# COMPARING MULTIPLE PROTEIN BINDING PROFILES IN CHIP-SEQ EXPERIMENTS

Hatice Gulcin Ozer[1,2], Jiejun Wu[3], Yi-Wen Huang[3], Jeffrey Parvin[1,2], Tim Huang[3], Kun Huang[1,2]

[1] Department of Biomedical Informatics, The Ohio State University
[2] The Ohio State University CCC Biomedical Informatics Shared Resource
[3] The Ohio State University Comprehensive Cancer Center
Columbus, OH 43210
{ozer, khuang}@bmi.osu.edu

New high-throughput sequencing technologies can generate millions of short sequences in a single experiment. As the size of the data increases, comparison of multiple experiments on different cell lines under different experimental conditions becomes a big challenge. In this paper, we investigate ways to compare multiple ChIP-seq experiments. We specifically studied epigenetic regulation of breast cancer and the effect of estrogen using 50 ChIP-seq data from Illumina Genome Analizer II. First, we evaluate the correlation among different experiments focusing on total number of reads in transcribed regions of the genome. Then, we adopt the method that is used to identify most stable genes in RT-PCR experiments to understand background signal across all experiments and to identify most variably transcribed regions of the genome. Gene ontology and function enrichment analysis on the 100 most variable genes demonstrate the biological relevance of the results. In this study, we present a method can effectively select diferentially transcribed regions based on protein binding profiles over multiple experiments using real data points without any normalization among the samples.

## 1. INTRODUCTION

During the past three years, with the rapid advancement in the next generation sequencing (NGS) technology, and related techniques such as chromatin immunoprecipitation (ChIP), researchers can investigate the protein-DNA binding relationship at high resolution using the ChIP-seq method. In a recent study, we have generated a large collection of ChIP-seq data for different cell lines and samples as well as for different proteins. The goal of this study is to understand the epigenetic regulation of breast cancer and the effect of estrogen. So far we have generated 46 lanes of ChIP-seq data from the Illumina Genome Analyzer II (GAII) for five different proteins in eight different samples (including the same samples under different conditions) with a total of more than 11 giga bases. The proteins include RNA polymerase II (Pol II), estrogen receptor α (ERα), and epigenetic markers including H3K4-dimethylation (H3K4me2), H3K9-dimethylation (H3K9me2), and H3K27-trimethylation (H3K27me3).

The accumulation of this large amount of data brings up an important question. How can we explore the difference of protein binding profiles among different cell types and multiple proteins. In another words, can we identify the genes with significantly different binding profiles for multiple proteins and cell types to differentiate different samples? To answer this questions, usually we need to develop effective comparison method such as data normalization between the data. However, since the data were generated from samples with distinctively different phenotypes and different proteins, normalization is not feasible or even meaningful. Instead, we need to resort to other approaches. In this paper, we adopted methods previously used for gene expression analysis in microarray and quantitative real-time PCR (qPCR) experiment. We first establish the similarity among different ChIP-seq experiments using well annotated genes. Then we further apply a feature selection method to pick genes that show differentiated protein binding profiles across multiple samples.

An advantage of this method is that it circumvented the process of normalization among the samples and is robust to experimental variations. Our results paved the way to carry out a whole genome screening in the future work.

## 2. METHODS

### 2.1. Dataset

ChIP-sequencing experiments are conducted on ten different samples with five different pull down proteins. Table 1 summarizes the types of samples and cell lines and pull down proteins. The samples include mammosphere from human breast cancer tissue, MCF10A (a non-tumerigenic human mammary epithelial cell line), MCF7 (a human breast cancer epithelial cell line), OHT (a taxmoifen resistant cell line

**Table 1**. Total of 50 ChIP-sequencing experiments are done on 10 different samples with 5 pull down proteins.

| Samples | Protein | | | | | Total Number of Lanes from Illumina GAII |
|---|---|---|---|---|---|---|
| | Pol II | ERα | H3K4me2 | H3K9me2 | H3K27me3 | |
| **mammosphere** | 2 | 1 | | | | 3 |
| **mammosphere+E2** | 2 | 1 | | | | 3 |
| **MCF10A w/ AKT knockout** | | | | 4 | 2 | 6 |
| **MCF10A** | | | | 4 | 2 | 6 |
| **MCF7** | 2 | 2 | 2 | 4 | | 10 |
| **MCF7+E2** | 2 | 2 | | | | 4 |
| **OHT** | 2 | 2 | 2 | | 4 | 10 |
| **OHT+E2** | 2 | 2 | | | | 4 |
| **PLC5** | | | | | 2 | 2 |
| **Huh7** | | | | | 2 | 2 |
| **Total** | **12** | **10** | **4** | **12** | **12** | **50** |

derived from MCF7). In addition, we also have data for two hepatoma (liver cancer) cell lines PLC5 and Huh7.

The treatments on the samples include AKT gene knockout (in MCF10A), application of E2 (also known as17-Estradial, to mammosphere and MCF7 and OHT cell lines). For each sample, ChIP-seq experiments focusing on different proteins have been carried out. The proteins include RNA polymerase II (Pol II), estrogen receptor α (ERα), and epigenetic markers including H3K4-dimethylation (H3K4me2), H3K9-dimethylation (H3K9me2), and H3K27-trimethylation (H3K27me3).

We used Illumina's Eland mapping algorithm to map sequence reads to Human reference genome hg18. We extracted transcribed region coordinates from RefGene database (1). For every experiment total number of sequence reads that are mapped to a gene region (as defined in RefGene) is counted. Here we will refer to total number sequence reads in a gene region as *gene count*.

## 2.2. Comparison of multiple ChIP-seq experiments

When the samples are clustered based on these gene counts, we observed that pull down protein and the sample type drives the clustering. Fig. 1 shows the heatmap of sample clustering based on all gene counts. This analysis demonstrate that gene count profiles of the experiments with same pull down protein on different samples or same samples with different treatment options are very similar. On the other hand, gene count profiles of experiments with certain pull down proteins are very different. For example, correlation between experiments with H3K9 Me2 and H3K27 Me3, H3K9 Me2 and ER are very strong, while correlation between experiments with Pol II and H3K27 Me3, and Pol II and H3K9 Me2 are very poor.

## 2.3. Select gene features based on protein binding profiles over multiple samples

Since these experiments are done on different cell lines and with different pull down proteins, background signal is not uniform across these experiments. We need a method to describe background variation across all experiments.

Vandesompele et al (2002) introduced a gene stability measure to indentify most stably expressed control genes in RT-PCR experiments. They defined gene stability measure *M* as the average pairwise variation between any particular gene and all other genes. Variation between two genes is calculated as standard deviation of log-transformed gene expression ratios. A small M value means expression of this gene is quite stable across all samples, while a large *M* value means this gene varies a lot across all samples (2).

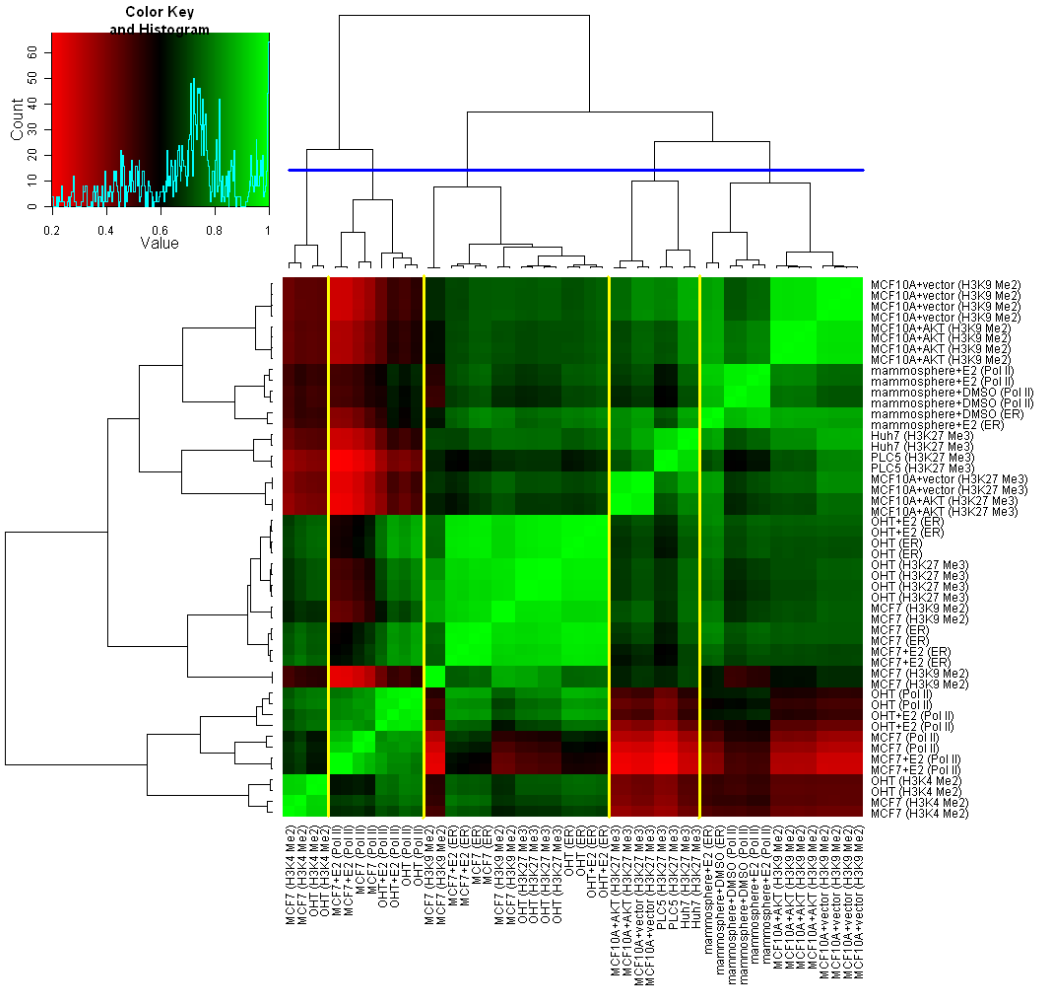


**Fig. 1.** Heatmap of sample clusering based on correlation of gene counts. Pearson correlation coefficient is calculated for every sample pair using read counts for 14,461 genes.

We adopted this method to determine least stable transcribed regions across all experiments as opposed to the original intent of the method to identify stable controls. When the transcribed regions were ranked by their *M* values, the small *M* values indicate the transcribed regions that do not vary in the samples or in response to experimental treatment in our 50 ChIP-sequencing experiments. This helped us to understand background variation and determine most variable regions across all experiments.

Out of *n* genes in total, for every combination of two genes *j* and *k* and in experiment *i*, log2-transformed ratios of gene counts $g_{ij}$ and $g_{ik}$ are calculated. Array $G_{jk}$ of *m* elements consists of these ratios across *m* experiments (Eq.1). Pairwise variation is calculated as the standard deviation of the $G_{jk}$ elements (Eq. 2). The gene-stability measure $M_j$ for a gene *j* is defined as the mean of all pairwise variations, $S_{jk}$ (Eq.3).

$$G_{jk} = \left\{ log_2\left(\frac{g_{1j}}{g_{1k}}\right), log_2\left(\frac{g_{2j}}{g_{2k}}\right), \ldots, log_2\left(\frac{g_{ij}}{g_{ik}}\right) \right\} \quad (1)$$

$$= \left\{ log_2\left(\frac{g_{ij}}{g_{ik}}\right) \right\}_{i=1 \to m}$$

$$S_{jk} = std\,(G_{jk}) \quad (2)$$

$$M_j = \frac{\sum_{k=1}^{n} S_{jk}}{n-1} \quad (3)$$

Out of 20,998 genes (transcribed regions) listed in RefGene database, we selected 14,461 genes that have 50 or more sequence reads on them at least in one of the experiments. Then we calculated *M* values for these 14,461 genes. We implemented *M* value calculations in R statistical data analysis language. Calculations are done on the computing cluster that consists of 72 computing nodes with dual quad core AMD Opteron 2378 processors. We submitted 145 jobs simultaneously (~100 genes per job) and each job is completed in 5 hours 48 minutes. Based on this timing, *M* value calculations for 14,461 genes would take about 35 days on a single processor.

## 3. RESULTS

### 3.1. Calculation of M values

First, we investigate the range of *M* value and its indications. Fig. 2 shows the histogram of *M* values for 14,461 genes. *M* values range between 1 and 2.8, and it is less than 2 for 96% of the genes. This indicates that there is a background pattern that most of the genes follow. Small *M* value means counts for this gene is quite stable across all experiments. On the other hand, large *M* value means counts for this gene is quite different than the background pattern.

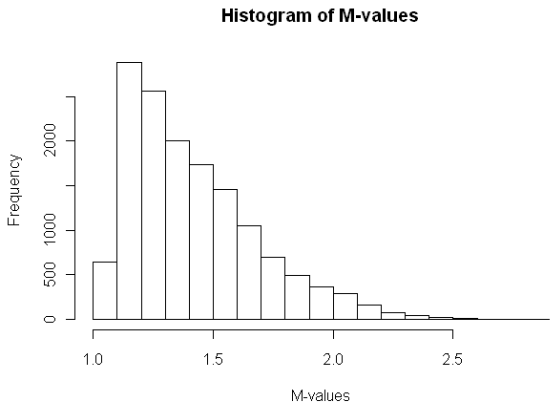


**Fig. 2.** Histogram of the *M* values of 14,461 transcribed regions of RefGene database across 50 ChIP-sequencing experiments.

### 3.2. Gene ontology and function enrichment analysis

In this study we selected top 100 genes as the most variable genes across all experiments (See appendix for the gene list). Counts for these genes are used to cluster genes and experiments. The heatmap in Fig. 3 shows clustering of these top 100 most variable genes across 50 experiments.

We performed functional analysis on the most variable 100 genes across 50 ChIP-seq experiments using Ingenuity Pathway Analysis (IPA) (3). Top functional groups are identified as cellular development, protein synthesis, gene expression, tissue development, RNA damage and repair, cancer, cellular growth and proliferation, cell cycle and breast cancer. Table 2 lists the IPA function and disease annotations, associated p-values (right-tailed Fisher's exact test) and total number of genes in that category.

## 4. CONCLUSION AND DISCUSSION

There are number of normalization methods introduced to compare pairs of sequencing experiments or series of same type of sequencing experiments (4-9). However, it is not feasible to apply these methods to compare tens, hundreds of sequencing experiments that conducted on different samples with different pull down proteins since the assumption on the same background level and noise model does not hold for different proteins. Our method provides a means to integrate ChIP-seq data from multiple experiments with different proteins, which can be of great importance in characterizing the regulome of the cells.

We observed that sample type and pull down protein are the major factors that determine the whole genome profile of the ChIP-seq datasets. As shown in Fig. 1, experiments are tightly clustered by protein and/or sample type. For example, two hepatoma cell lines PLC5 and Huh7 are clustered with a non-tumerigenic human mammary epithelial cell line MCF10A with H3K27me3 epigenetic marker. In this case, H3K27me3 determines the background signal. On the other hand, human epithelial cell lines of MCF7 (breast cancer) and MCF10A (non-tumerienic mammary) do not cluster by H3K9me3 but tightly

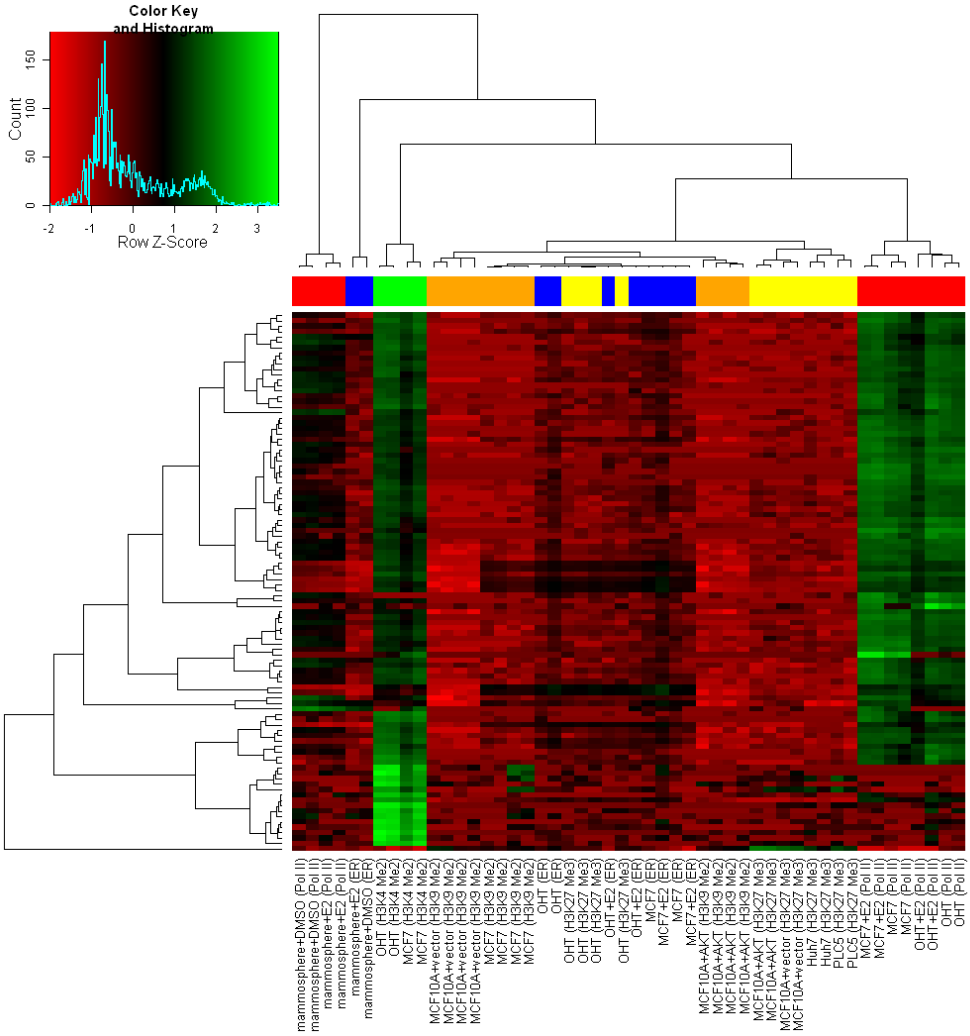


**Fig. 3.** Clustering of top 100 most variable genes and 50 ChIP-sequencing experiments. The distance function for hierarchical clustering is 1-correlation. Colors along the columns indicate 5 pull down proteins.

cluster by sample cell line. These observations confirms that whole genome profile of different samples and proteins are not the same. Therefore, we need a method to compare multiple ChIP-seq datasets without changing the observed data.

In this study, we introduced a new method to compare multiple ChIP-seq experiments without normalization. For every gene we calculated average pairwise variation between itself and all other genes as the stability measure, *M* value. By ranking all of the genes by *M* values, we identified the 100 most variable genes across 50 ChIP-seq experiments.

The functional and disease enrichment analyses on these 100 genes demonstrate the biological relevance of the results. First, most of the samples are from cancer cell lines or samples and our gene list confirms this fact with several cancer related functional groups being highly enriched such as carcinoma (35 genes) and breast cancer (23 genes). Second, most of the cell types in this study are epithelial cells and the most enriched functional term from IPA analysis is related to the developmental process of the epithelial cells. Thirdly, the samples contain two different tissues – breast tissue (MCF7, MCF10A, mammosphere, OHT) and liver tissue (PLC5 and Huh7). Out of the 100 genes, 7 of

them are identified as liver morphology genes, which implies that these genes can be applied to differentiate the two tissues. In addition to these functional groups that confirms our expectation on the cancer related samples, several other terms may also lead to more biological insight. For instance, two functional groups are related to fibroblasts (ie. interphase of fibroblasts, cell death of fibroblast cell lines). Given the complicated roles of fibroblast in cancer development and its existence in mammosphere samples, we can explore more about the roles of the fibroblast cell division and death genes in the different cancer stage. In addition, the 18 genes related to the formation of pulmonary artery may also be related to the angiogenesis process in regular tumor development and tumor microenvironment component interactions.

This approach can also be used to evaluate the quality of series of experiments. For example, MCF7 and OHT samples for ERα protein do not cluster with the mammosphere samples for the same protein. For these particular samples we know that MCF7 and OHT samples for ERα protein had some problems in sample preparation.

As a conclusion, we introduced a method to effectively compare multiple ChIP-seq experiments without changing the real observations. The proposed method does not require normality or homogeneity of variance for the data points (2). We identified least stable gene regions across large set of experiments using the stability measure *M*. The presented results confirm the feasibility of the method and necessitates whole genome screening as the future work.

**Table 2.** IPA function and disease annotation terms associated with the top 20 with functional categories (replicated terms are removed, 17 terms are left).

| IPA Function/Disease Annotation | p-value | # Genes |
|---|---|---|
| developmental process of epithelial cells | 1.06E-09 | 26 |
| synthesis of protein | 8.49E-09 | 16 |
| transcription of DNA | 3.03E-06 | 26 |
| developmental process of tissue | 3.90E-06 | 19 |
| catabolism of mRNA | 3.92E-06 | 4 |
| Carcinoma | 6.57E-06 | 35 |
| proliferation of connective tissue cells | 8.33E-06 | 20 |
| interphase of fibroblasts | 9.28E-06 | 16 |
| breast cancer | 9.61E-06 | 23 |
| Dermatitis | 1.47E-05 | 9 |
| cell death of fibroblast cell lines | 1.55E-05 | 32 |
| formation of pulmonary artery | 2.46E-05 | 8 |
| cytolysis of trophoblast giant cells | 2.46E-05 | 16 |
| tumorigenesis of large-cell diffuse lymphoma | 2.46E-05 | 11 |
| formation of pulmonary artery | 2.46E-05 | 18 |
| quantity of endometrial cancer cell lines | 2.46E-05 | 6 |
| morphology of liver | 3.30E-05 | 7 |

## Appendix

List of 100 most variable genes:

MALAT1, CLDN4, KRT19, TXNIP, SNHG1, ACTG1, ZFP36L1, BRD2, RPS2, SLC38A2, PRMT6, DKFZP686I15217, FOXH1, RPL10, ATF4, HIST1H1C, RAB26, NPPC, EEF1A1, TWIST1, ZNF580, HEXIM1, GADD45B, TUBD1, HSPA1A, RPLP0, C11orf83, JMJD8, EIF4A2, JUN, MRPL41, JUNB, LTB, TNFRSF12A, PSMD6, AREG, FLJ11235, GAS5, HOXD8, PARD6B, RPRML, TLCD1, HES7, RPL12, RPL13, H2AFZ, MYL6, MRPS34, BRIP1, ZFP36, SNHG8, C17orf82, ZNF217, DKFZp779M0652, RNFT1, KRT18, SNHG5, HES1, NCRNA00173, RPL27A, ZFP36L2, IER5L, MCM7, NEAT1, HLA-H, PNRC2, HIST2H2AC, HAGHL, DDIT4, RPL41, TBX2, XBP1, CDKN2BAS, IER5, FOXA1, S100A11, CDKN2A, CDK5R2, PROKR2, HIST1H2AE, APRT, PPAN-P2RY11, SFRS2, C11orf48, TMEM107, IER2, HIST1H1B, VGF, HIST2H2BE, PHPT1, HIST1H3B, HIST1H2AM, RPS6KB1, MFSD3, NOG, MIDN, HIST1H3D, MYC, NEUROD2, C10orf114.

## References


1. RefGene Database: http://hgdownload.cse.ucsc.edu/goldenPath/hg18/database/refGene.txt.gz
2. Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A, Speleman F: Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol* 2002, **3**:research0034.1-0034.11.
3. Ingenuity Pathway Analysis, IPA: http://www.ingenuity.com
4. Taslim C, Wu J, Yan P, Singer G, Parvin J, Huang T, Lin S, Huang K: Comparative study on ChIP-seq data: normalization and binding pattern characterization. *Bioinformatics* 2009 25(18):2334-2340.
5. Balwierz PJ, Carninci P, Daub CO, Kawai J, Hayashizaki Y, Van Belle W, Beisel C, van Nimwegen E: Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepCAGE data. *Genome Biol* 2009, 10:R79.
6. Cloonan N, Forrest AR, Kolle G, Gardiner BB, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G, Robertson AJ, Perkins AC, Bruce SJ, Lee CC, Ranade SS, Peckham HE, Manning JM, McKernan KJ, Grimmond SM: Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat Methods* 2008, 5:613-619.
7. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 2008, 5:621-628.
8. Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D, Schmidt D, O'Keeffe S, Haas S, Vingron M, Lehrach H, Yaspo ML: A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* 2008, 321:956-960.
9. Robinson MD and Oshlack A: A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology* 2010, 11:R25.