dHealth 2019 – From eHealth to dHealth
D. Hayn et al. (Eds.)

doi:10.3233/978-1-61499-971-3-105

A Comprehensive FXR Signaling Atlas Derived from Pooled ChIP-seq Data

Emilian JUNGWIRTH[a,b,d,e,1], Katrin PANZITT[b], Hanns-Ulrich MARSCHALL[c], Martin WAGNER[b,d,e] and Gerhard G. THALLINGER[a,d,e]
[a]*Institute of Computational Biotechnology, Graz University of Technology, Austria*
[b]*Research Unit for Translational Nuclear ReceptorResearch, Division of Gastroenterology and Hepatology, Medical University Graz, Graz, Austria*
[c]*Department of Molecular and Clinical Medicine, University of Gothenburg and Sahlgrenska University Hospital, Gothenburg, Sweden*
[d]*OMICS Center Graz, Graz, Austria*
[e]*BioTechMed-Graz, Graz, Austria*

Abstract. Background: ChIP-seq is a method to identify genome-wide transcription factor (TF) binding sites. The TF FXR is a nuclear receptor that controls gene regulation of different metabolic pathways in the liver. **Objectives:** To re-analyze, standardize and combine all publicly available FXR ChIP-seq data sets to create a global FXR signaling atlas. **Methods:** All data sets were (re-)analyzed in a standardized manner and compared on every relevant level from raw reads to affected functional pathways. **Results:** Public FXR data sets were available for mouse, rat and primary human hepatocytes in different treatment conditions. Standardized re-analysis shows that the data sets are surprisingly heterogeneous concerning baseline quality criteria. Combining different data sets increased the depth of analysis and allowed to recover more peaks and functional pathways. **Conclusion:** Published single FXR ChIP-seq data sets do not cover the full spectrum of FXR signaling. Combining different data sets and creating a "FXR super-signaling atlas" enhances understanding of FXR signaling capacities.

Keywords. ChIP-seq, FXR, ENCODE

1. Introduction

Transcription factors (TF) bind to distinct recognition sites on the DNA and thereby regulate gene transcription. Chromatin immunoprecipitation sequencing (ChIP-seq) is a method to identify genome-wide binding sites of a specific TF and to gain information about transcriptional regulation, affected genes and pathways. Nuclear receptors (NRs) are a class of TFs, which are directly activated/inactivated by agonistic/antagonistic ligands. The NR farnesoid X receptor (FXR) is activated by bile acids, thereby controlling gene regulation of different metabolic pathways mainly in the liver (*e.g.* bile acid-, lipid- and glucose metabolism). FXR recently attracted attention as a novel drug target for various metabolic liver diseases. Therefore, understanding precise genomic FXR binding and transactivation of genes is important to fully reconstruct FXR signaling, particularly when used as therapeutic drug.

[1] Corresponding Author: Emilian Jungwirth, Medical University of Graz, A-8010 Graz, Stiftingtalstrasse 24, E-Mail: emilian.jungwirth@medunigraz.at

Several FXR ChIP-seq data sets for different species, conditions and cell lines have been reported, none so far for human liver tissue. Our aim was to re-analyze these publicly available data sets with a standardized method and combine these data sets for further extended downstream analysis of FXR signaling properties. In addition, we compared the available public data sets to our own human biopsy material.

2. Methods

We searched public sources for available FXR-ChIP-seq data sets to determine a common set of generally applicable quality criteria based on the ones proposed in the ENCODE- and other authoritative ChIP-seq guidelines [1, 2]. Furthermore, we investigated different parameter settings and control sample variants. A combined mouse FXR ChIP-seq data set was generated by pooling mapped reads of the available four standardized mouse data sets to gain a higher sequencing depth. Enriched regions a.k.a. peaks represent putative FXR binding sites. A *de novo* motif analysis and motif scan was performed on all called peaks. The potentially regulated genes were determined using proximity to peaks. Those genes were used to identify enriched pathways.

2.1. Data sets

In public repositories, FXR-ChIP-seq data sets were available for mouse, rat and a cell line of primary human hepatocytes. We also had access to our own FXR-ChIP-seq data set from human liver tissue (Table 1). Raw reads were available for all data sets except "Mouse-Guo" and "Mouse-Osborne". For the "Mouse-Osborne" data set only mapped read tracks were available. In case of the "Mouse-Guo" data sets only the called peak tracks were available.

Table 1. Available data sets for this study. Naming of the data sets is based on the species and the last author of the paper where the data was first published.

Paper	Samples	Name	Ref
Genome-wide tissue-specific farnesoid X receptor binding in mouse liver and intestine.	2	Mouse-Guo	[3]
Genome-wide interrogation of hepatic FXR reveals an asymmetric IR-1 motif and synergy with LRH-1.	1	Mouse-Osborne	[4]
Metformin interferes with bile acid homeostasis through AMPK-FXR crosstalk.	4	Mouse-Lefebvre	[5]
Gene expression profiling in human precision cut liver slices in response to the FXR agonist obeticholic acid.	4	Mouse-Kersten	[6]
Genomic analysis of hepatic farnesoid X receptor binding sites reveals altered binding in obesity and direct gene repression by farnesoid X receptor in mice.	4	Mouse-Kemper	[7]
Toxicogenomic module associations with pathogenesis: a network-based approach to understanding drug toxicity.	6	Rat-Stevens	[8]
Genome-wide binding and transcriptome analysis of human farnesoid X receptor in primary human hepatocytes.	2	PHH-Guo	[9]
Unpublished observation: FXR ChIP-seq in normal vs cholestatic patients	2	Human-Wagner	-

2.2. ChIP-seq analysis

We created our own ChIP-seq analysis pipeline (Fig 1). The quality of the data samples is assessed at relevant steps of the analysis. Most of the data processing was performed using a locally available Galaxy [10] instance. The analysis comprises three major steps:

Raw read handling: Most of the data sets were single-end (SE) Illumina reads. Trimmomatic (version 0.36.5) [11] was used to trim and filter overrepresented sequences such as Illumina adapter. Additional parameters to the ILLUMACLIP were a SLIDINGWINDOW of 4 bases with an average quality of 28 and a minimum length of 80% of the raw read length to ensure a high read quality. FastQC [12] was used to confirm the quality.

Mapping and peaks calling: Filtered reads were mapped to the human genome version hg19, mouse genome version mm10 and rat genome version rn6 using Bowtie 2 (version 2.3.4.2) [13, 14] with default parameters.

To determine putative FXR binding sites model-based analysis of ChIP-seq version 2 (MACS2 version 2.1.1) [15, 16] was used. Various parameter combinations were used to evaluate their effects on the outcome and determine the most reliable parameter combination. The parameters were: q-value of 0.01 or 0.05, using input, IgG or no control sample, having a fixed or estimated fragment length and the two different standard effective genome sizes for human (2.45 and 2.7Gbp).

Downstream analyses: For the top 500 scoring peaks a *de novo* motif analysis was performed using Multiple Em for Motif Elicitation MEME SUITE (version 4.12.0.0) [17]. The sequences flanking the peak summit by 100bp on either side were examined. Apart from the number of motifs which was set to 10 the default parameters were used. Additionally, a motif scan for the canonical IR1 FXR motif (AGGTCAxTGACCT) [18] was performed using the tool FIMO from MEME SUITE. The scan was performed for the HOMER FXR motif across the narrow peaks and wider peak regions. The wider peak region was defined as 1000bp up- and downstream from the peaks summit.

Peaks were annotated to UCSC knownGenes using the R package ChIP-Seeker [19]. Each gene was defined as potentially regulated by FXR if a peak summit is located in the promotor (defined as +/-1kbp around TSS), intron or exon region of that gene. Genes were subjected to a REACTOME [20] pathways analysis; a q-value of less than 0.05 was considered statistically significant.

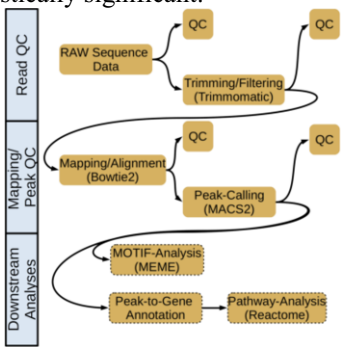

Figure 1. ChIP-seq analysis pipeline: The three major steps of a ChIP-seq analysis are (i) Read quality control (QC), (ii) Mapping and peak calling, and (iii) Downstream-analyses such as a motif- and a pathway-analysis.

2.3. Pooling the single data sets

A combined mouse data set "Mouse-pooled" was generated by pooling the filtered and mapped reads of 13 individual mouse samples from 4 different mouse data sets to gain higher sequencing depth. By pooling the samples on the read level, a summation of the individual FXR-signals is achieved. This summation of the FXR-signals allows the detections of weaker FXR binding sites, which could not be detected in single data sets. Because all data sets are from different laboratories only limited summation of noise is expected to occur. This analytic procedure combined with the strict filtering of the raw reads is expected to lead to a high quality virtually deep sequenced FXR ChIP-seq data set.

Subsamples were created to further investigate the saturation of FXR-related peaks/genes. The subsamples were created by randomly selecting reads from the entire combine data set. The subsamples size reached from 1/20 to 2/3 of the entire pooled reads. For each subsample size five distinct subsamples were created.

2.4. Comparison

The comparison between the data sets on a read and peak level was based on the quality metrics proposed in ENCODE- and other authoritative ChIP-seq guidelines [1, 2] (Table 2).

Table 2. Metrics used to assess the quality of the ChIP-seq samples. NSC/RSC were calculated using the phantompeakqualtools package version 2 [21, 22].

Quality metric	Abbriviation
Ratio of uniquely mapped reads to total number of reads	UMR/TNR
Ratio of uniquely mapped reads to total number of mapped reads	UMR/TMR
Non-Redundant Fraction	NRF
PCR Bottleneck Coefficient 1	PBC1
PCR Bottleneck Coefficient 2	PBC2
Normalized Strand Cross-correlation coefficient	NSC
Relative Strand Cross-correlation coefficient	RSC
Fraction of reads, which are in peak regions	FRiP
Percentage of peaks with foldchange greater than 5	%fc>5
Percentage of peaks, which are in Dnase I HS sites	% Dnase I HS

The similarity between the various peak calling results and the corresponding genes was determined using the Jaccard distance [23]. The pairwise Jaccard distances were visualized with a heatmap. It was necessary to map the genes to their orthologues of the other species to correctly estimate the similarity between different species. Mouse and rat genes were mapped to their corresponding human genes.

A dotplot was used to illustrate enrichment of pathways across samples. Some samples did not show any enriched pathways under the defined settings. Additional pathway trees for each sample with enriched pathways were created, to investigate the branch and subtree differences between the samples.

3. Results

In public repositories, FXR-ChIP-seq data sets from three different species are available: five for mice, one for rat and one for human primary hepatocytes. Most data sets include baseline FXR binding and binding events under pharmacological treatment (i.e. FXR activation with different ligands) or diseased conditions (i.e. diet-induced non-alcoholic fatty liver disease, bile duct ligation induced cholestasis). No public data sets are available for human liver tissue (Table 1). Our analysis shows that these data sets are heterogeneous concerning baseline quality criteria (Table 3).

Table 3. Evaluation of ChIP-seq quality for the available data sets. The number of samples/analysis results which pass the quality metric in respect total number of samples/analysis results is presented. Peak calling was performed with multiple parameter combinations; thereby the number of peak calling results is a multiple of the number of samples.

Study / Quality metric	Threshold value	Mouse-Guo	Mouse-Osborne	Mouse-Lefebvre	Mouse-Kersten	Mouse-Kemper	Mouse-pooled	Rat-Stevens	PHH-Guo	Human-Wagner
UMR/TNR	50%	-	-	4/4	4/4	4/4	-	4/6	2/2	2/2
UMR/TMR	50%	-	-	4/4	4/4	4/4	-	6/6	2/2	2/2
NRF	50%	-	-	4/4	4/4	4/4	1/1	6/6	2/2	1/2
PBC1	50%	-	1/1	4/4	4/4	2/4	1/1	6/6	2/2	1/2
PBC2	1,00	-	1/1	4/4	4/4	4/4	1/1	6/6	2/2	2/2
NSC	1,05	-	-	0/4	0/4	4/4	-	6/6	2/2	2/2
RSC	0,8	-	-	0/4	0/4	0/4	-	6/6	2/2	2/2
FRiP	1%	-	-	16/16	15/32	27/32	4/4	24/24	8/16	22/24
%fc>5	50%	-	8/8	16/16	32/32	32/32	0/4	24/24	16/16	24/24
%Dnase I HS	80%	1/2	8/8	0/16	0/32	0/32	0/4	-	2/16	5/24

When analyzed with the various analysis parameters in a standardized manner, the number of called FXR peaks and associated genes ranges from 103 to 40,080 and 6 to 12,873 in the single data sets, respectively. For the combined data set, the number of called peaks reached from 24,747 to 59,319 and the number of associated genes from 10,038 to 13,826 for the different parameter combinations. The called peaks/genes of the combined data sets represent more than just the simple addition of binding sites/genes from the single data sets and can be explained by enhancement of weak signals after virtually increasing sequencing depth.

The comparison of the public data sets to our human data set revealed that the quality of the human data set (although derived from surgical tissue) is in many regards at least as good as published data sets. The human data set passed the RSC quality criteria, which is crucial for the correct estimation of the fragment length by MACS2. The human data set also included an input and IgG control sample, which was critical to analyze the impact of different control samples in ChIP-seq experiments.

The most prevalent motif identified by the *de novo* search within the top 500 peaks was the canonical FXR IR-1 motif (AGGTCAxTGACCT). It was present in 2 to 54% of narrow peaks and 20 to 64% in wider peak regions for the different data sets.

The similarity between the samples was determined using the Jaccard distance based on the identified genes. Samples of the same data set group together rather than samples from the same condition/treatment from different data sets (Fig 2). Based on the quality criteria and the pairwise Jaccard similarities the parameter combination of:

q-value 0.05, no control sample, fixed set fragment length (if the estimated fragment length was unrealistic) and - for the human samples - an effective genome size of 2.7Gbp was considered as the most reliable parameter settings. Only peak calling results from those parameters were used for all further analysis.

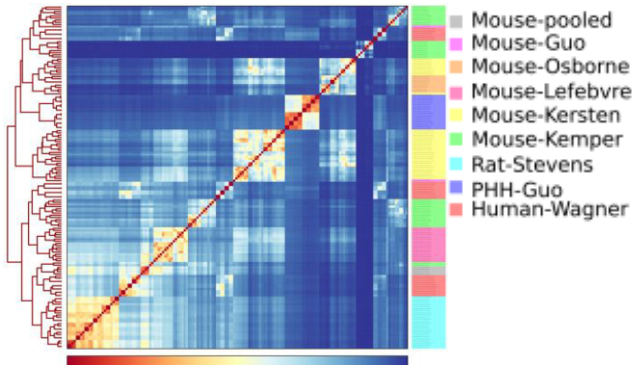

Figure 2. Heatmap based on the pairwise Jaccard distance. The samples are colored based on the data sets. The cluster tendency seems to be towards data sets rather than sample conditions.

Peaks were assigned to the closest annotated genes. Based on the assigned genes enriched REACTOME pathways were identified (Fig. 3). The combined analysis revealed additional significant pathways, which are not present in any of the single mouse data sets. Some of those additional pathways are also present in samples of other species. This demonstrates both a conservation of the FXR dependency of that pathway across multiple species and validity of the additional pathways identified by the combined data set.

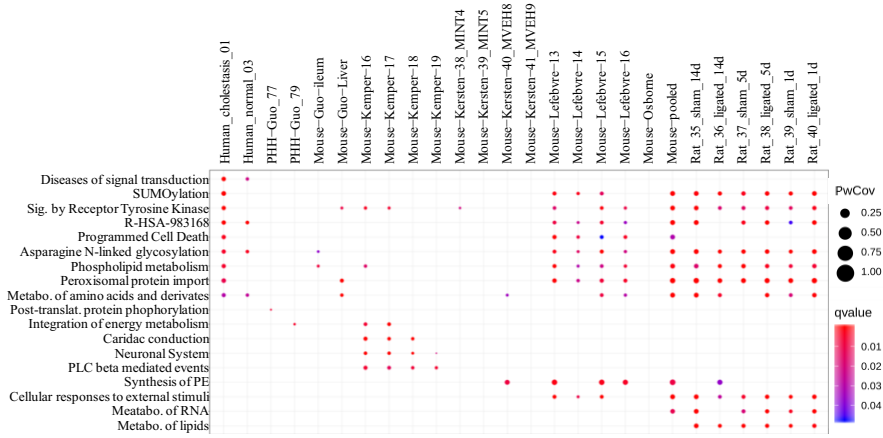

Figure 3. Top enriched RACTOME pathways represented in a dotplot. The dot color relates to the q-value and the size to the pathway coverage (number of pathway genes found in the pathway / total number of genes in the pathways). For some samples no enriched pathways were found under the defined settings.

3.1. Insights in FXR binding events revealed by the combined data set

The combined mouse data set shows many additional peaks, genes and pathways which were not present in any of the individual samples (e.g the 'Translocation of SLC2A4 (GLUT4) to the plasma membrane' pathway is one of 33 pathways which are only

present in the combined mouse data set). Similarly, some peaks, genes and pathways present in one or more individual mouse samples are not present in the combined data set (e.g. the 'Tspy-ps' gene is not present in the combined mouse data set although it is present in 8 of the individual mouse samples). This indicates that the signal for those peaks is not conserved across all samples. This could be explained either by a signal that is only present under very specific conditions, which were only met in a single sample, or by incorrectly called peaks due to noise. Peaks are more prevalent in the vicinity of TSSs, which is expected for a TF ChIP-seq experiment.

Interestingly, over 96% of the liver FXR ChIP-seq genes from the "Mouse-Guo" data set are present in the combined data set although the "Mouse-Guo" was not included in the pool because only the peak tracks were available. Furthermore 70% of the "Mouse-Guo" genes which are not present in any other single mouse sample are present in the pooled data set. This indicates that the pooling of FXR signal allowed the detection of weaker signals. Although the combined data sets revealed many new potential FXR related binding sites, saturation appears not to be reached. This is demonstrated by subsampling the combined data set (Fig. 4).

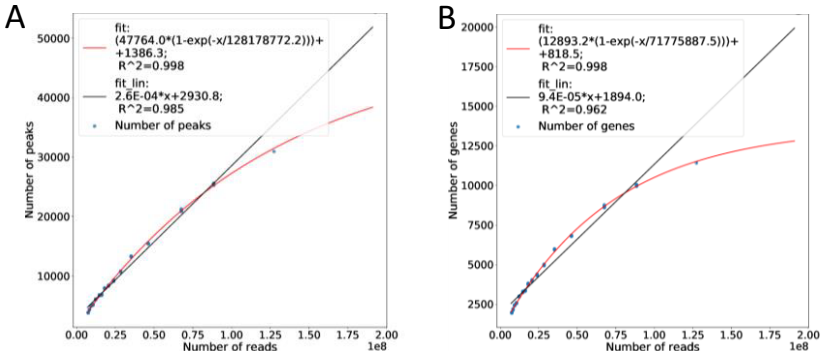

Figure 4. The number of reads with respect to the number of peaks (A) and number of genes (B) for the "Mouse-pooled" data set and its subsamples. The blue points represent the number of peaks/genes for either the entire "Mouse-pooled" data set or of its subsamples. A linear (black) and exponential (red) fitting curve was created for the data points; the exponential curve represents a much better fit.

4. Discussion

Several FXR ChIP-seq data sets are publicly available for various species and conditions. Standard ENCODE quality criteria are usually not reported for those data sets. We observe that the analysis results are sensitive to settings of certain analysis parameters such as the effective genome size and most prominently to the choice of control sample, which is generally underappreciated in most studies. A low-quality control sample can have a significant impact on the peak calling results even if the ChIP-seq sample is of good quality. Influences of control samples on the peak calling results were also reported in other studies [24]. Therefore, an analysis without a control sample should be considered. Interestingly, the human in vivo samples were more similar to rodent in vivo samples than to in vitro human primary hepatocytes.

Individual data sets often exhibit a too low sequencing depth to identify weak/rare binding sites, therefore we combined all available mouse reads to create a "FXR-super-signaling-atlas" for a profound downstream analysis of FXR signaling capacities. This data set allowed to detect more binding sites, genes and connected pathways. However, even the combined data set did not reach the theoretical determined saturation.

In conclusion, this meta-analysis of these different data sets with standardized methods should help to get a comprehensive and global overview of FXR binding events, FXR binding motifs, FXR-dependent gene regulation and affected pathways across various species. Combining standardized public data sets allows for more profound detection of binding events and signaling capacities.

References

[1] Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res 2012;22(9), 1813-31.
[2] Shin H, Liu T, Duan X, Zhang Y, Liu XS. Computational methodology for ChIP-seq analysis. Quantitative Biology 2013;1(1), 54-70.
[3] Thomas AM, Hart SN, Kong B, Fang J, Zhong X, Guo GL. Genome- wide tissue- specific farnesoid X receptor binding in mouse liver and intestine. Hepatology 2010;51(4), 1410-9.
[4] Chong HK, Infante AM, Seo Y, Jeon T, Zhang Y, Edwards PA, et al. Genome-wide interrogation of hepatic FXR reveals an asymmetric IR-1 motif and synergy with LRH-1. Nucleic Acids Res 2010;38(18), 6007-17.
[5] Lien F, Berthier A, Bouchaert E, Gheeraert C, Alexandre J, Porez G, et al. Metformin interferes with bile acid homeostasis through AMPK-FXR crosstalk. J Clin Invest 2014;124(3), 1037-51.
[6] Ijssennagger N, Janssen AW, Milona A, Pittol JMR, Hollman DA, Mokry M, et al. Gene expression profiling in human precision cut liver slices in response to the FXR agonist obeticholic acid. J Hepatol 2016;64(5), 1158-66.
[7] Lee J, Seok S, Yu P, Kim K, Smith Z, Rivas- Astroza M, et al. Genomic analysis of hepatic farnesoid X receptor binding sites reveals altered binding in obesity and direct gene repression by farnesoid X receptor in mice. Hepatology 2012;56(1), 108-17.
[8] Sutherland J, Webster Y, Willy J, Searfoss G, Goldstein K, Irizarry A, et al. Toxicogenomic module associations with pathogenesis: A network-based approach to understanding drug toxicity. The Pharmacogenomics Journal 2017;18(3), 377-90.
[9] Zhan L, Liu H, Fang Y, Kong B, He Y, Zhong X, et al. Genome-wide binding and transcriptome analysis of human farnesoid X receptor in primary human hepatocytes. PloS One 2014;9(9), e105930.
[10] Afgan E, Baker D, Van den Beek M, Blankenberg D, Bouvier D, Čech M, et al. The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Res 2016;44(W1), W3-W10.
[11] Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for illumina sequence data. Bioinformatics 2014;30(15), 2114-20.
[12] Andrews S. FastQC A quality control tool for high throughput sequence data. <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>. Accessed 2018 10/10.
[13] Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. Nature Methods 2012;9(4), 357.
[14] Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 2009;10(3), R25.
[15] Feng J, Liu T, Qin B, Zhang Y, Liu XS. Identifying ChIP-seq enrichment using MACS. Nature Protocols 2012;7(9), 1728.
[16] Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-seq (MACS). Genome Biol 2008;9(9), R137.
[17] Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in bipolymers. Proc Int Conf Intell Syst Mol Biol. 1994;2, 28-36.
[18] Laffitte BA, Kast HR, Nguyen CM, Zavacki AM, Moore DD, Edwards PA. Identification of the DNA binding specificity and potential target genes for the farnesoid X-activated receptor. J Biol Chem 2000;275(14), 10638-47.
[19] Yu G, Wang L, He Q. ChIPseeker: An R/bioconductor package for ChIP peak annotation, comparison and visualization. Bioinformatics 2015;31(14), 2382-3.
[20] Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, et al. The reactome pathway knowledgebase. Nucleic Acids Res 2017;46(D1), D649-55.
[21] Kundaje A, Jung LY, Kharchenko P, et al. Assessment of ChIP-seq data quality using cross-correlation analysis. <http://code.google.com/p/phantompeakqualtools>. Accessed 2018 08/23.
[22] Kharchenko PV, Tolstorukov MY, Park PJ. Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat Biotechnol 2008;26(12), 1351.
[23] Jaccard P. Lois de distribution florale dans la zone alpine. Bull Soc Vaudoise Sci Nat 1902;38, 69-130.
[24] Marinov GK, Kundaje A, Park PJ, Wold BJ. Large-scale quality analysis of published ChIP-seq data. G3 (Bethesda) 2014;4(2), 209-23.