damidseq_pipeline: an automated pipeline for processing DamID sequencing datasets

Owen J Marshall[1]* and Andrea H Brand[1]

[1]Wellcome Trust/Cancer Research UK Gurdon Institute, Cambridge, CB2 1QN, UK

Associate Editor: Dr. Inanc Birol

ABSTRACT

Summary: DamID is a powerful technique for identifying regions of the genome bound by a DNA-binding (or DNA-associated) protein. Currently no method exists for automatically processing next-generation sequencing DamID (DamID-seq) data, and the use of DamID-seq datasets with normalisation based on read-counts alone can lead to high background and the loss of bound signal. DamID-seq thus presents novel challenges in terms of normalisation and background minimisation. We describe here damidseq_pipeline, a software pipeline that performs automatic normalisation and background reduction on multiple DamID-seq FASTQ datasets.

Availability and implementation: Open-source and freely available from http://owenjm.github.io/damidseq_pipeline. The damidseq_pipeline is implemented in Perl and is compatible with any Unix-based operating system (e.g. Linux, Mac OSX).

Contact: o.marshall@gurdon.cam.ac.uk

1 INTRODUCTION

DamID is a well-established technique for discovering regions of DNA bound by or associated with proteins (van Steensel and Henikoff, 2000). It has been used to map the genome-wide binding of transcription factors, chromatin proteins, nuclear complexes associated with DNA and RNA pol II (for e.g. Choksi *et al.*, 2006; Filion *et al.*, 2010; Southall *et al.*, 2013; Singer *et al.*, 2014). The technique can be performed in cell culture, whole organisms (van Steensel and Henikoff, 2000) or with cell-type specificity (Southall *et al.*, 2013), and requires no fixation or antibody purification.

DamID involves the fusion of a bacterial DNA adenine methylase (Dam) to any DNA-associated protein of interest. The bacterial Dam protein methylates adenine in the sequence GATC and, given that higher eukaryotes lack native adenine methylation, the DNA binding footprint of the protein of interest is uniquely detectable through isolating sequences flanked by methylated GATC sites. However, a major consideration with DamID is that any Dam protein within the nucleus will non-specifically methylate adenines in GATC sequences at accessible regions of the genome. For this reason, DamID is always performed concurrently with a Dam-only control, and the final DNA binding profile is typically presented as a $\log_2$(Dam-fusion/Dam-only) ratio.

Although the majority of published DamID experiments have used tiling microarrays for data analysis, next-generation sequencing (NGS) allows greater sensitivity and higher accuracy. While several recent studies have used NGS with DamID (Wu and Yao, 2013; Lie-A-Ling *et al.*, 2014; Clough *et al.*, 2014; Carl and Russell, 2015), these have relied upon a comparison of peak binding intensities between read-count-normalised Dam-fusion and Dam samples. Depending on the characteristics of the Dam-fusion protein (see below) this approach may lead to real signal being lost, and correct normalisation of the datasets is required to detect all binding by many Dam-fusion proteins. Here, we describe a software pipeline for the automated processing of DamID-sequencing (DamID-seq) data, including normalisation and background reduction algorithms.

2 ALGORITHMS

Although DamID-seq data can be aligned and binned as per all NGS data, two issues arise that are specific to DamID. The first major consideration is the correct normalisation of the Dam-fusion and Dam-control samples. The greatest contribution to many Dam-fusion protein datasets is the non-specific methylation of accessible genomic regions (e.g. Fig. 1B), with a mean correlation between Dam alone and Dam-fusion datasets of 0.70 (n=4, Spearman's correlation). Representing the data as a (Dam-fusion/Dam) ratio in theory negates such non-specific methylation. However, strong methylation signals at highly bound regions in the Dam-fusion dataset will reduce the relative numbers of reads present at accessible genomic regions in this dataset (see, for example, the occupancy of Dam-RNA Pol II over the *eyeless* gene in Fig. 1), and normalising the data based on read counts alone can therefore produce a strong negative bias to the ratio file (Fig. 1B (iii), Fig. S5A). Depending on the characteristics of the fusion protein, this negative bias can lead to real signal being lost (Fig. 1). Although microarray data inadvertently overcame this bias through the manual adjustment of laser intensities during microarray scanning, until now no method has existed for correctly normalising DamID-seq datasets.

In order to correct for this negative bias we use the read counts from accessible genomic regions—as determined from the Dam-only dataset—as the basis for normalisation, while avoiding regions likely to contain real signal in the Dam-fusion sample. We use the following algorithm to adjust the Dam-fusion dataset.

1. Given the GATC-site resolution of DamID, we divide the read counts into GATC fragments.
2. All GATC fragments lacking read counts are excluded. The remaining GATC fragments are divided into deciles.

*to whom correspondence should be addressed

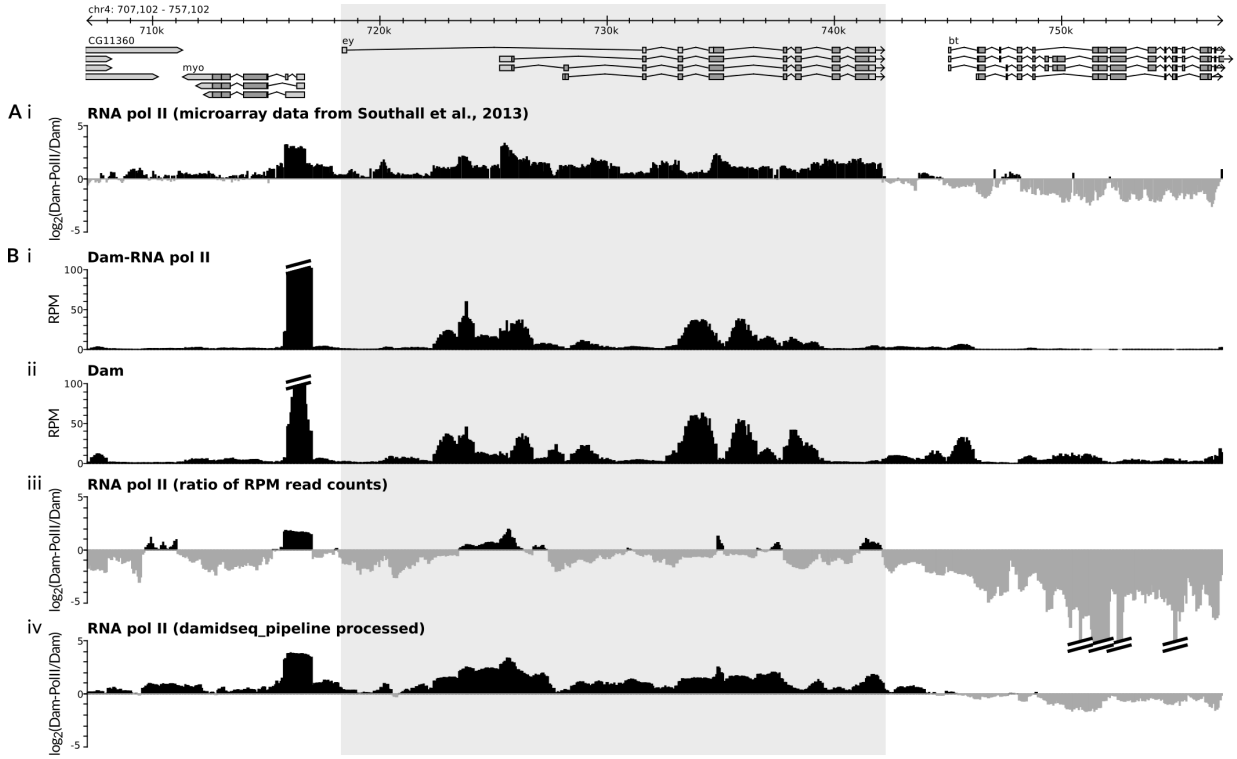

Fig. 1. Results of the damidseq_pipeline. (A) The gene *eyeless (ey)* (highlighted) is expressed in *D. melanogaster* laval neural stem cells (Southall *et al.*, 2013) and previously-published microarray DamID in these cells (i) shows RNA polymerase II occupancy (Southall *et al.*, 2013). (B) Performing DamID-seq in the same cell type illustrates the high correlation between Dam-Pol II (i) and Dam alone (ii) in terms of RPM (read counts/million mapped reads). Taking the ratio of the two RPM-normalised datasets fails to show significant RNA pol II occupancy at *ey* (iii); however, processing via the damidseq_pipeline software successfully recovers the RNA pol II occupancy profile while minimising background (iv). See Supplementary methods for experimental details.

3. Given the high probability that the highest 10% of Dam-fusion read counts represent bound signal rather than background signal, we exclude fragments that have scores in this decile.
4. The first 3 deciles of the Dam sample can generate inconsistent normalisation values if included (Table S2), so we exclude fragments that lie within this range.
5. The distribution of the $\log_2$(Dam-fusion/Dam) ratio $(x_1, x_2, \ldots, x_n)$ for all remaining fragments is determined via the gaussian kernel density estimate $\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x_i-x)^2}{2h^2}\right)$, where h is the bandwidth, estimated via the method of Silverman (1986): $h = 0.9\frac{\min(\sigma, IQR)}{1.34}n^{-1/5}$ (where σ is the standard deviation of the sample and IQR the interquartile range). For speed considerations, we estimate kernel density over 300 equally spaced points within the interval $[\max(-5, \min(x)), \min(5, \max(x))]$.
6. The point of maximum kernel density represents the point of maximum correspondence between Dam-fusion and Dam values; if both samples are correctly normalised this value should equal 0. We therefore normalise all Dam-fusion values by $1/\left(2^{\arg\max(\hat{f}_h(x))}\right)$.

In addition to ensuring correct normalisation, a second important consideration is the reduction of background noise. Regions without specific methylation will have randomly distributed background counts that, when a ratio file is generated, will generate a large degree of noise. Such noise can potentially obscure peak detection. In order to mitigate this effect we add pseudocounts to both datasets. In order to maintain equivalence between replicates with differing numbers of reads (assuming that $\text{genome}_{\text{bound}} \ll \text{genome}_{\text{unbound}}$) the number of pseudocounts added is proportional to the sequencing coverage, thus $c\frac{reads}{bins}$, where c is a constant. (see Table S1 for a comparison of gene calls with different read-depths). Adding pseudocounts increases the number and the total genomic coverage of detected peaks and increases the signal:noise ratio (Fig. S1-4).

The combination of these two methods compares favorably with previously published microarray data (Fig. 1B (iv)) or DamID-seq data (Fig. S1-4; Fig. S5).

3 IMPLEMENTATION

The damidseq_pipeline software is implemented in Perl, and will process multiple single-end read sequencing files in FASTQ or BAM format. The pipeline can match sequencing adaptors to sample names, automatically identifies the Dam-only control, and performs alignment, read-length extension, normalisation,

background reduction and ratio file generation. (See Supplementary Methods for details.)

A large number of user-configurable options are provided, including the ability to adjust the normalisation algorithm parameters, generate read-count normalised files and add a user-specified number of pseudocounts. Parameters specified on the command-line can be saved as defaults if the user desires.

The damidseq_pipeline software is open-source and freely available at http://owenjm.github.io/damidseq_pipeline. A detailed set of installation and usage instructions are provided at the above website, along with a small example dataset.

ACKNOWLEDGEMENTS

We thank Charles Bradshaw for helpful comments on the software. This work was supported by the BBSRC [BB/L00786X/1] and Wellcome Trust [092545]. The Gurdon Institute is supported by core funding from the Wellcome Trust [092096] and CRUK [C6946/A14492].

REFERENCES

Carl, S. H. and Russell, S. (2015). Common binding by redundant group B Sox proteins is evolutionarily conserved in Drosophila. *BMC Genomics*, **16**(1), 1–22.

Choksi, S. P., Southall, T. D., Bossing, T., Edoff, K., de Wit, E., Fischer, B. E., van Steensel, B., Micklem, G., and Brand, A. H. (2006). Prospero acts as a binary switch between self-renewal and differentiation in Drosophila neural stem cells. *Dev Cell*, **11**(6), 775–789.

Clough, E., Jimenez, E., Kim, Y.-A., Whitworth, C., Neville, M. C., Hempel, L. U., Pavlou, H. J., Chen, Z.-X., Sturgill, D., Dale, R. K., Smith, H. E., Przytycka, T. M., Goodwin, S. F., Van Doren, M., and Oliver, B. (2014). Sex- and Tissue-Specific Functions of Drosophila Doublesex Transcription Factor Target Genes. *Dev. Cell*, **31**(6), 761–773.

Filion, G. J., van Bemmel, J. G., Braunschweig, U., Talhout, W., Kind, J., Ward, L. D., Brugman, W., de Castro, I. J., Kerkhoven, R. M., Bussemaker, H. J., and van Steensel, B. (2010). Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. *Cell*, **143**(2), 212–24.

Lie-A-Ling, M., Marinopoulou, E., Li, Y., Patel, R., Stefanska, M., Bonifer, C., Miller, C., Kouskoff, V., and Lacaud, G. (2014). RUNX1 positively regulates a cell adhesion and migration program in murine hemogenic endothelium prior to blood emergence. *Blood*, **124**(11), e11–20.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

Singer, R., Atar, S., Atias, O., Oron, E., Segal, D., Hirsch, J. A., Tuller, T., Orian, A., and Chamovitz, D. A. (2014). Drosophila COP9 signalosome subunit 7 interacts with multiple genomic loci to regulate development. *Nucleic Acids Res.*, **42**(15), 9761–70.

Southall, T. D., Gold, K. S., Egger, B., Davidson, C. M., Caygill, E. E., Marshall, O. J., and Brand, A. H. (2013). Cell-Type-Specific Profiling of Gene Expression and Chromatin Binding without Cell Isolation: Assaying RNA Pol II Occupancy in Neural Stem Cells. *Dev. Cell*, **26**(1), 101–12.

van Steensel, B. and Henikoff, S. (2000). Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase. *Nat. Biotechnol.*, **18**(4), 424–8.

Wu, F. and Yao, J. (2013). Spatial compartmentalization at the nuclear periphery characterized by genome-wide mapping. *BMC Genomics*, **14**(1), 591.