Genomics Data 2 (2014) 302–304

Contents lists available at ScienceDirect

Genomics Data

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

Data in Brief

The role of Cockayne Syndrome Protein B in transcription regulation

Jieun Jeong

Epigenetics Program, Department of Cell and Developmental Biology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

ARTICLE INFO

Article history:
Received 9 August 2014
Received in revised form 24 August 2014
Accepted 26 August 2014
Available online 19 September 2014

Keywords:
Cockayne Syndrome
Regulation of gene expression
Enhancers
Promoters
Motif analysis

ABSTRACT

We investigated the question if CSB (Cockayne Syndrome complementation B) protein actively regulates gene transcription and how mutations in CSB gene affect that regulatory role.
Here we describe how we processed and interpreted ChIP-seq data (deposited in Gene Expression Omnibus with accession number GSE50171) obtained during an investigation of that question, and how this analysis assisted in the generation of hypothesis that were subsequently validated using other types of experiment.

Specifications	
Organism/cell line/tissue	*Homo sapiens*
Sex	*Male*
Sequencer or array type	*Illumina HiSeq 2000™ system*
Data format	*Raw files: ChIP-seq FASTQ files* *Processed files: peaks in bed format*
Experimental factors	*mutated CSB vs. wild type*
Experimental features	*ChIP-seq and bioinformatics analysis*
Consent	*n/a, in vitro experiment*
Sample source location	*Philadelphia, Pennsylvania, USA*

Direct link to deposited data *[provide URL below].*
http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=vfmfryagmygcony&acc=GSE50171.

Experimental design, materials and methods

Purpose of collecting and analyzing ChIP-seq data

Cockayne Syndrome complementation B (CSB, official symbol ERCC6) is an ATP-dependent chromatin remodeling protein, known for its role in lesion-driven DNA repair. Breaks in DNA chain can result from mutagenic factors like ultraviolet radiation or oxidizing compounds in nucleoplasm, but when they hinder gene expression by obstructing the RNA polymerase activity, DNA chain can be spliced back in a process that takes several stages. Some stages of that process require that DNA is detached from nucleosomes or attached back, and CSB participates in that activity [3].

That function of CSB explains why mutations of CSB gene may cause hypersensitivity to UV radiation and other mutagenic factors, but fails to fully explain the process of Cockayne Syndrome that entails premature aging and various developmental defects [4,5]. This motivated investigation for other functions of CSB. This note relates to the investigation of the role of CSB in regulating gene expression in cells not affected by DNA damage. The question was if CSB is binding to chromatin in a sequence-driven manner (as opposed to lesion-driven binding during DNA repair), and if yes, what sequence specific transcription factors control the binding of CSB and what are epigenetic and transcriptional consequences. The results of the complete investigation are in [1].

Cell culture

CS1AN-sv cells were maintained in DMEM-F12 with 10% FBS. The CS1AN primary cells have mutations in both CSB alleles, but only one of these alleles was retained after SV40 immortalization [7]; the resulting CS1AN cell line is, therefore, hemizygous for CSB. The retained allele contains an A to T transversion at position 1088, which introduces a premature stop codon at amino acid 337.

Stable cell lines expressing CSB were generated by infecting CS1AN-sv cells with CSB-expressing lentivirus (pLenti-PGK-Neo, Addgene) [6]. Stable cell lines expressing CSB or harboring the empty vector were selected within 600 μg/ml G418. CSBΔN1 was expressed from the pSVL vector [6]. CS1AN-sv cells stably expressing CSBΔN1 were generated by co-transfection with pLenti-PGK-neo.

http://dx.doi.org/10.1016/j.gdata.2014.08.009

ChIP-seq data preparation

To increase ChIP efficiency we removed soluble CSB before cross-linking DNA and proteins [6,9,11]. Cells were collected in Buffer B (150 mM NaCl, 0.5 mM MgCl2, 20 mM HEPES pH 7.8, 10% Glycerol, 0.5% Triton X-100) and soluble CSB was separated from chromatin by centrifugation at 15,000 rpm for 5 min at 4 °C. The resulting pellets were resuspended in Buffer B and fixed with 1% formaldehyde for 10 min at room temperature. Cross-linked cells were sonicated at 40% amplitude (30 s on, 90 s off, for 24 min total) using the Branson 101-135-126 Sonifier. Chromatin IP (ChIP) was performed using a monoclonal anti-CSB antibody (1B1) that recognizes the N-terminal 507 amino acids of CSB [9,10].

10 ng of ChIPed DNA was used to prepare libraries for deep sequencing using the multiplexed ChIP-seq sample preparation protocol described on the website of the Next-Generation Sequencing Core, Perelman School of Medicine, University of Pennsylvania (http://ngsc.med.upenn.edu/).

Initial data analysis

The Next-Generation Sequencing Core at the University of Pennsylvania performed DNA sequencing using Illumina hiSeq2000 sequencers for single-end sequencing with a read length of 50 bps. The resulting sequencing reads were mapped to the human genome (HG19 assembly) using Bowtie version 0.12.7. Peaks were identified using HOMER (Hypergeometric Optimization of Motif EnRichment) version 4.1 with a default option (FDR = 0.001 and Poisson p-value cutoff = 0.0001) on ChIPed samples against matching input DNA samples. Raw and processed files (GSE50171) have been deposited at the Gene Expression Omnibus (GEO) repository [2].

We removed from the analysis chromosome M because it has abnormally high coverage (as a cell contains many copies of that chromosome) and "non-standard chromosomes", parts of the chromosomes that were not assembled in HG19. Using HOMER package, we have identified 3607 peaks of CSBΔN1 and 17,779 peaks of CSB. Then we investigated further if peaks of CSB and CSBΔN1 have different properties.

Co-location of CSB peaks indicates associations with sequence specific transcription factors

We tested how frequent peaks of our two types are in the same locations. For each peak ***p*** of one type (CSB, CSBΔN1) we find the distance ***D(p)*** to the nearest peak of the other type. Most peaks have ***D(p)*** at least 10,000 bp, but the percentage of lower values is much higher than what we would get for a uniform random placement of one or both sets of peaks (for CSB it would be 2.9% and for CSBΔN1 it would be 14.2%. Moreover, in this model the percentages of distances under 1000 bp and under 100 bp would be respectively 10 and 100 times smaller) (Table 1).

The fact that distance under 100 bp form more than one third of distances under 10,000 bp indicates that in many cases CSB peaks are positioned by the same sequence-specific transcription factors as CSBΔN1.

Table 1
Peaks of the two types classified according to the distance from the nearest peak of the other kind, counts and percentages (in the additional columns).

$D(p) \leq$	100 bp	%	1 kbp	%	10 kbp	%
CSB	662	3.72	943	5.30	1892	10.64
CSBΔN1	662	18.35	927	25.70	1563	43.33

Next, we investigated what is the reason for the much lower number of binding sites identified for CSBΔN1 as opposed to CSB. More precisely, we asked which of the following is more likely:

A. N1 deletion removed the ability of CSB to bind to a locus.
B. After N1 deletion this ability remains but is somewhat weaker, resulting in concentration of reads that while above average, does not reach the level required by our peak calling program..

To test that, for each peak we have computed rpm of both CSB ChIP and CSBΔN1 ChIP, and we defined a peak to be specific to one of the read sets if the respective rpm count was at least 4 times larger than the other. With that definition, 36% of the CSB peaks were specific to CSB and 24% of the CSBΔN1 peaks were specific to CSBΔN1. There was only one case each of CSB peak being specific to CSBΔN1 and vice versa, so the majority of peaks were non-specific. We could conclude that B was the dominant pattern, with the number of peaks specific to CSBΔN1 being not very significant. This conclusion is still true even if we make more relaxed definition of "specific", e.g., the ratio of rpm counts being at least 3 rather than at least 4 (Table 2).

Classification of the loci of CSB peaks indicates its role in regulating gene expression

To find clues about possible functions of CSB peaks we classified them into two ways: using gene annotations that give gene starts and ends, exon lists etc., and using the information on chromatin modifications in model cell types.

We have used CEAS package to find if the peaks have significant relation to chromosome regions defined by gene annotations, namely intergenic, promoter/TSS, 3′end/TES, intron and exon, the results are in Tables S2 and S3 of [1]. The most notable enrichment was for promoter regions that form 1.1% of the annotated genome and are occupied by 1.5% of CSB peaks and 3.1% of CSBΔN1 peaks. CEAS package also produces relevant pie-charts and p-values [8].

We used supplementary data of Ernst and Kellis (2011) [12] and custom scripts (attached in supplementary materials) to find the relation to regions defined by epigenetic modifications. These authors used a large panel of histone modifications known to have impact on gene regulation and defined 15 classes of genomic regions using an unsupervised learning algorithm based on HMM. These classes have a straightforward interpretation due to the prior knowledge of their role in gene expression. Ernst et al. provided classification for a number of cell types that included two types of fibroblasts, sister types to the CS1AN-sv cells used in our experiments. We classified each peak using the location of its center. In this discussion we refer to regions established for NHLF (normal human lung fibroblast). As we show in Table 3, the results are similar

Table 2
CSB and CSBΔN1 peaks classified according to the ratio between W, the normalized number of reads of CSB (the number of reads mapped to the surrounding 200 bp, divided by the number of all mapped reads) and M, the normalized number of reads CSBΔN1. We show both counts and percentages (in the additional rows below the rows with counts).

Peak type	M/W ≥				W/M ≥			
	4	3	2	1	1	2	3	4
CSB (W)	1	1	16	300	3307	4459	3329	6403
%	0.01	0.01	0.09	1.68	18.56	25.03	18.69	35.94
CSBΔN1 (M)	879	361	662	1251	440	45	5	1
%	24.12	9.91	18.17	34.33	12.07	1.23	0.14	0.03

Table 3
CSB peaks in regions defined by epigenetic marks in different cell types. Ernst et al. have three types of "Promoter regions" and four types of "Enhancer regions". Percentage refers to the proportion of peaks, and enrichment, to the ratio "percentage of peaks/percentage of genome". For comparison, we also give statistics for p300 peaks found in hESC cells.

		Promoters		Enhancers	
		% of peaks	Enrichment	% of peaks	Enrichment
CSB	NHLF	2.840%	2.108	19.084%	4.198
	HMEC	2.172%	1.789	21.941%	3.763
	hESC	3.137%	1.875	9.171%	2.312
p300	hESC	47.267%	28.253	22.605%	5.700

for HMEC (human mammary epithelial cells) but as expected, the relationship to the regions established for hESC (embryonic stem cells) is much weaker.

Transcribing promoter classes cover 1.1% of the genome, 2.7% of CSB peaks and 7.3% of CSBΔN1 peaks. Note that an epigenetically defined transcribed promoter, a locus with high level of H3K4me3 is absent in many genes at their 5′ end, while in other genes it occupies a longer region than the average, which explains a stronger relationship (enrichment, p-value) of CSB peaks with promoters defined in that fashion, rather than with promoters defined by fixed intervals around TSS sites. However, the strongest relationship exists with all enhancer classes, as they cover 4.4% of the genome (both intergenic and intragenic), and contain 19% of CSB peaks and 25% of CSBΔN1 peaks.

The strong association with enhancer regions strongly argues that CSB is important for the regulation of gene expression. However, this relationship is weaker than for p300 that has peaks almost exclusively in promoters and enhancers. That suggests that CSB is engaged in a wider variety of chromatin activities than the highly specialized histone acetyltransferase p300. However, p300 binding occurs both in short "peaks", with less than 1000 bps, and long "regions" that cover ca. 20% of the genome, and our comparison is restricted to peaks, so this suggestions should be qualified.

Motif analysis reveals the role of AP-1 complex in sequence specific placing of CSB

We used HOMER package to search for occurrence and enrichment of known binding motifs of various TFs (transcription factors). To eliminate irrelevant motifs we removed from considerations motifs that occur in less than 10% of peaks and have enrichment factor below 2. The largest group of motifs matched so-called TPA response element, i.e., consensus sequence TGASTCA (where S denotes a G or C). This response element is associated with a number of TFs, of which AP-1, or Fos-cJun, is most frequently involved in gene regulation. The physical association of CSB and cJun was verified with chromatin immunoprecipitation. A smaller group of motifs (in terms of the number of occurrences) had a consensus matching known motif of CTCF.

Forming specific conjecture and validation

Using the identified peaks, their annotations with putative target genes, expression levels of those genes and the presence of TPA response elements we identified loci most likely to regulate the expression of the nearby genes. Targeted experiments confirmed most of those functional associations. Notably, the impact of gene expression was in some cases enhancing the expression, and repressing in other cases. This is in keeping with general observations about the so-called enhancers. While initially they were associated with enhancing the gene expression, they may be more accurately described as regions where the binding of TFs is easier, with some TFs binding exclusively in loci with that type of histone modifications. However, these chromatin enzymes that form complexes with TFs may have different functions.

Other experiments confirmed that in most cases the regulatory activity of CSB required its ability as nucleosome remodeler, but removing that ability with a genetic modification retained the regulatory impact of certain binding loci. Most probably, CSB participates in a number of complexes with a variety of functions, and the full elucidation of its role will require further investigations.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gdata.2014.08.009.

Acknowledgments

All members of Fan lab contributed to the design of the analysis described here. I would like to specifically thank Dr. Hua-Ying Fan who took the time from her busy schedule to review this manuscript.

References

[1] R.J. Lake, E.L. Boetefuer, P.F. Tsai, J. Jeong, I. Choi, K.J. Won, H.Y. Fan, The sequence-specific transcription factor c-Jun targets Cockayne Syndrome Protein B to regulate transcription and chromatin structure. PLoS Genet. 10 (2014) e1004284.
[2] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=vfmfryagmygcony&acc=GSE50171.
[3] R.J. Lake, H.Y. Fan, Structure, function and regulation of CSB: a multi-talented gymnast. Mech. Ageing Dev. 134 (2013) 202–211.
[4] J.C. Newman, A.D. Bailey, A.M. Weiner, Cockayne Syndrome group B protein (CSB) plays a general role in chromatin maintenance and remodeling. Proc. Natl. Acad. Sci. U. S. A. 103 (2006) 9613–9618.
[5] A.D. Bailey, L.T. Gray, T. Pavelitz, J.C. Newman, K. Horibata, et al., The conserved Cockayne syndrome B-piggyBac fusion protein (CSB-PGBD3) affects DNA repair and induces both interferon-like and innate antiviral responses in CSB-null cells. DNA Repair (Amst) 11 (2012) 488–501, http://dx.doi.org/10.1016/j.dnarep.2012.02.004.
[6] I. Cho, P.F. Tsai, R.J. Lake, A. Basheer, H.Y. Fan, ATP-dependent chromatin remodeling by Cockayne Syndrome Protein B and NAP1-like histone chaperones is required for efficient transcription-coupled DNA repair. PLoS Genet. 9 (2013) e1003407.
[7] C. Troelstra, A. Van Gool, J. de Wit, W. Vermeulen, D. Bootsma, et al., ERCC6, a member of a subfamily of putative helicases, is involved in Cockayne's Syndrome and preferential repair of active genes. Cell 71 (1992) 939–953.
[8] H. Shin, T. Liu, A.K. Manrai, X.S. Liu, CEAS: cis-regulatory element annotation system. Bioinformatics 25 (2009) 2605–2606.
[9] R.J. Lake, A. Geyko, G. Hemashettar, Y. Zhao, H.Y. Fan, UV-induced association of the CSB remodeling protein with chromatin requires ATP-dependent relief of N-terminal autorepression. Mol. Cell 37 (2010) 235–246.
[10] L.T. Gray, K.K. Fong, T. Pavelitz, A.M. Weiner, Tethering of the conserved piggyBac transposase fusion protein CSB-PGBD3 to chromosomal AP-1 proteins regulates expression of nearby genes in humans. PLoS Genet. 8 (2012) e1002972.
[11] J.H. Dennis, H.Y. Fan, S.M. Reynolds, G. Yuan, J.C. Meldrim, et al., Independent and complementary methods for large-scale structural analysis of mammalian chromatin. Genome Res. 17 (2007) 928–939.
[12] J. Ernst, M. Kellis, ChromHMM: automating chromatin-state discovery and characterization. Nat. Methods 9 (2012) 215–216.