further pruned to contain the top 30%.

For the remaining genes, we used two definitions of transcription start sites, generating two tables. The first table was generated with the curated gene list using RefSeq annotated gene start sites. The

second table was generated with the same gene list using transcription start sites defined by the scRNA signal [1]. This required the *bedgraph2cppmatch.pl*, *cppmatch*, *deduplicate.pl*, and *tss\_reannotation.R*. *Bedgraph2cppmatch.pl* reorders and reformats the scRNA bedgraph file so that it can be read by *cppmatch*. *Cppmatch* requires the output of *bedgraph2cpp* and a table with gene and desired interval information (-100/+200 nt from the gene start). This table contains the gene identification, chromosome number, strand identity, and absolute coordinates of the start and end of the interval around the transcription start site for each gene. *Cppmatch* will return all hits in the bedgraph file falling within the interval for each gene in the above-defined gene list.

The transcription start site for each gene was reannotated based on the coordinate within the -100/+200 nt interval that contained the highest number of scRNA 5'-ends on the sense strand, using *cppmatch* and *deduplicate.pl* run on the combined scRNA bedgraph file. Only sense strand matches were counted (R1 scRNA forward strand hits were considered for plus strand genes and R1 scRNA reverse strand hits were considered for minus strand genes). The coordinate returned by *deduplicate.pl* for each gene was reported as the reannotated gene start site. Finally, the *tss\_reannotation.R* script uses this list to produce a table containing the updated transcription start sites formatted so that the downstream applications are able to properly center the reads for visualization.

For calculating promoter Pol II enrichment, *cppmatch* was run with the merged ChIP-seq Pol II bedgraph file and Pol II hits were summed within +/- 500nt interval of each gene.

#### 2.1.6. Heatmap generation

The primary functional program, *make\_heatmap*, converts a single nucleotide binned bedgraph file into a matrix. This matrix contains the number of reads, by gene, found within a specified interval around a genetic feature. This output can be further customized by editing the size of the interval, the position of the feature, as well as the number and size of bins to report within the interval.

The *make\_heatmap* code is 'strand aware' and will orient all the count information in the correct down/upstream direction regardless of the strand on which a gene is located. The columns in the matrix represent the count information collected by the program binned in user-defined sizes and encompassing an interval around the genomic feature also supplied by the user. For this experiment we generated matrices with both 10 nt and 1 nt bins within an interval of +/- 500 nt from the TSS. As the input file for ChIP-seq samples, we used the bedgraph file merged from two biological replicates and containing centerpoints of R1 and R2 strand matches, using the physical start option to define genes (-l p) and returning matches around both plus and minus strand genes (-s b) (*make\_heatmap* -b c -h g -l p -a s -v t -d g -s b <fileinfo> -500 10 100 (10 nt bin) or -500 1 1000 (1 nt bin)). This generated a single matrix file per ChIP-seq sample. For scRNAs, the forward and reverse bedgraph files were used separately to build matrices for genes annotated on plus and on minus strands in *hg19* (*make\_heatmap* -b c -h g -l s -a s -v t -d g -s s (sense) or -s o (antisense) <fileinfo> -500 10 100 (10 nt bin) or -500 1 1000). This generated four matrix files for each forward and reverse bedgraph input file pair (either R1 or R2). 10-bin tables were used for heat map plots and 1-bin tables were used for metagene plots. Hits from the forward read file mapping within the +/- 500 nt interval of plus strand genes and hits from the reverse read file matching to minus strand genes were classified as sense matches and the two matrix files were merged as the combined sense strand matrix file. Accordingly, reads from the forward read file around minus strand genes and reads from the reverse strand file around the plus strand genes were classified as antisense matches and the respective two matrix files were merged to make the combined antisense strand matrix file. The resultant two combined strand matrix files were used to make heat map images (10-nt bins) and metagene plots (1-nt bins). This analysis was done separately for R1 (5'-ends) and R2 (3'-ends) of scRNAs to generate four heatmaps

and four metagene plots for each sample (sense 5', antisense 5', sense 3' and antisense 3').

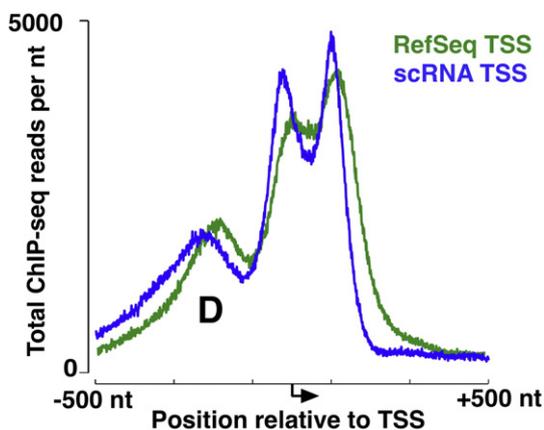
All visualization was done in R [5] using gplots [6], RcolorBrewer [7], and base functions. Reshape2 [8] and plyr [9] functions were also used. The *heatmap\_gen.R* script produces three output files. First, a csv of the data, sorted by the total signal across the entire interval is produced as a reference. The function *heatmap\_gen\_list* draws and saves heatmaps from each matrix given as input to the script [2]. Last, the *metagene\_gen\_list* function produces the metagene plot for each matrix (Fig. 3).

Versions of all software used in this analysis are available for download here: <https://github.com/kzarns/ChIP-Seq/tree/master/tools>.

### 3. Discussion

We describe the analysis of whole-genome paired end Pol II ChIP-seq and paired end short capped RNA (scrRNA) data in human breast cancer MCF-7 cells [1]. For ChIP-seq, we report paired end analysis of sequencing performed on a MiSeq instrument. With sufficient amount of starting cell material and optimized washing conditions to obtain sufficient signal-to-noise ratio as controlled by qPCR with positive and negative control primers prior to library preparation and in the final library, approximately 12 M reads per replicate are sufficient for ChIP-seq in the human genome with N-20 anti-Pol II antibody. As a result, two ChIP-seq samples can be combined in a single V3 MiSeq run. The use of paired end sequencing in conjunction with small (~300 nt total fragment, 180 nt mean insert size in the library) and uniform library sizes selected by agarose gel electrophoresis makes it possible to increase the resolution of ChIP-sequencing and directly visualize the double peaks of Pol II signal previously reported from re-analysis of single-end ChIP-seq data [10]. Because it is a MiSeq instrument, there is no additional cost associated with paired-end sequencing.

For scrRNAs, we report paired end analysis of short capped RNA sequencing. In the past, single-end sequencing of scrRNA using short reads was reported [2,11]. Paired end sequencing allows one to directly determine the 5'-ends and the 3'-ends of scrRNAs without relying on adapter trimming and to use short, 50-nt or fewer, sequencing length per side. The percentage of mappable reads for scrRNA datasets is lower than for ChIP-seq likely because of carryover of stable RNAs such as rRNA [12] (Table 1, Fig. 1). Because the 5'-cap selection procedure efficiently selects for transcription start sites of genes, a relatively low coverage might be sufficient, such that up to four samples can be combined in a single V3 MiSeq run even without additional changes



**Fig. 3.** Centering of Pol II ChIP-sequencing signal around transcriptional start sites. The plot shows metagene analysis as in [1], except that metagene plots for ChIP-seq signal are overlaid. Plots are centered around RefSeq annotated transcription start sites (TSSs) (Green) and scrRNA-defined TSSs (Blue). Defining TSSs based on scrRNAs sharpens the Pol II signal. We suggest that the previously observed double peaks of Pol II at the promoters [10] are reproducible, but likely arise from sonication-based ChIP-sequencing [1], and are distinct from the broad peak corresponding to divergent transcription (marked with "D").

**Table 1**

Alignment statistics for all samples included in the analysis.

Sample	Total reads	Aligned reads	Multi-mapped reads (suppressed)
ChIP 1	15,652,676	13,446,295 (85.90%)	540,939 (3.46%)
ChIP 2	16,947,644	14,635,302 (86.36%)	561,951 (3.32%)
scrRNA 1	4,815,685	1,837,112 (38.15%)	624,503 (12.97%)
scrRNA 2	10,169,519	4,980,050 (48.97%)	2,667,232 (26.23%)
scrRNA 3	3,186,513	1,610,543 (50.54%)	1,120,507 (35.16%)

in RNA size selection. Further optimization of scrRNA size selection towards recovering shorter RNA species, in 25–50 nt range for human cells, and use of pre-selection by removal of known stable RNA species [13], is expected to increase the percentage of uniquely mappable reads and further improve coverage with comparable depth of sequencing. We note that in contrast to the Tru-Seq small RNA kit, the libraries prepared with NEB Illumina-compatible small RNA kit are not compatible with paired-end sequencing because of a different sequence of the 3'-end adapter, and can only be used in a single-end sequencing format.

The use of scrRNAs allows one to re-annotate the positions of gene start sites [1,2,11,14]. While annotations of low-activity genes and less stable transcripts will require higher coverage of sequencing, using MCF-7 data, we clearly see the increase in sharpness of the ChIP-seq data peaks in metagene analyses when the start sites were reannotated using our existing short capped data [1] (Fig. 3). The use of scrRNAs for mapping promoter-proximal Pol II pausing and annotation of TSSs will continue to prove useful in future analyses of Pol II-related datasets.

### Acknowledgments

This work was supported by the UND School of Medicine pilot grant and by the National Institutes of Health [5P20GM104360 to N.S.].

### Appendix A. Supplementary data

Supplementary data to this article can be found online at <http://dx.doi.org/10.1016/j.gdata.2015.06.021>.

### References

- [1] A. Samarakody, A. Abbas, A. Scheidegger, J. Warns, O. Nnoli, B. Jokinen, K. Zarns, B. Kubat, A. Dhasarathy, S. Nechaev, RNA polymerase II pausing can be retained or acquired during activation of genes involved in the epithelial to mesenchymal transition. *Nucleic Acids Res.* 43 (2015) 3938–3949, <http://dx.doi.org/10.1093/nar/gkv263>.
- [2] S. Nechaev, D.C. Fargo, G. dos Santos, L. Liu, Y. Gao, K. Adelman, Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in *Drosophila*. *Science* 327 (2010) 335–338, <http://dx.doi.org/10.1126/science.1181421>.
- [3] W.J. Kent, BLAT—the BLAST-like alignment tool. *Genome Res.* 12 (2002) 656–664, <http://dx.doi.org/10.1101/gr.229202> Article published online before March 2002.
- [4] B. Langmead, C. Trapnell, M. Pop, S.L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10 (2009) R25, <http://dx.doi.org/10.1186/gb-2009-10-3-r25>.
- [5] Team, R.D.C. (2008) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org>
- [6] Gregory R. Warnes, B.B., Lodewijk Bonebakker, Robert Gentleman, Wolfgang Huber Andy Liaw, Thomas Lumley, Martin Maechler, Arni Magnusson, Steffen Moeller, Marc Schwartz and Bill Venables (2015) gplots: Various R Programming Tools for Plotting Data. R package version 2.16.0. <http://CRAN.R-project.org/package=gplots>
- [7] E. Neuwirth, RColorBrewer: ColorBrewer palettes. R package version 1.0–1.5, 2011 <http://CRAN.R-project.org/package=RColorBrewer>.
- [8] H. Wickham, reshape2: Flexibly reshape data: a reboot of the reshape package. R package version 1, 2012 <http://cran.r-project.org/web/packages/reshape2/index.html>.
- [9] H. Wickham, The split-apply-combine strategy for data analysis. *J. Stat. Softw.* 40 (2011) 1–29 <http://www.jstatsoft.org/v40/i01/>.
- [10] M. Quinodoz, C. Gobet, F. Naef, K.B. Gustafson, Characteristic bimodal profiles of RNA polymerase II at thousands of active mammalian promoters. *Genome Biol.* 15 (2014) R85, <http://dx.doi.org/10.1186/gb-2014-15-6-r85>.
- [11] T. Henriques, D.A. Gilchrist, S. Nechaev, M. Bern, G.W. Muse, A. Burkholder, D.C. Fargo, K. Adelman, Stable pausing by RNA polymerase II provides an opportunity to target and integrate regulatory signals. *Mol. Cell* 52 (2013) 517–528, <http://dx.doi.org/10.1016/j.molcel.2013.10.001>.

- [12] L.J. Core, A.L. Martins, C.G. Danko, C.T. Waters, A. Siepel, J.T. Lis, Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat. Genet.* 46 (2014) 1311–1320, <http://dx.doi.org/10.1038/ng.3142>.
- [13] A. Mayer, J. di Iulio, S. Maleri, U. Eser, J. Vierstra, A. Reynolds, R. Sandstrom, J.A. Stamatoyannopoulos, L.S. Churchman, Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution. *Cell* 161 (2015) 541–554, <http://dx.doi.org/10.1016/j.cell.2015.03.010>.
- [14] W. Gu, H.C. Lee, D. Chaves, E.M. Youngman, G.J. Pazour, D.J. Conte, C.C. Mello, CapSeq and CIP-TAP identify Pol II start sites and reveal capped small RNAs as *C. elegans* piRNA precursors. *Cell* 151 (2012) 1488–1500, <http://dx.doi.org/10.1016/j.cell.2012.11.023>.