Computational cleaning of noisy 5' end tag sequencing data sets from rare in vivo cells

Johannes Eichler Waage[1], Ilka Hoof[1], Jette Bornholdt[1], Esben Pedersen[2], Mette Jørgesen[1], Kim Theilgaard[3], Cord Brakebusch[1], Bo Porse[4], Albin Sandelin[1] ✉

[1]Bioinformatics Centre, University of Copenhagen, Denmark
[2]Biomedical Institute, BRIC, University of Copenhagen, Copenhagen, Denmark
[3]Biotech Research and Innovation Centre, University of Copenhagen, Denmark
[4]The Finsen Laboratory, Rigshospitalet, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark

Motivation and Objectives

CAGE-seq, cap analysis of gene expression followed by next-generation sequencing, allows for precise profiling of the promoterome (Plessy *et al.,* 2010) Here, we present a data filtration and processing pipeline for analysis of nanoCAGE-seq, a variant of the method allowing for very small amounts of input material (~50 ng per sample), and thus expanding the number of tissue- and cell types available for proteome profiling. We show, however, that low-intensity signal across exons, mRNA degradation and other method-specific noise is common to this technique, obfuscating true promoters in the dataset. Rigorous filter methods, including tag clustering, cluster width and profile filtering, and variance filtering rescue bona fide promoters, allowing for detection of promoter usage, inter-sample promoter switching and detection of new putative promoters. These types of filtering methods could potentially also be used on other noisy next-generation data sets. Here, we present result from nanoCAGE from two different studies; data from a mouse melanoma skin cancer model, as well as data from human acute promyelocytic leukemic blast populations.

Methods

For both studies, samples were sequenced in biological triplicates on the Illumina Genome Analyzer II and the Illumina HiSeq 2000, quality validated by fastqc (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and trimmed when necessary, and were mapped to the *mus* (mm9) and *homo* (hg19) genomes by Bowtie (Langmead *et al.,* 2009).

For *homo* and *mus* nanoCAGE data from the Rac1 project, single tags were removed, and all sample were merged followed by consensus generation by merging all tags within 20bp. Next, we required a 2/1 cluster height to width ratio. Clusters having a width of 5 tags or less were removed to filter for PCR-amplification artifacts. Tags were counted in consensus clusters per sample, and expression values were quantified as TPM (tags per million mapped). Clusters with <5 TPM in the highest sample were removed to filter out noise in the lowest band. The intra-replicate coefficient of variance (CV) was calculated for all samples, and clusters with a CV higher than 1 were removed.

All statistical analyses were performed in the statistical package R (Ihaka *et al.,* 1996), and the Bioconductor package edgeR (Robinson *et al.,* 2010) was used for differential testing.

Results and Discussion

We present preliminary results from nanoCAGE in two different studies. First, we show data from nanoCAGE of epidermal cells in a model of melonama skin cancer, harvested from Rac1 KO vs. WT mice, treated with or without the proliferative agent tetradecanoylphorbol acetate (TPA). This four-way experiment allows a detailed characterization of promoter usage of treated vs. untreated mice, and how the Rac1 gene contributes to the gene expression in hyperplastic vs. normal skin cells. After the initial stringent filtering, we present a confined set of high confidence promoters (figure 1) and their interactions between the samples and treatments. Secondly, we present preliminary results from nanoCAGE-seq of blast cells of human acute promyelocytic leukemic populations versus the corresponding normal hematopoietic progenitor cell, revealing, among other things, a pattern of promoter switching from full length transcript to shorter transcripts in the cancer cells.

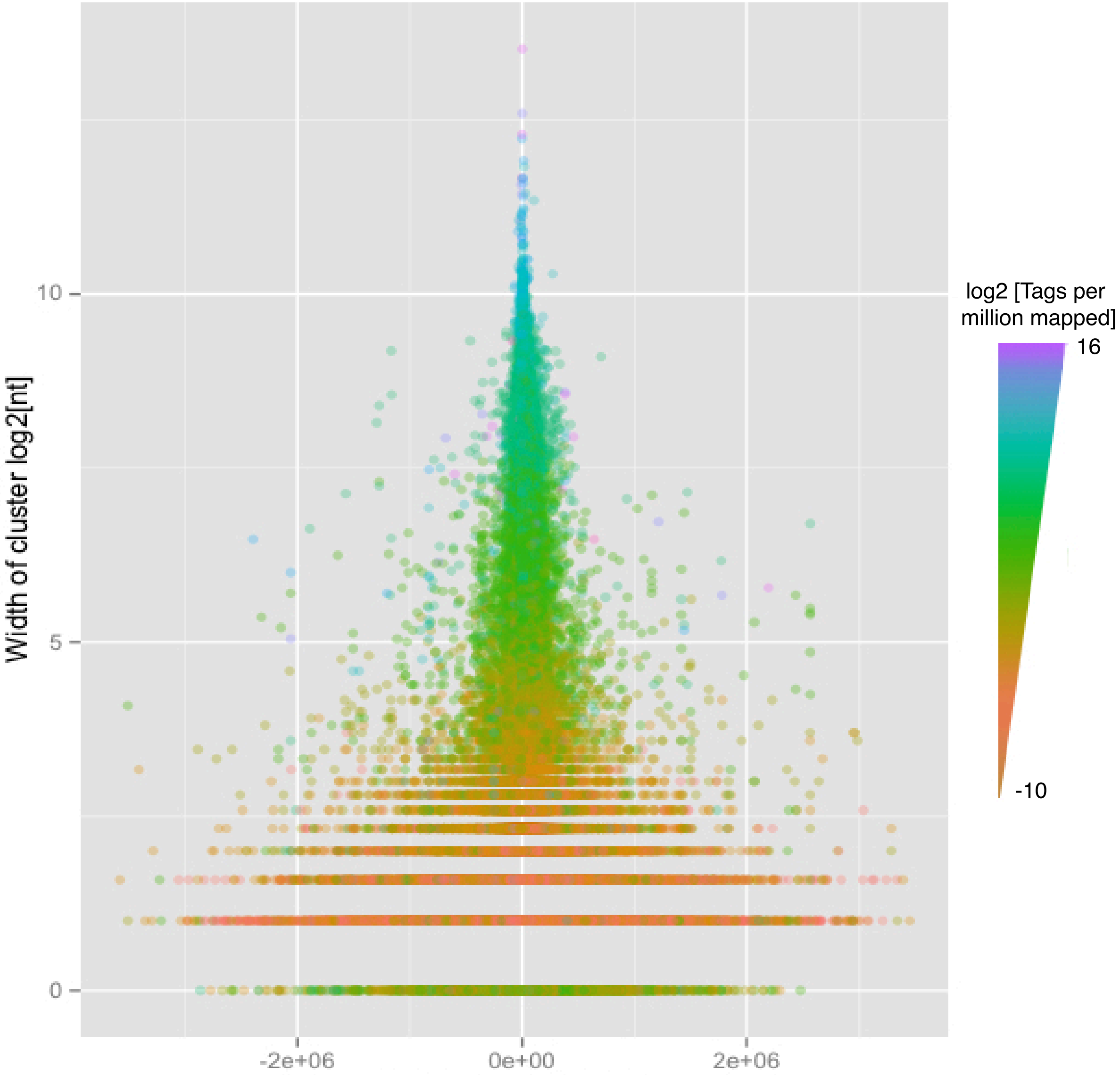

Figure 1. nanoCAGE-seq data requires rigourous filtering. Scatterplot of all clusters before filtering. X-axis: distance to nearest UCSC knownGene transcription start site, y-axis: width of cluster in nt (log2). Clusters are color-coded by expression amount (TPM). As evident, higher expressed clusters are closer to the TSS and wider, while the lowest expressed clusters, much of if noise, are spread across the genome and are slim.

References

Ihaka, Ross, and Robert Gentleman. (1996) R: A language for data analysis and graphics. *Journal of computational and graphical statistics* **5**(3), 299-314.

Langmead B *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**(3), R25.

Plessy C *et al.* (2010) Linking promoters to functional transcripts in small samples with nanoCAGE and CAGEscan. *Nature methods* **7**(7), 528-534.

Robinson MD, McCarthy DJ, and Smyth GK. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**(1), 139-140.