Genomics Data 7 (2016) 144–147

Contents lists available at ScienceDirect

Genomics Data

journal homepage: www.elsevier.com/locate/gdata

Data in Brief

Transcriptional profiling of the epigenetic regulator Smchd1

Ruijie Liu [a], Kelan Chen [a,b], Natasha Jansz [a,b], Marnie E. Blewitt [a,b,*], Matthew E. Ritchie [a,b,c,*]

[a] Molecular Medicine Division, The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria 3052, Australia
[b] Department of Medical Biology, The University of Melbourne, Parkville, Victoria 3010, Australia
[c] School of Mathematics and Statistics, The University of Melbourne, Parkville, Victoria 3010, Australia

ARTICLE INFO

Article history:
Received 22 December 2015
Accepted 30 December 2015
Available online 31 December 2015

Keywords:
RNA-sequencing
voom
Sample variability
Epigenetics

ABSTRACT

Smchd1 is an epigenetic repressor with important functions in healthy cellular processes and disease. To elucidate its role in transcriptional regulation, we performed two independent genome-wide RNA-sequencing studies comparing wild-type and *Smchd1* null samples in neural stem cells and lymphoma cell lines. Using an R-based analysis pipeline that accommodates observational and sample-specific weights in the linear modeling, we identify key genes dysregulated by Smchd1 deletion such as clustered protocadherins in the neural stem cells and imprinted genes in both experiments. Here we provide a detailed description of this analysis, from quality control to read mapping and differential expression analysis. These data sets are publicly available from the Gene Expression Omnibus database (accession numbers GSE64099 and GSE65747).

Specifications	
Organism/cell line/tissue	*Mus musculus*, C57BL/6J strain for lymphoma cell lines, FVB/N/C57BL/6J F1 for neural stem cell lines.
Sex	Male.
Sequencer or array type	NSC data: Libraries prepared with the Illumina TruSeq Total Stranded RNA kit and sequenced on an Illumina HiSeq 2000 with Illumina TruSeq SBS Kit v3-HS reagents as 100 bp paired-end reads. Lymphoma data: Libraries prepared with the Illumina TruSeq RNA Sample Preparation Kit v2 and sequenced on an Illumina HiSeq 2000 with Illumina TruSeq SBS Kit v3-HS reagents as 100 bp reads (paired and single-end).
Data format	Raw (fastq) and summarized counts.
Experimental factors	RNA was obtained from Smchd1 null and wild-type samples.
Experimental features	Neural stem cells were derived from E14.5 male embryos. Lymphoma cells were derived from lethally irradiated mice transplanted with fetal liver cells from E14.5 male embryos at the time when animals for sacrificed due to end-stage lymphoma. The lymphoma cells were plated out for growth in vitro and resulting cell lines analyzed here.
Consent	All animal experiments were carried out in accordance with the Walter and Eliza Hall Institute of Medical Research Animal Ethics Committee guidelines (AEC 2011.027).
Sample source location	Melbourne, Australia.

* Corresponding authors at: Molecular Medicine Division, The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria 3052, Australia.
E-mail addresses: blewitt@wehi.edu.au (M.E. Blewitt), mritchie@wehi.edu.au (M.E. Ritchie).

1. Direct link to deposited data

http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE64099
http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE65747

2. Introduction

Smchd1 (structural maintenance of chromosomes hinge domain containing 1) is an important epigenetic modifier that has a critical role in X inactivation [1,2] and genomic imprinting [3,4]. Although initial studies of Smchd1 used these two classic models of epigenetic control, it has become clear that Smchd1 has a broader role in regulating gene expression during normal development [5], in cancer [6] and in the development of facioscapulohumeral muscular dystrophy (FSHD) [7–9].

We were particularly interested to look at the role of Smchd1 in regulating gene expression via RNA sequencing (RNA-seq), as Smchd1 is a repressor protein, and so the very low level of expression of Smchd1 repressed genes best lends itself to RNA-seq over array-based platforms. To this end, we conducted RNA-seq experiments in two model systems, the first was in neural development using neural stem cells and the second was in a cancer model using lymphoma cell lines. In both experiments, samples with wild-type levels of Smchd1 are compared to samples with a null allele of this gene. This article describes our analyses of these two data sets, using a consistent, R-based pipeline that can deal with both observational and sample-level heterogeneity.

http://dx.doi.org/10.1016/j.gdata.2015.12.027

3. Experimental design, materials and methods

3.1. Mouse strains and sample information

MommeD1 mutant mice were maintained on the FVB/N inbred background, and backcrossed with C57BL/6 mice for more than 15 generations to produce C57BL/6 MommeD1 congenic mice (as previously described in [1]). Neural stem cells were isolated and cultured from the brains of FVB/C57BL/6J F1 E14.5 male embryos, homozygous or wild-type for the *Smchd1*$^{\mathrm{MommeD1}}$ mutation as described in [5]. Lymphoma cell lines were derived from a gene trap allele of Smchd1, described in [6]. This allele was backcrossed onto C57BL/6J, then crossed onto the Eμ-Myc transgenic background to generate *Smchd1*$^{gt/gt}$ *Eμ-MycTg*/+ embryos and their wild-type controls, for transplant and generation of lymphomas. Genotyping was carried out as described in [1,2] and [6]. Experimental animals were treated in accordance with the Australian Government National Health and Medical Research Council guidelines under the approval from the animal ethics committees of the Walter and Eliza Hall Institute (WEHI AEC 2011.027).

3.2. RNA-seq sample preparation and sequencing

Qiagen RNeasy Mini kits were used to extract RNA from *Smchd1*$^{\mathrm{MommeD1/MommeD1}}$ and *Smchd1*$^{+/+}$ wild-type NSCs according to the manufacturer's instructions. RNA was quantified using the NanoDrop 1000 Spectrophotometer (Thermo Scientific) and RNA integrity assessed with the Agilent Bioanalyzer 2100 (Agilent Technologies). Illumina's TruSeq total RNA sample preparation kit was used to prepare libraries for sequencing, which was performed by the Australian Genome Research Facility (Melbourne, Australia) on the Illumina HiSeq 2000 platform to obtain 100 bp paired-end reads.

For the Lymphoma data set, Qiagen RNeasy Mini kits were used to extract RNA from *Smchd1*$^{MommeD1/MommeD1}$;*E*μMycTg/+ and *Smchd1*$^{+/+}$;*E*μMycTg/+ lymphoma cells. Samples were prepared for sequencing at the Australian Genome Research Facility where quality control, library preparation (using Illumina's TruSeq RNA sample preparation kit) and sequencing on the Illumina HiSeq 2000 platform was performed to obtain 100 bp paired-end (for 6 out of 7 samples) or single-end (for 1 sample) reads.

3.3. Quality control and data pre-processing

The FastQC software [10] was used to assess the quality of the raw sequence data. Fig. 1 displays the distribution of sequencing quality (Phred) scores at each base position across reads from a representative RNA-seq sample from each data set. Although variation in base quality is observed across the read, with slightly lower quality at the beginning and end, median quality is above 34 (corresponding to a probability of an incorrect base call below 0.0004) for the entire read. Similar boxplots of base quality scores were observed for other samples (data not shown).

Sequences were then mapped to the mouse reference genome (mm10) using the *Rsubread* program [11] and gene-level counts were obtained by the *featureCounts* procedure [12].

Further analysis was carried out using the *edgeR* [13] and *limma* [14] R/Bioconductor packages. Counts-per-million (CPM) were calculated for each gene to standardize for differences in library-size and filtering was carried out to retain genes with a baseline expression level of at least 0.5 CPM in 3 or more samples. For each data set, TMM normalization [15] was applied and a multidimensional scaling (MDS) plot based on the $\log_2$(CPM) was generated to show relationships between samples (Fig. 2). In both data sets, we observe samples that do not cluster well with their respective replicates of the same genotype. Sample 6 in the NSC data (Fig. 2A) and samples 1 and 7 in the Lymphoma data (Fig. 2B) are more variable than the other replicates of the same type. For NSC sample 6 and Lymphoma sample 7, there was no experimental factor that could be identified to explain this phenomenon. Lymphoma sample 1 on the other hand was the only single-end sample in this experiment that was processed on a different day to the other samples, leading us to conclude that batch processing differences was the likely cause of the additional variation.

3.4. Differential expression analysis

Based on inspection of the MDS plots, which showed variability between replicate samples, linear models [16] with combined observational and sample weights [17,18] were fitted to the $\log_2$(CPM) to summarize over replicate samples. This strategy, implemented in the *voomWithQualityWeights* function, down-weights low abundance observations, which are systematically more variable (Fig. 2C) and

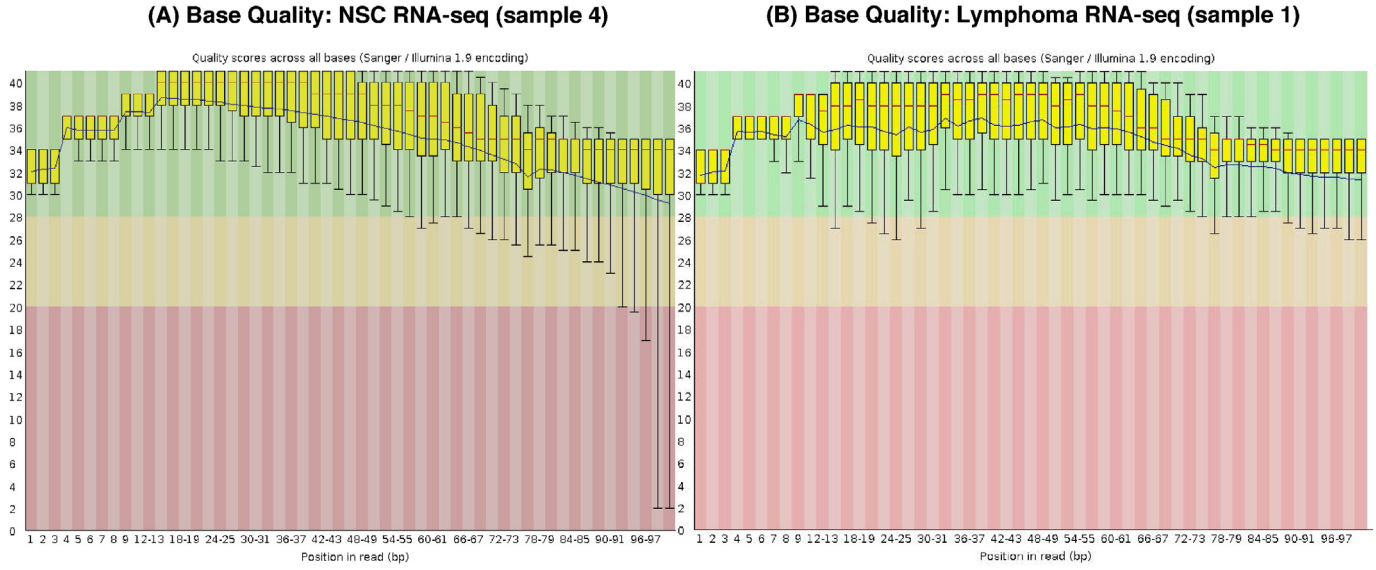

Fig. 1. Quality assessment at the read level. Boxplots of base-calling Phred scores at different base positions across all the reads in representative libraries from NSC RNA-seq (A) and Lymphoma cell line RNA-seq (B) experiments generated by FastQC. The box represents 25% and 75% quantiles of the scores with median score marked by the red line. Whiskers mark the 10% and 90% quantiles and blue lines show the mean quality score.

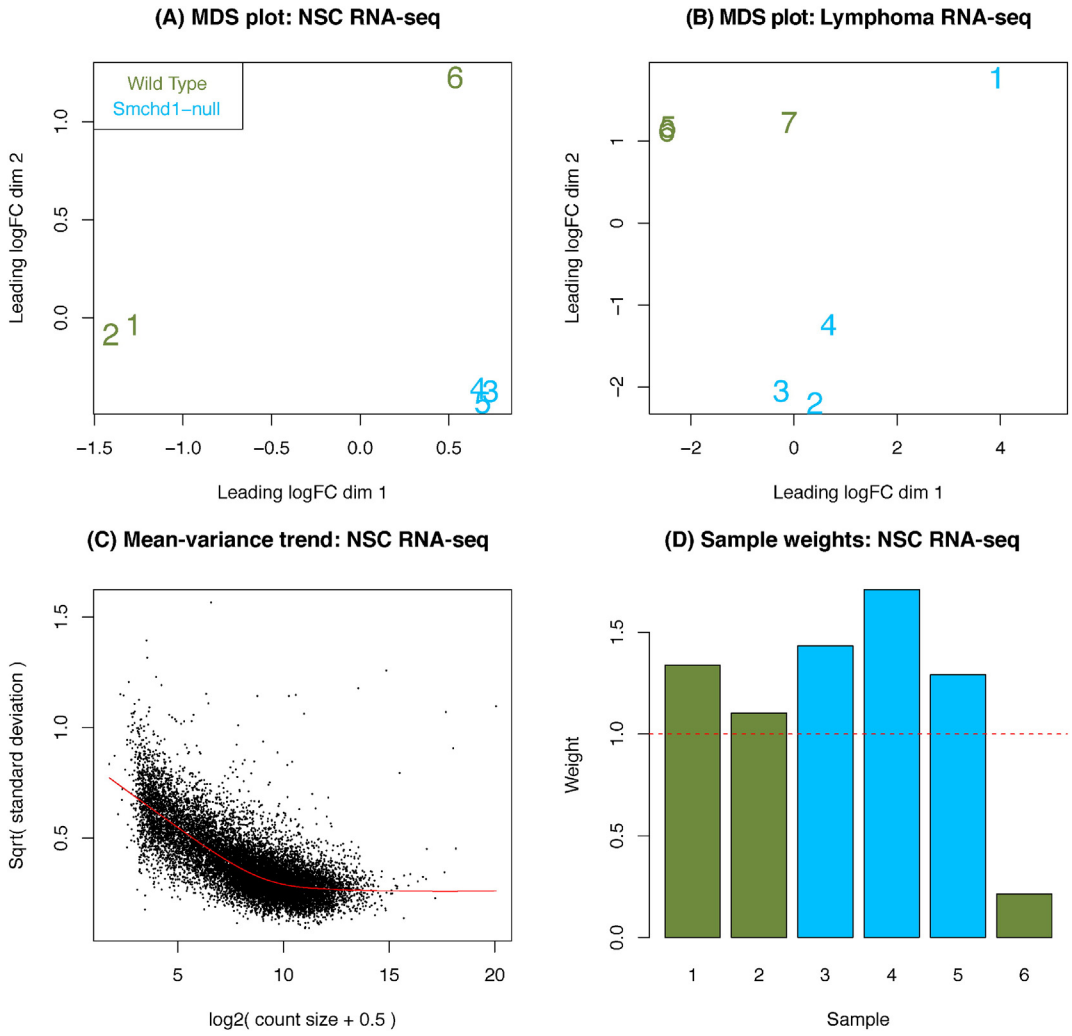

Fig. 2. Quality assessment at the sample level. Multi-dimensional scaling (MDS) plots of the NSC (A) and Lymphoma (B) data sets, with samples numbered and color coded by genotype. Distances correspond to the mean $\log_2$fold-change for the top 500 genes that best discriminate each pair of samples. In both experiments, one or more samples cluster poorly with replicates of the same genotype, motivating the use sample weights (D) in the regression modeling to detect differential expression. Panel C shows a scatterplot of the mean–variance relationship in abundance estimated from biological replicates from the NSC data set using the *voom* method. Panel D shows the sample weights estimated for the NSC data set that are combined with *voom*'s abundance-related weights in the *voomWithQualityWeights* function and used in the linear model analysis to detect differentially expressed genes.

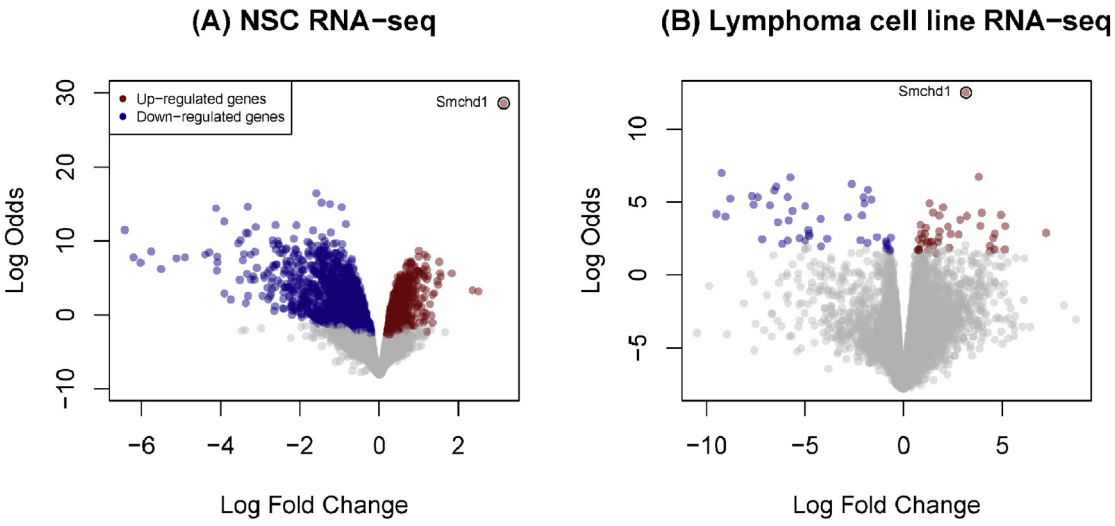

Fig. 3. Summary of the RNA-seq results. Volcano plot representation of differential expression analysis of genes in the *Smchd1* wild-type versus *Smchd1* null comparison for the NSC (A) and Lymphoma RNA-seq (B) data sets. Red and blue points mark the genes with significantly increased or decreased expression respectively in *Smchd1* wild-type compared to *Smchd1* null samples (FDR < 0.01). The x-axis shows $\log_2$fold-changes in expression and the y-axis the log odds of a gene being differentially expressed. In both data sets, *Smchd1* is the top ranked gene.

observations from entire samples that show higher variation (Fig. 2D) to get more precise estimates of gene expression and increase power to detect changes. Moderated *t*-statistics were used to assess differential expression between *Smchd1*$^{+/+}$ wild-type and *Smchd1*$^{MommeD1/MommeD1}$ samples, with genes ranked according to their false discovery rate [19]. Log-odds of differential expression [20] were also calculated. Both raw and summary-level count data for these experiments are available under GEO series accession numbers GSE64099 and GSE65747.

4. Results

At a false discovery rate (FDR) cut-off of 1%, there are 2838 differentially expressed genes (1282 up-regulated and 1556 down-regulated) in the comparison of *Smchd1* wild-type and *Smchd1* null NSC samples. The same comparison in the Lymphoma data set detected 90 genes (45 up-regulated and 45 down-regulated). These genes are highlighted in Fig. 3A and B respectively. In both analyses, *Smchd1* is the top ranked gene with $\log_2$fold-change greater than 3.1.

The NSC analysis revealed that a number of protocadherin genes, especially those from the alpha and beta clusters, were significantly differentially expressed, with down-regulation of 11 alpha cluster genes and 20 beta cluster genes. This finding is in line with studies performed in other tissues and cell lines where Smchd1-deficiency is concomitant with increased expression of protocadherin genes [3,4,6]. However, the widespread impacts observed in this analysis suggest that Smchd1 plays a critical role in regulating the protocadherin clusters in NSCs. Imprinted genes, such as *Ndn*, *Mkrn3* and *Peg12* were down-regulated by almost 2-fold, indicative of loss of imprinting in the absence of Smchd1, also in agreement with results of previous studies [3,4,6].

Genes uncovered in the Lymphoma analysis are consistent with previous reports in a different system that profiled male embryos [2], where the expression of imprinted genes such as *Peg12* and *Mkrn3* was shown to be disturbed in the absence of Smchd1. However it is interesting to note that *Peg12* and *Mkrn3* are much more strikingly down-regulated in the Lymphoma data set than in the NSCs as they are normally only very lowly expressed in the lymphoma cell lines. This may represent not just loss of imprinting, as has been shown previously [2,3], but also potential activation independent of imprinting status.

The modest number of differentially expressed genes identified in the Lymphoma data set is influenced in part by a suspected batch processing difference mentioned earlier, but also by the increased genetic heterogeneity present in profiles obtained from tumor samples. In contrast, many more genes are detected in the NSC experiment, where the samples are genetically equivalent and much less heterogeneous and genetically unstable than the lymphoma cell lines.

5. Discussion

In this report we provide a detailed description of the analysis of the RNA-seq data from [5,18] made possible using an R-based processing pipeline in the *Rsubread* and *limma* packages. In particular, the *voomWithQualityWeights* function in *limma* allows more variable samples to be down-weighted in the analysis [18]. In each case, the decision to use this approach was guided by inspection of the MDS plot to assess how well replicate samples clustered. This methodology is generally applicable to analyses of designed RNA-seq experiments, where variations in sample quality are frequently observed and the source of such variation is generally unknown. Scripts and data to reproduce this analysis are available from http://bioinf.wehi.edu.au/folders/smchd1/.

Acknowledgments

This work was supported in part by an Australian National Health and Medical Research Council grant to MB and MR (GNT1045936). MB is an Australian Research Council Queen Elizabeth II fellow (DP1096092). KC and NJ hold Australian Postgraduate Awards from the Australian National Health and Medical Research Council. This work was made possible through Victorian State Government Operational Infrastructure Support and the Australian Government NHMRC IRIISS.

References

[1] M.E. Blewitt, et al., SmcHD1, containing a structural-maintenance-of-chromosomes hinge domain, has a critical role in X inactivation. Nat. Genet. 40 (5) (2008) 663–669.
[2] A.V. Gendrel, et al., Smchd1-dependent and -independent pathways determine developmental dynamics of CpG island methylation on the inactive X chromosome. Dev. Cell 23 (2) (2012) 265–279.
[3] A.W. Mould, et al., Smchd1 regulates a subset of autosomal genes subject to monoallelic expression in addition to being critical for X inactivation. Epigenetics Chromatin 6 (1) (2013) 19.
[4] A.V. Gendrel, et al., Epigenetic functions of Smchd1 repress gene clusters on the inactive X chromosome and on autosomes. Mol. Cell. Biol. 33 (16) (2013) 3150–3165.
[5] K. Chen, et al., Genome-wide binding and mechanistic analyses of Smchd1-mediated epigenetic regulation. Proc. Natl. Acad. Sci. U. S. A. 112 (27) (2015) E3535–E3544.
[6] H.S. Leong, et al., Epigenetic regulator Smchd1 functions as a tumor suppressor. Cancer Res. 73 (5) (2013) 1591–1599.
[7] R.J. Lemmers, et al., Digenic inheritance of an SMCHD1 mutation and an FSHD-permissive D4Z4 allele causes facioscapulohumeral muscular dystrophy type 2. Nat. Genet. 44 (12) (2012) 1370–1374.
[8] S. Sacconi, et al., The FSHD2 gene SMCHD1 is a modifier of disease severity in families affected by FSHD1. Am. J. Hum. Genet. 93 (4) (2013) 744–751.
[9] M. Larsen, et al., Diagnostic approach for FSHD revisited: SMCHD1 mutations cause FSHD2 and act as modifiers of disease severity in FSHD1. Eur. J. Hum. Genet. 23 (6) (2015) 808–816.
[10] S. Andrews, FastQC: A Quality Control Tool for High Throughput Sequence Data. 2015 Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.
[11] Y. Liao, G.K. Smyth, W. Shi, The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Res. 41 (10) (2013), e108.
[12] Y. Liao, G.K. Smyth, W. Shi, featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30 (7) (2014) 923–930.
[13] M.D. Robinson, D.J. McCarthy, G.K. Smyth, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26 (1) (2010) 139–140.
[14] M.E. Ritchie, et al., limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 43 (7) (2015), e47.
[15] M.D. Robinson, A. Oshlack, A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 11 (3) (2010) R25.
[16] G.K. Smyth, Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat. Appl. Genet. Mol. Biol. **3** (2004) (p. Article3).
[17] C.W. Law, et al., voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 15 (2) (2014) R29.
[18] R. Liu, et al., Why weight? Modelling sample and observational level variability improves power in RNA-seq analyses. Nucleic Acids Res. 43 (15) (2015), e97.
[19] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate — a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B Methodol. 57 (1) (1995) 289–300.
[20] I. Lonnstedt, T.P. Speed, Replicated microarray data. Stat. Sin. 12 (2002) 31–46.