Genomics Data 5 (2015) 346–351

Contents lists available at ScienceDirect

Genomics Data

journal homepage: http://www.journals.elsevier.com/genomics-data/

Data in Brief

Transcriptome and H3K27 tri-methylation profiling of Ezh2-deficient lung epithelium

Aliaksei Z. Holik [a,b,*], Laura A. Galvis [a], Aaron T.L. Lun [b,c], Matthew E. Ritchie [b,d,e], Marie-Liesse Asselin-Labat [a,b,*]

[a] ACRF Stem Cells and Cancer 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] Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria 3052, Australia
[d] Molecular Medicine Division, The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria 3052, Australia
[e] School of Mathematics and Statistics, The University of Melbourne, Parkville, Victoria 3010, Australia

ARTICLE INFO

Article history:
Received 9 July 2015
Accepted 12 July 2015
Available online 16 July 2015

Keywords:
Ezh2
Lung
Development
RNA-seq
ChIP-seq

ABSTRACT

The adaptation of the lungs to air breathing at birth requires the fine orchestration of different processes to control lung morphogenesis and progenitor cell differentiation. However, there is little understanding of the role that epigenetic modifiers play in the control of lung development. We found that the histone methyl transferase Ezh2 plays a critical role in lung lineage specification and survival at birth. We performed a genome-wide transcriptome study combined with a genome-wide analysis of the distribution of H3K27 tri-methylation marks to interrogate the role of Ezh2 in lung epithelial cells. Lung cells isolated from Ezh2-deficient and control mice at embryonic day E16.5 were sorted into epithelial and mesenchymal populations based on EpCAM expression. This enabled us to dissect the transcriptional and epigenetic changes induced by the loss of Ezh2 specifically in the lung epithelium. Here we provide a detailed description of the analysis of the RNA-seq and ChIP-seq data, including quality control, read mapping, differential expression and differential binding analyses, as well as visualisation methods used to present the data. These data can be accessed from the Gene Expression Omnibus database (super-series accession number GSE57393).

Specifications	
Organism/cell line/tissue	*Mus musculus*, C57BL/6J strain carrying transgenic Ezh2 and Shh alleles
Sex	Mixed gender
Sequencer or array type	*RNA-seq*: Libraries were prepared with the Illumina TruSeq Total Stranded RNA kit with Ribo-Zero and sequenced on HiSeq 2000 with TruSeq SBS Kit v3 — HS reagents (Illumina) as 100 bp single end reads. *ChIP-seq*: Libraries were prepared with the Illumina TruSeq Nano kit and sequenced on HiSeq 2500 with TruSeq Rapid SBS Kit — HS reagents as 100 bp single end reads.
Data format	Raw (fastq) and analysed
Experimental factors	*RNA-seq*: RNA was obtained from epithelial and mesenchymal cell populations from control and Ezh2 deficient lungs. *ChIP-seq*: H3K27me3 ChIP was performed on epithelial cells from control and Ezh2-deficient lungs.
Experimental features	Epithelial and mesenchymal cell populations were derived from the lungs of Ezh2-deficient and control mouse embryos at day E16.5. Both cell populations were profiled for gene expression (Total RNA-seq). Epithelial cells were additionally subjected to genome-wide analysis of H3K27 tri-methylation (ChIP-seq).
Consent	All animal experiments were carried out in accordance with the Walter and Eliza Hall Institute of Medical Research Animal Ethics Committee guidelines (AEC 2010.017).
Sample source location	Melbourne, Australia

1. Direct link to deposited data

http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE57393.

2. Experimental design, materials and methods

2.1. Mouse strains

Shh-cre mice expressing cre recombinase under the control of Sonic Hedgehog (*Shh*) promoter [1] were purchased from the Jackson Laboratory. Mice bearing a loxP-targeted *Ezh2* allele ($Ezh2^{fl}$) were obtained

* Corresponding authors at: ACRF Stem Cells and Cancer Division, The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria 3052, Australia.
E-mail addresses: aliakseiholik@gmail.com (A.Z. Holik), labat@wehi.edu.au (M.-L. Asselin-Labat).

http://dx.doi.org/10.1016/j.gdata.2015.07.006

from Prof Tarakhovsky at Rockfeller University, NY [2]. Animals were genotyped as described in the respective publications. All animal experiments were carried out in accordance with the Walter and Eliza Hall Institute of Medical Research Animal Ethics Committee guidelines (AEC 2010.017).

We generated *Shh-cre*;*Ezh2*$^{fl/fl}$ mice, in which the catalytic SET domain of *Ezh2* is excised from day E9.5 in the epithelium of the lung primordia. Since the *cre* transgene is knocked into the *Shh* locus, effectively rendering *Shh-cre* animals heterozygous for the *Shh* allele, we used *Shh-cre*;*Ezh2*$^{fl/+}$ animals as controls.

2.2. RNA-seq: sample preparation and sequencing

Lungs from *Shh-cre*;*Ezh2*$^{fl/fl}$ and control embryos were harvested at day E16.5 and separated into epithelial (EpCAM$^+$) and mesenchymal (EpCAM$^-$) cell populations as described previously [3]. If necessary, EpCAM$^+$ cells were pooled from several embryos of the same genotype to obtain a minimum of 10^5 cells. Each genotype/cell type combination had 3 biological replicates. Total RNA was extracted with the Total RNA Purification Kit (Norgen) according to manufacturer instructions. RNA integrity was assayed on the Tapestation machine (Agilent) using R6K screentape.

RNA-seq libraries were prepared from 150 ng of total RNA using the TruSeq Stranded Total RNA kit with Ribo-Zero (Illumina) according to the kit guidelines. Libraries were quantified using Tapestation (Agilent) to estimate the average fragment size and Broad Range Qubit reagent (Life Technologies) to accurately estimate library concentration. Quantified libraries were pooled at equimolar concentrations and sequenced as 100 bp single-end reads on a HiSeq 2000 machine with TruSeq SBS Kit v3 — HS reagents (Illumina) at the Australian Genome Research Facility (AGRF).

2.3. ChIP-seq: sample preparation and sequencing

EpCAM$^+$ lung cells from *Shh-cre*;*Ezh2*$^{fl/fl}$ and control embryos were collected as described above for the RNA-seq experiment. Chromatin immunoprecipitation using the H3K27me3 antibody (Millipore #07-449) was carried out, as described previously (see Supplementary materials in Galvis et al. [3] for a detailed description of the procedure). DNA concentration was quantified using Broad Range Qubit reagent (Life Technologies). 20–30 ng of immunoprecipitated DNA from each of the samples (two biological replicates for each genotype) was subjected to NGS library preparation using the TruSeq Nano DNA Sample Preparation Kit (Illumina) according to the kit manual. We made the following amendments to the library preparation protocol: fragmentation and size selection steps were omitted, such that we started the protocol from the end-repair step, proceeding directly to the 3′-adenylation step. Additionally, 2 extra amplification cycles were included during the fragment enrichment step (10 amplification cycles in total) to compensate for reduced input amount of immunoprecipitated samples (~1/4 of the recommended amount). Resulting libraries were size selected using the Pippin Prep DNA Size Selection System (1.5% cassette, Sage Science) to ensure that fragment sizes were below 900 bp. Libraries were quantified as described above for RNA-seq, pooled at equimolar concentrations and sequenced as 100 bp single end reads on a HiSeq 2500 machine using TruSeq Rapid SBS Kit — HS reagents (Illumina) at the AGRF.

2.4. Sequencing quality

Quality control of sequencing output was carried out using the FastQC software [4]. Fig. 1 displays the distribution of sequencing quality (Phred) scores at each base position across all reads in a representative RNA-seq (Fig. 1A) and ChIP-seq (Fig. 1B) library. Although the median sequencing quality is reduced towards the 3′-end of the read, the majority of sequencing scores are above 30 across the length of the read, corresponding to a probability of an incorrect base call below 0.001. A similar pattern of sequencing quality scores was observed across all libraries, in both the RNA-seq and ChIP-seq experiments.

2.5. Read mapping and summarisation

Reads from both experiments (RNA-seq and ChIP-seq) were mapped to the *mm10* build of the mouse reference genome using the Rsubread aligner (version 1.13.25) [5] with default settings. Notably, only unique reads were retained and the Hamming distance was used to break the ties for the reads with more than one best mapping location. Resulting BAM files were sorted and indexed using the SAMtools software suite [6].

Mapped RNA-seq reads were summarised at the gene level according to the NCBI RefSeq annotation (*mm10* genome assembly) in a strand-specific manner. Read summarisation was performed using featureCounts [7] function in the Rsubread package with default settings, except with the *strandSpecific* option set to 2 (reversely stranded).

In the RNA-seq experiment, the number of reads per sample ranged from 26 to 35 million. On average, 84% of all reads mapped to the reference genome (range 80–89%) and 58% of all reads mapped to known genomic features (range 56–61%). For the ChIP-seq experiment, the number of reads per sample ranged from 21 to 27 million, with an average mapping rate of 84%. The mapping rate was slightly lower for Ezh2-deficient samples compared to control samples (82% and 86% respectively).

2.6. RNA-seq analysis

In order to avoid excessive variability associated with lowly expressed genes, we removed genes that failed to achieve a count per million (CPM)

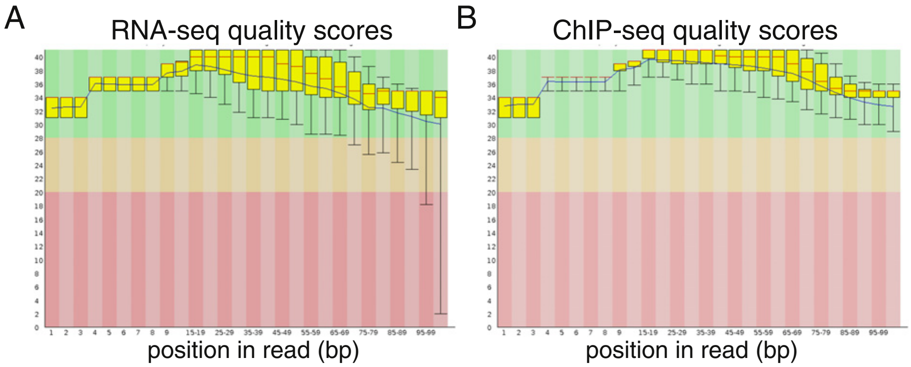

Fig. 1. Distribution of base-calling Phred scores at different base positions across all the reads in representative libraries from RNA-seq (A) and ChIP-seq (B) experiments. The box represents 25% and 75% quantiles of the scores with median score marked by the red line. Whiskers demarcate 10% and 90% quantiles. Blue line represents mean quality score.

above 0.5 in at least 3 libraries. To further limit the analysis space, we also removed predicted and pseudo-genes, genes without annotation and genes that mapped to the Y or mitochondrial chromosomes, leaving 14,831 genes available for the differential expression analysis. Normalisation factors for library sizes were calculated using the trimmed mean of M-values (TMM) method [8] from the edgeR package (version 3.10.2) [9]. A multi-dimensional scaling plot of normalised samples revealed strong clustering according to genotype and tissue of origin, as well as good reproducibility among biological replicates (Fig. 2A).

We used the voom method [10] to transform the count data and derive observational-level weights. These were used to fit gene-wise linear models, followed by differential expression tests using empirical Bayes moderated t-statistics [11]. Fig. 2B illustrates the general relationship between gene expression levels and their variability. The plot shows a characteristic trend of decrease in variance with increase in average expression. Notably, there is no drop in variance at the low end of the mean expression values suggesting that we have successfully filtered out lowly expressed genes.

A conventional approach in differential expression analyses is to test for any differential expression (i.e. the null hypothesis tested is that there is no change in gene expression) and combine the resulting false discovery rate (FDR) with an arbitrary fold-change cut-off in order to limit the findings to a 'biologically relevant' subset of genes. This approach, however, fails to properly control the type I error rate [12]. In order to avoid this problem, we used the TREAT function [12] from the limma package [13] to formally test if the expression change was greater than a biologically relevant threshold (in this case, 1.2-fold). We controlled the false discovery rate (FDR) at 5% by applying the Benjamini–Hochberg method [14] across the TREAT p-values.

Comparison between the epithelial cells from Ezh2-deficient and control epithelium yielded 1623 differentially expressed genes, 1150 of which were up-regulated and 473 down-regulated (Fig. 2C). Although Ezh2 deletion in *Shh-cre*;*Ezh2*$^{fl/fl}$ lungs is confined to epithelial cells, we also detected a small number of differentially expressed genes in the mesenchymal cell population with 5 genes increasing and 6 genes decreasing in expression (Fig. 2D). That number increased to 43 and 93 up- and down-regulated genes respectively, if we tested for any changes in expression (i.e. the gene expression fold change is significantly different from 1).

2.7. ChIP-seq analysis

To perform a differential binding analysis of H3K27 tri-methylation between control and Ezh2-deficient lung epithelium we used the csaw package [15]. We counted reads from each immunoprecipitated sample into adjoining 2 kb bins spanning the entire genome. Reads mapped to genomic regions marked as repeat sequences by RepeatMasker [16] from the UCSC server (http://hgdownload.cse.ucsc.edu/goldenPath/mm10/bigZips/chromOut.tar.gz) were excluded. To remove low-abundance bins corresponding to putative regions of non-specific binding, we calculated the average log-count per million (logCPM) for each bin across all samples using the aveLogCPM function in the edgeR package. We excluded low-abundance bins with an average logCPM below 0.5, yielding 157,664 bins for further analysis.

In order to correct for potential composition bias, we counted the reads into contiguous 10 kb bins for each library and used these counts to compute normalisation factors with the TMM method [8]. These factors, in turn, were used to calculate effective library sizes for the subsequent analysis with the 2 kb bins. Fig. 3A contains log$_2$ fold-change (logFC) versus average abundance plots between one of the control libraries and every other library prior to normalisation, as well as the log-ratio of normalisation factors between each pair of libraries (red

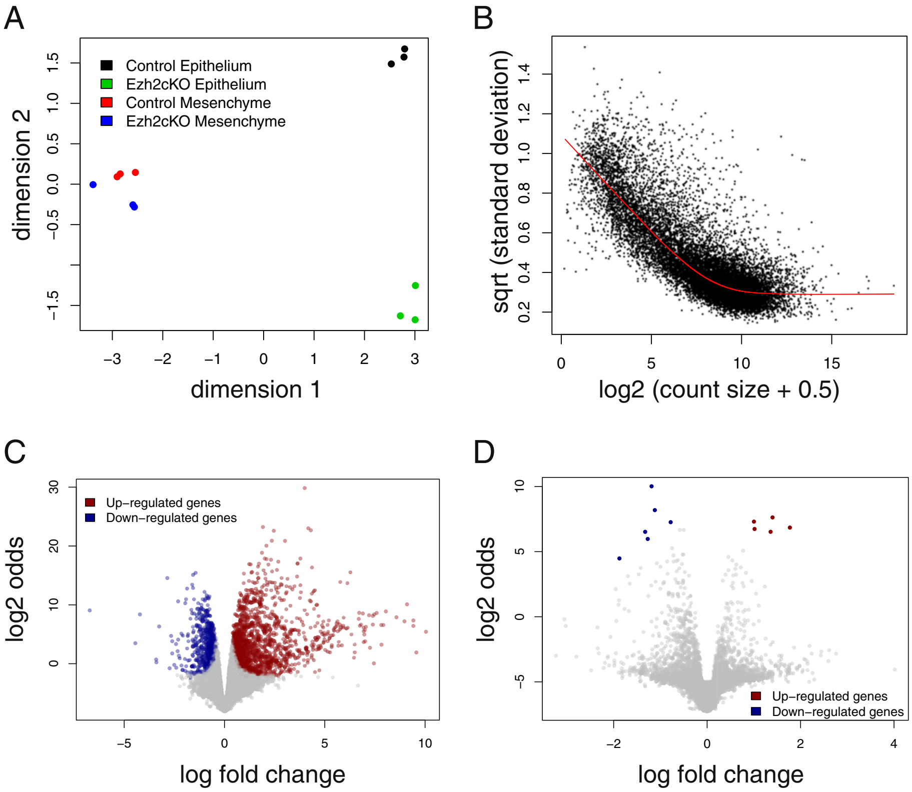

Fig. 2. RNA-seq analysis. A. Unsupervised clustering of filtered and normalised libraries by multi-dimensional scaling revealed a strong segregation of samples according to genotype and cell type. Distances on the plot correspond to the mean log$_2$ fold-change for the top 500 genes that discriminate each pair of samples. B. Scatterplot representing mean-variance relationship in the count data estimated from biological replicates. C–D. Volcano plot representation of differential expression analysis of epithelial (C) and mesenchymal (D) samples. Red and blue points mark the genes with significantly increased and reduced expression respectively in Ezh2-deficient lungs compared to control samples. The x axis shows log$_2$ fold-change in expression and the y axis gives log odds of significance p values.

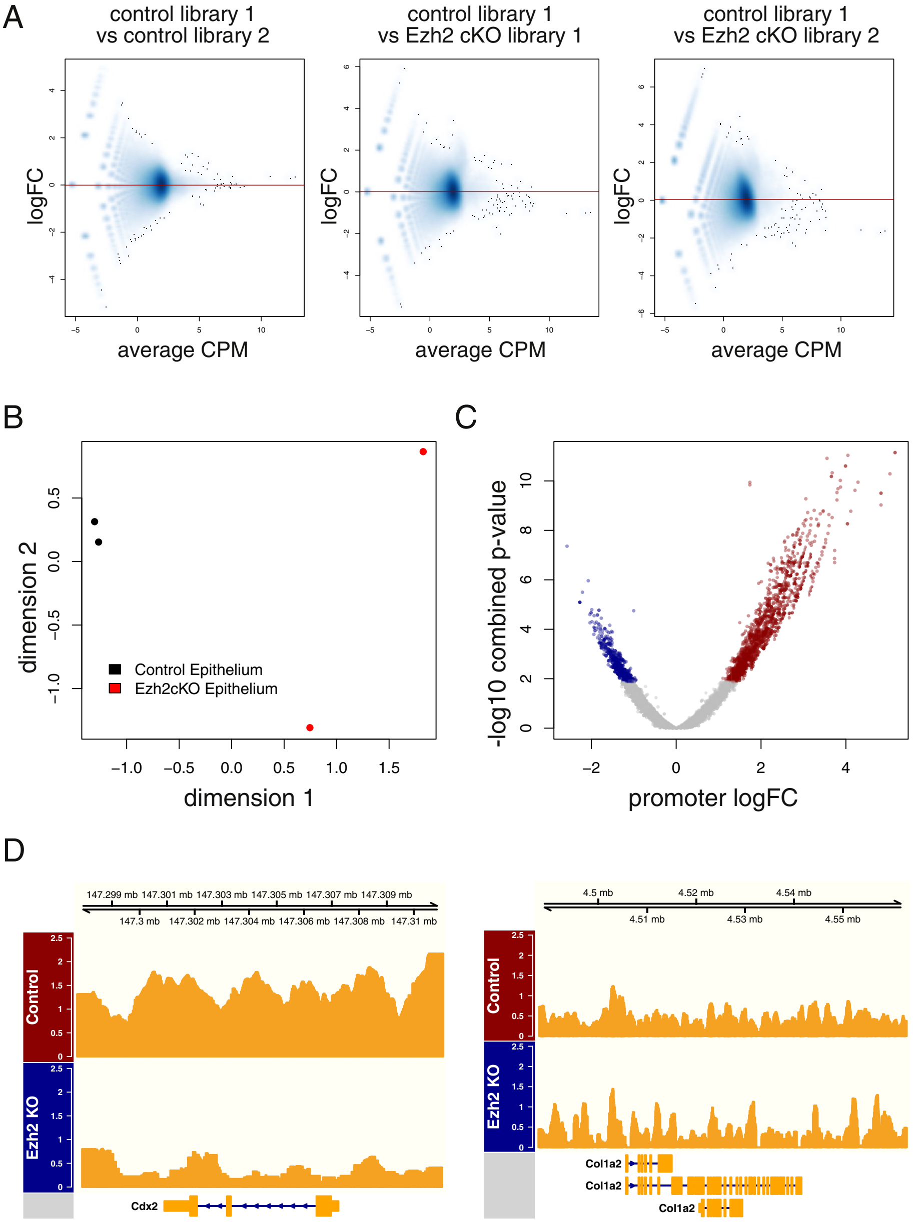

Fig. 3. ChIP-seq analysis. A. Smoothed scatter plots of means (x-axis) and differences (y-axis) between $\log_2$ counts per million of one of the control libraries and every other library. The high-density cloud of points in the centre of each plot corresponds to 10 kb bins containing background regions. The red line represents a between-libraries scaling factor to be used for normalisation of library sizes. B. Unsupervised clustering of samples by multi-dimensional scaling based on the $\log_2$ fold-change of top 500 2 kb bins that distinguish each pair of samples. C. Volcano plot representation of differential marking (DM) analysis between control and Ezh2-deficient epithelium. Each point represents a promoter; the x-axis represents the $\log_2$ FC of each promoter, defined as the $\log_2$ FC of the overlapping bin with the lowest p-value; the y-axis contains $-\log_{10}$-transformed combined p value for each promoter. Promoters that display significantly higher or lower levels of H3K27 tri-methylation in control epithelium compared to Ezh2-deficient samples are marked in red and blue respectively. D. Examples of H3K27me3 marking around the genes constitutively repressed (*Cdx2* — marked by H3K27me3) and constitutively expressed (*Col1a2* — unmarked by H3K27me3) in the lung epithelium. The y-axis represents read coverage (in counts per 10 million) at each position with a smoothing window of 1 kb.

line). A multidimensional scaling plot using filtered and normalised bin counts revealed a strong separation between the genotypes, as well as good clustering of the control libraries, while the Ezh2-deficient libraries were less consistent (Fig. 3B).

To detect the 2 kb bins that were differentially marked between control and Ezh2-deficient epithelium, we used the quasi-likelihood (QL) negative binomial framework [17] in the edgeR package [9]. Briefly, we first estimated an abundance-dependent trend in the negative binomial (NB) dispersions across all bins [18]. Using the trended NB dispersion, we fitted a generalised linear model (GLM) to the counts for each bin. The log-effective library sizes were used as offsets during GLM fitting. We estimated the QL dispersion for each bin from the GLM deviance, and fit an abundance-dependent trend to these estimates across all bins [17]. We then shrunk the QL dispersion estimate for each bin towards the trend, using a robust empirical Bayes strategy. Finally, a p-value was computed for each bin using the QL F-test.

To summarise the bins at the promoter level, we identified sets of bins overlapping NCBI RefSeq gene promoters, defined as the 6 kb region centred around each transcription start site (TSS). We limited the promoters to those genes that were present in our filtered RNA-seq dataset, resulting in 14,847 promoters. We then computed a combined p-value for each promoter from the p-values of all overlapping bins using Simes' method [19]. We applied the Benjamini–Hochberg correction method [14] to control the FDR across promoters at 5%. The logFC for each promoter was defined as the logFC of the bin with the lowest p-value. We then defined differentially marked promoters as those that had FDR < 0.05 and a positive logFC. Using this definition, we identified 1214 genes with promoters that were significantly enriched for H3K27 tri-methylation in control epithelium compared to the Ezh2-deficient samples (Fig. 3C). We obtained similar results when aggregating the bins over the gene bodies (defined as a region including 3 kb upstream of TSS and the rest of the gene).

In order to assess how the loss of Ezh2 affected the expression of genes marked with H3K27me3, we performed a gene set test using the ROAST method [20] from the limma package with genes differentially marked by H3K27me3 as the target gene set. Consistent with the repressive role of H3K27 tri-methylation, these genes were significantly over-represented among the genes up-regulated in the Ezh2-deficient epithelium (p value — 0.003).

2.8. ChIP-seq visualisation

Reads from a genomic locus of interest were extracted from individual BAM files and read coverage at each position was calculated using the GenomicAlignments package [21]. To account for differences in library sizes, we divided the coverage in each library by the corresponding effective library sizes converting the coverage into counts per 10 million reads. Mean group coverage was then calculated by averaging coverage from the samples of the same genotype. Normalised read coverage for control and Ezh2-deficient samples was visualised using the Gviz package (version 1.12.1) [22]. Given the broad nature of peaks from the H3K27me3 mark, read coverage was plotted with a 1 kb smoothing window. Fig. 3D displays mean read coverage for the known Ezh2 target gene, *Cdx2*, as well as a *Col1a2* gene constitutively expressed in lung epithelium. For *Cdx2* the strong H3K27 tri-methylation signal seen in the control epithelium is largely lost in the Ezh2-deficient sample. At the same time, there was no change in signal over the promoter of *Col1a2*, consistent with its active expression.

3. Discussion

In this report we provide a detailed description of the bioinformatics analysis that we carried out on the original transcriptome and H3K27 tri-methylation data from Ezh2-deficient lung epithelium [3]. One of the notable features of that study is separation of the embryonic lung tissue into cell populations enriched for either epithelial or mesenchymal cells, based on the expression of the epithelial surface marker EpCAM. The use of purified cell populations for RNA-seq profiling enabled us to detect gene expression changes specific to the tissue targeted for Ezh2 deletion (epithelium). It also improved our ability to detect expression changes that would have been otherwise obscured by the expression patterns in the mesenchymal cell population. For example, one of the most interesting findings of the original report was increased expression of *Igf1* in the epithelium of Ezh2-deficient lung, a gene highly expressed in the wild-type lung mesenchyme. This expression pattern is a likely reason why the increase in *Igf1* expression was overlooked in a similar study using RNA extracted from the whole lung of Ezh2-deficient embryos [23].

We extended the use of purified cell populations to our ChIP-seq analysis, providing a map of H3K27 tri-methylation specific to lung epithelium. We employed a relatively little-used approach in our ChIP-seq experimental design, by comparing immunoprecipitated DNA from wild type cells and cells derived from a sample where H3K27me3 marks have been depleted through genetic targeting of Ezh2, the histone methyl-transferase largely responsible for the deposition of tri-methylation marks on H3K27. We feel that this is an improvement on a more conventional design, where immunoprecipitated sample is compared to either a sample immunoprecipitated with a non-specific antibody (IgG control) or to a whole cell extract (input) sample. The "input" approach does not capture the potentially crucial biases introduced by the immunoprecipitation step. While these technical biases are addressed by the "IgG control" approach, it presents a substantial challenge for library preparation and sequencing due to the limited amount of material that can be immunoprecipitated. The additional benefit of the differential binding approach we employed is that it allows for the detection of more relevant genomic regions, where the level of epigenetic mark changes upon the loss of an epigenetic mediator. In this case, the use of immunoprecipitated sample from Ezh2-deficient lung epithelium enabled us to distinguish the genes that lost H3K27 tri-methylation from the loci where the mark remained unaffected by the loss of Ezh2, leading to a more biologically relevant interpretation of the gene expression data.

We also used a novel, window-based method to analyse the ChIP-seq data, which enabled us to detect differentially bound regions without relying on peak calling or limiting the analysis to predefined genomic regions, such as promoters or gene bodies. For integration with gene expression analysis, we summarised the windows (or bins) at the promoter level. This approach is superior to simply counting the reads across the entire promoter region, which may fail to detect differential marking if it only occurs in a fraction of the promoter. This is especially important for H3K27 tri-methylation, which is less confined to the promoter regions compared, for example, to H3K4me3 or H3K9ac marks.

While repeating some of our analyses during preparation of this manuscript, we noted some differences with the results in the original publication due to minor changes in the latest versions of the underlying software. In particular, in this report we detected two extra differentially expressed genes in the RNA-seq analysis of Ezh2-deficient epithelium. To ensure reproducibility of this analysis, we provide the script used to perform all steps described in this report. The script is provided within an R project with version control enabled by packrat [24], a dependency management system for R to ensure that appropriate versions of software packages are available to anybody wishing to reproduce our analysis. The scripts and versions of R packages used in the analysis are available from http://bioinf.wehi.edu.au/folders/ezh2lung/. The most up-to-date versions of R (3.2.1) [25] and Bioconductor (version 3.1) [26] available at the time of writing were used for the analysis. The only exception to that rule is read alignment and summarisation, which were carried out using an earlier version of Rsubread (1.13.25).

Acknowledgments

The authors thank Stephen Wilcox and Prof Gordon Smyth for genomics and bioinformatics expertise. MLAL is supported by Australian Research Council Queen Elizabeth II Fellowships. ATLL is supported by an Elizabeth and Vernon Puzey Scholarship from the University of Melbourne. LAG is supported by Strathmore Community College Scholarship. This work was supported in part by an Australian National Health and Medical Research Council grant to MLAL and MR (APP1079756). This work was made possible through Victorian State Government Operational Infrastructure Support, Australian Government NHMRC IRIISS, and the Ian Potter Foundation.

References

[1] B.D. Harfe, P.J. Scherz, S. Nissim, H. Tian, A.P. McMahon, et al., Evidence for an expansion-based temporal Shh gradient in specifying vertebrate digit identities. Cell 118 (2004) 517–528.

[2] S. Ih, A. Basavaraj, A.N. Krutchinsky, O. Hobert, A. Ullrich, et al., Ezh2 controls B cell development through histone H3 methylation and Igh rearrangement. Nat. Immunol. 4 (2003) 124–131.

[3] L.A. Galvis, A.Z. Holik, K.M. Short, J. Pasquet, A.T.L. Lun, et al., Repression of Igf1 expression by Ezh2 prevents basal cell differentiation in the developing lung. Development 142 (2015) 1458–1469.

[4] S. Andrews, FastQC: a quality control tool for high throughput sequence dataURL 2014. http://www.bioinformatics.babraham.ac.uk/projects/fastqc (Last accessed: 08/08/2014).

[5] Y. Liao, G.K. Smyth, W. Shi, The subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Res. 41 (2013) e108-e108.

[6] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, et al., The sequence alignment/map format and SAMtools. Bioinformatics 25 (2009) 2078–2079.

[7] Y. Liao, G.K. Smyth, W. Shi, featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30 (2014) 923–930.

[8] M.D. Robinson, A. Oshlack, A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 11 (2010) R25.

[9] M.D. Robinson, D.J. McCarthy, G.K. Smyth, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26 (2010) 139–140.

[10] C.W. Law, Y. Chen, W. Shi, G.K. Smyth, voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biology 15 (2014) R29.

[11] G.K. Smyth, Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Stat. Appl. Genet. Mol. Biol. 3 (2004) 1–25.

[12] D.J. McCarthy, G.K. Smyth, Testing significance relative to a fold-change threshold is a TREAT. Bioinformatics 25 (2009) 765–771.

[13] M.E. Ritchie, B. Phipson, D. Wu, Y. Hu, C.W. Law, et al., limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research (2015) gkv007.

[14] 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 (1995) 289–300.

[15] A.T.L. Lun, G.K. Smyth, De novo detection of differentially bound regions for ChIP-seq data using peaks and windows: controlling error rates correctly. Nucleic Acids Research (2014) gku351.

[16] A.F.A. Smit, R. Hubley, P. Green, RepeatMasker Open-4.0. 2015. URL http://www.repeatmasker.org (Last accessed: 15/06/2014).

[17] S.P. Lund, D. Nettleton, D.J. McCarthy, G.K. Smyth, Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunken dispersion estimates. Stat. Appl. Genet. Mol. Biol. 11 (2012).

[18] D.J. McCarthy, Y. Chen, G.K. Smyth, Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. 40 (2012) 4288–4297.

[19] R.J. Simes, An improved Bonferroni procedure for multiple tests of significance. Biometrika 73 (1986) 751–754.

[20] D. Wu, E. Lim, F. Vaillant, M.-L. Asselin-Labat, J.E. Visvader, et al., ROAST: rotation gene set tests for complex microarray experiments. Bioinformatics 26 (2010) 2176–2182.

[21] M. Lawrence, W. Huber, H. Pagès, P. Aboyoun, M. Carlson, et al., Software for computing and annotating genomic ranges. PLoS Comput. Biol. 9 (2013) e1003118.

[22] F. Hahne, S. Durinck, R. Ivanek, A. Mueller, S. Lianoglou, G. Tan, L. Parsons, Gviz: plotting data and annotation information along genomic coordinates. R package version 1.12.1URL 2015. http://bioconductor.org/packages/release/bioc/html/Gviz.html (Last accessed 15/05/2015).

[23] M.E. Snitow, S. Li, M.P. Morley, K. Rathi, M.M. Lu, et al., Ezh2 represses the basal cell lineage during lung endoderm. Development 142 (2015) 108–117.

[24] K. Ushey, J. McPherson, J. Cheng, J.J. Allaire, packrat: a Dependency Management System for Projects and their R Package Dependencies. R package version 0.4.4URL 2015. https://rstudio.github.io/packrat (Last accessed 29/06/2015).

[25] R Core Team, R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2015. (URL: http://www.Rproject.org Last accessed 15/06/2015).

[26] R.C. Gentleman, V.J. Carey, D.M. Bates, B. Bolstad, M. Dettling, et al., Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 5 (2004) R80.