Genomics Data 7 (2016) 29–31

Contents lists available at ScienceDirect

Genomics Data

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

Data in Brief

p53 transcriptional programs in B cells upon exposure to genotoxic stress *in vivo*: Computational analysis of next-generation sequencing data

Claudia Tonelli [a], Bruno Amati [a,b], Marco J. Morelli [b,*]

[a] Department of Experimental Oncology, European Institute of Oncology (IEO), Via Adamello 16, 20139 Milan, Italy
[b] Center for Genomic Science of IIT@SEMM, Fondazione Istituto Italiano di Tecnologia (IIT), Via Adamello 16, 20139 Milan, Italy

ARTICLE INFO

Article history:
Received 28 October 2015
Accepted 6 November 2015
Available online 7 November 2015

Keywords:
p53
ChIP-Seq
RNA-Seq
Genotoxic stress
Motif analysis

ABSTRACT

The transcriptional programs activated by p53 in B cells *in vivo* following exposure to ionizing radiation were studied through the integrated analysis of various types of next-generation sequencing data: genome-wide profiling of p53 binding sites, mapping of histone marks and open chromatin regions and quantification of gene expression. Moreover, the binding of p53 was associated to a series of specific motifs on the DNA, which were directly inferred from the data. Here, we describe in detail the computational analysis of the datasets associated with our study (Tonelli et al., Oncotarget 6 (2015), 24611-26), deposited in the GEO archive (accession code GSE71180), and we provide the R scripts needed to generated the figures of the paper.

Specifications	
Organism/cell line/tissue	Mouse (C57/Bl6 B cells and non-B cells; p53KO B cells and non-B cells)
Sex	Not applicable
Sequencer or array type	Illumina Hi-Seq 2000
Data format	Raw and analyzed
Experimental factors	Spleens from C57/Bl6 and p53KO mice were collected 4 h after exposure to 7 Gy total body irradiation and from a control cohort of mice. After pressing the spleens through nylon cell strainers and hypotonic lysis of red blood cells, the cell suspensions were incubated with B220 MicroBeads (Miltenyi Biotec) and B cells were enriched by magnetic cell sorting (MACS), according to the manufacturer's instructions (Miltenyi Biotec). The remaining fraction constituted the non-B cell populations used in this study.
Experimental features	Previously described cell types were used for ChIP-Seq (for p53), RNA-Seq and DNase-Seq experiments.
Consent	n/a
Sample source location	Milan, Italy

1. Direct link to deposited data

http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE71180.

2. Experimental design, materials and methods

The GEO submission SuperSeries GSE71180, associated with the Tonelli et al. study [1], contains a total of 32 NGS samples, divided in three series: GSE71175, containing 6 ChIP-Seq samples (5 ChIP against p53 and one Input); GSE71176, containing 24 RNA-Seq samples (4 conditions with 2 replicates each for the p53 KO cells, 4 conditions with 4 replicates each for the C57/Bl6 cells); GSE71177, containing a DNase-Seq sample and the corresponding input. The datasets are summarized in Table 1.

These samples allowed studying the genomic occupancy and the transcriptional changes induced by p53 activation in B and non-B cells *in vivo*, following DNA damage produced by ionizing radiation. Cells from p53 null mice were analyzed to define the p53-dependent response.

3. Data analysis

We complement the methods of the original publication and the instructions deposited in the GEO archive with the source code used to produce the Figures from the Next-Generation Sequencing (NGS) data files. Under the accession number GSE71180, we provided the raw data files (sequencing reads, in fastq format), plus a series of processed data files: for the ChIP-Seq and DNase-Seq samples (excluding the inputs), we supplied the locations of the bound genomic regions in BED format, as obtained with the MACS [2] peak caller (v. 2.0.9), while for the RNA-Seq samples, we provided the quantification of the expression of each gene, *i.e.* the number of reads assigned to every gene,

* Corresponding author.
E-mail address: marco.morelli@iit.it (M.J. Morelli).

http://dx.doi.org/10.1016/j.gdata.2015.11.006

Table 1
Summary of the 32 samples available in the GSE71180 SuperSeries.

Sample ID	Sample name	Replicate	Data type
GSM1828855	p53.wt.Bcells.mock	1/1	ChIP-Seq
GSM1828856	p53.wt.Bcells.IR	1/1	ChIP-Seq
GSM1828857	p53.wt.nonBcells.mock	1/1	ChIP-Seq
GSM1828858	p53.wt.nonBcells.IR	1/1	ChIP-Seq
GSM1828859	p53.null.spleen.IR	1/1	ChIP-Seq
GSM1828860	Input	1/1	ChIP-Seq
GSM1828861	p53.null.Bcells.mock.1	1/2	RNA-Seq
GSM1828862	p53.null.Bcells.mock.2	2/2	RNA-Seq
GSM1828863	p53.null.nonBcells.mock.1	1/2	RNA-Seq
GSM1828864	p53.null.nonBcells.mock.2	2/2	RNA-Seq
GSM1828865	p53.null.Bcells.IR.1	1/2	RNA-Seq
GSM1828866	p53.null.Bcells.IR.2	2/2	RNA-Seq
GSM1828867	p53.null.nonBcells.IR.1	1/2	RNA-Seq
GSM1828868	p53.null.nonBcells.IR.2	2/2	RNA-Seq
GSM1828869	p53.wt.Bcells.mock.1	1/4	RNA-Seq
GSM1828870	p53.wt.Bcells.mock.2	2/4	RNA-Seq
GSM1828871	p53.wt.Bcells.mock.3	3/4	RNA-Seq
GSM1828872	p53.wt.Bcells.mock.4	4/4	RNA-Seq
GSM1828873	p53.wt.nonBcells.mock.1	1/4	RNA-Seq
GSM1828874	p53.wt.nonBcells.mock.2	2/4	RNA-Seq
GSM1828875	p53.wt.nonBcells.mock.3	3/4	RNA-Seq
GSM1828876	p53.wt.nonBcells.mock.4	4/4	RNA-Seq
GSM1828877	p53.wt.Bcells.IR.1	1/4	RNA-Seq
GSM1828878	p53.wt.Bcells.IR.2	2/4	RNA-Seq
GSM1828879	p53.wt.Bcells.IR.3	3/4	RNA-Seq
GSM1828880	p53.wt.Bcells.IR.4	4/4	RNA-Seq
GSM1828881	p53.wt.nonBcells.IR.1	1/4	RNA-Seq
GSM1828882	p53.wt.nonBcells.IR.2	2/4	RNA-Seq
GSM1828883	p53.wt.nonBcells.IR.3	3/4	RNA-Seq
GSM1828884	p53.wt.nonBcells.IR.4	4/4	RNA-Seq
GSM1828885	p53.wt.Bcells.DNaseI	1/1	DNase-Seq
GSM1828886	Input.DNaseI	1/1	DNase-Seq

normalized to the gene length and to the total number of reads aligned on any exon of any gene. We called this quantification exonic RPKM, or eRPKM, to distinguish it for the conventional normalization of read counts to the total number of aligned reads (anywhere on the genome). Most information needed to produce the figures is already available in the processed data, with the exception of four fields for the ChIP-Seq peaks: 1) annotation, 2) enrichment, 3) summit and 4) motif annotation. Here, we provide the complete resources needed to reproduce the figures of the main paper, and the instructions to generate the missing information. Finally, the genomic regions associated with previously published histone modifications [3–4] are also attached for convenience.

3.1. Analysis environment

Data analysis was entirely performed in R, the widely-used open-source environment for statistical computing and data analysis. The main package used for the analysis is CompEpiTools v1.2.6 [5], which is part of the BioConductor project [6] and it can be installed from the URL http://www.bioconductor.org/packages/release/bioc/html/compEpiTools.html. CompEpiTools is a flexible and user-friendly package to perform basic analyses of NGS data.

3.2. Description of the source files

The source code TonelliEtAl2015_sourceCode.zip is composed of 5 files and 2 directories, described below:

- filemapping_GEO.R

This file contains the links between the R objects used to produce the figures and the files deposited on the GEO archive, listed in Table 1. In particular, ChIP-Seq BED files are converted to GRanges and gene expression quantifications are organized in a data frame. This code also arranges in a list ChIP-Seq alignment (BAM) files, which must be obtained from the raw sequencing files (fastq) following the instructions deposited on the GEO archive.

- analysisEnvironment.R

This R script loads all the data files needed to produce the figures of the main paper [1], and contains all the libraries and functions invoked in the R scripts contained in the file TonelliEtAl2015_Figures.R. In particular, the environment requires the following libraries: compEpiTools, gplots, VennDiagram, lattice, flashClust, TxDb.Mmusculus.UCSC.mm9.knownGene, and org.Mm.eg.db (see sessionInfo.log).

- TonelliEtAl2015_Figures.R

This R script sources the instructions contained in the file analysisEnvironment.R to load the pre-generated data objects associated with the main paper [1], and contains the code used to produce all the figures referring to the computational analyses of NGS data. Occasionally, some figures require computing tag density on genomic intervals, and therefore require alignment (BAM) files: in these cases, a pre-computed table was used.

- prepareDatasets.R

This collection of R scripts allows complementing the processed ChIP-Seq files available on GEO with the extra fields required for the generation of the final figures. The output of these scripts is contained in the ChIPpeaks.rds datafile in the data directory, under the form of a list of genomic ranges, which is automatically loaded in the analysisEnvironment.R script. The scripts contained in prepareDatasets.R use several external tools (MEME [7], TOMTOM [8] FIMO [9]), the mm9 reference genome, available in Bioconductor in the library BSgenome.Mmusculus.UCSC.mm9 (v. 1.4.0), and may require a consistent amount of time (6–24 h) to complete, depending on the platform used. In order to execute the scripts, the processed ChIP-Seq files should first be downloaded from GEO and organized according to the instructions contained in the filemapping.R file. Subsequently, alignment (BAM) files must be generated from the raw fastq files deposited in GEO, following the instructions on the archive.

The extra fields consist in:

1) the genomic annotation of the p53 ChIP-Seq peaks: a peak overlapping with a [−5 kb, +2 kb] window around a standard promoter is considered “promoter”, those overlapping with an H3K4me1 peak, “enhancers”, otherwise they are classified as “distal”;
2) the enrichment of the peak: computed with the GRenrichment function in the compEpiTools suite;
3) the summit of the peak: computed with the GRcoverageSummit function in the compEpiTools suite;
4) the motif annotation of the peak, which is obtained through five main steps: i) the generation of a FASTA file containing the sequences of the top 1000 enriched genomic regions spanned by the peaks of the p53.wt.Bcells.IR sample; ii) the estimation of the unspaced p53 motif from these sequences using MEME [7] (we verify that the estimated motif coincides with the p53 canonical motif contained in the Jaspar Core Vertebrata database [10] using TOMTOM [8]); iii) the creation of the motifs with spacers, obtained by inserting sequences with constant probability over the 4 nucleotides (spacers) between the two half decameric sites; iv) the scoring of these motifs against the mouse genome with FIMO [9]; v) the assignment of the motifs to the ChIP-Seq peaks.

- SessionInfo.log

A log file containing the output of the R sessionInfo() command, specifying all the versions of all the libraries used in the analysis environment.

- data folder

This folder contains all the R objects needed to produce the figures of the main paper.

- figures folder

This folder contains all the figures of the main paper, in pdf format, obtained by running the scripts contained in TonelliEtAl2015_Figures.R.

Acknowledgments

This work was supported by funding from grants from the European Science Council (ERC) (268271), the Italian Health Ministry, Fondazione Cariplo (2009-2593) and the Italian Association for Cancer Research (AIRC) (13182) to B.A.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gdata.2015.11.006.

References

[1] C. Tonelli, M.J. Morelli, S. Bianchi, L. Rotta, T. Capra, A. Sabò, S. Campaner, B. Amati, Genome-wide analysis of p53 transcriptional programs in B cells upon exposure to genotoxic stress in vivo. Oncotarget 6 (2015) 24611–24626.
[2] Y. Zhang, T. Liu, C.A. Meyer, J. Eeckhoute, D.S. Johnson, B.E. Bernstein, C. Nusbaum, R.M. Myers, W. Li, X.S. Liu, Model-based analysis of ChIP-Seq (MACS). Genome Biol. 9 (2008) R137.
[3] A. Sabò, T.R. Kress, M. Pelizzola, S. de Pretis, M.M. Gorski, A. Tesi, M.J. Morelli, P. Bora, M. Doni, A. Verrecchia, C. Tonelli, G. Fagà, V. Bianchi, A. Ronchi, D. Low, H. Müller, E. Guccione, S. Campaner, B. Amati, Selective transcriptional regulation by Myc in cellular growth control and lymphomagenesis. Nature 511 (2014) 488–492.
[4] M. Pelizzola, M.J. Morelli, A. Sabò, T.R. Kress, S. de Pretis, B. Amati, Selective transcriptional regulation by Myc: experimental design and computational analysis of high-throughput sequencing data. Data Brief 3 (2015) 40–46.
[5] K. Kishore, S. de Pretis, R. Lister, M.J. Morelli, V. Bianchi, B. Amati, J.R. Ecker, M. Pelizzola, methylPipe and compEpiTools: a suite of R packages for the integrative analysis of epigenomic data. BMC Bioinf. 16 (2015) 313.
[6] R.C. Gentleman, V.J. Carey, D.M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A.J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J.Y.H. Yang, J. Zhang, Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 5 (2004) R80.
[7] T.L. Bailey, M. Gribskov, Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, AAAI Press, Menlo Park 1994, pp. 28–36.
[8] S. Gupta, J.A. Stamatoyannopoulos, T.L. Bailey, W. Stafford Noble, Quantifying similarity between motifs. Genome Biol. 8 (2007) R24.
[9] C.E. Grant, T.L. Bailey, W. Stafford Noble, FIMO: scanning for occurences of a given motif. Bioinformatics 27 (2011) 1017–1018.
[10] E. Portales-Casamar, S. Thongjuea, A.T. Kwon, D. Arenillas, X. Zhao, E. Valen, D. Yusuf, B. Lenhard, W.W. Wasserman, A. Sandelin, JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. Nucleic Acids Res. 38 (2010) D105.