Genomics Data 3 (2015) 15–18

Contents lists available at ScienceDirect

Genomics Data

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

Data in Brief

Mapping of genomic double-strand breaks by ligation of biotinylated oligonucleotides to forum domains: Analysis of the data obtained for human rDNA units

N.A. Tchurikov *, O.V. Kretova, D.M. Fedoseeva, V.R. Chechetkin, M.A. Gorbacheva, A.A. Karnaukhov, G.I. Kravatskaya, Y.V. Kravatsky

Engelhardt Institute of Molecular Biology, Moscow, Russia

ARTICLE INFO

Article history:
Received 30 October 2014
Accepted 30 October 2014
Available online 12 November 2014

Keywords:
Forum domains
Double-strand breaks
Fragile sites
rDNA
Bioinformatics
HEK293T

ABSTRACT

DNA double-strand breaks (DSBs) are associated with different physiological and pathological processes in different organisms. To understand the role of DSBs in multiple cellular mechanisms, a robust method for genome-wide mapping of chromosomal breaks at one-nucleotide resolution is required. Many years ago, we detected large DNA fragments migrating from DNA-agarose plugs in pulsed-field gels, which we named 'forum domains' [1,2]. Recently, we developed a method for genome-wide mapping of DSBs that produces these 50–150 kb DNA domains using microarrays or 454 sequencing (Tchurikov et al., 2011; 2013). Now we have used Illumina sequencing to map DSBs in repetitive rDNA units in human HEK293T cells. Here we describe in detail the experimental design and bioinformatics analysis of the data deposited in the Gene Expression Omnibus with accession number GSE49302 and associated with the study published in the Journal of Molecular Cell Biology (Tchurikov et al., 2014).

Specifications	
Organism/cell line/tissue	Homo sapiens/HEK 293T cells
Sex	Female
Sequencer or array type	Illumina Genome Analyzer IIx
Data format	Raw and processed. Raw data: FASTQ reads. Processed data: BED, WIG, and text table files. Metadata in SOFT and MINiML formats are supplied by GEO for automated processing.
Experimental factors	HEK293T cells were seeded in 10 cm culture plates 1–2 days before experiments in DMEM containing 10% FBS, and were used at approximately 60–80% confluency.
Experimental features	DNA domains, migrating in 0.8% agarose mini-gels from the DNA-agarose plugs, were electroeluted. Biotinylated oligonucleotides were ligated to DNA sequences at DSB sites.
Consent	*Level of consent allowed for reuse if applicable (typically for human samples).*
Sample source location	Moscow 119334, Russia

* Corresponding author at: Engelhardt Institute of Molecular Biology, Vavilov str. 32, Moscow, 119334, Russia. Tel.: +7 499 135 97 53; fax: +7 499 135 14 05.
E-mail address: tchurikov@eimb.ru (N.A. Tchurikov).

Direct link to deposited data

http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE49302.

Experimental design, materials and methods

Sample preparation

DNA-agarose plugs were prepared as described previously [2–6]. The major steps are illustrated in Fig. 1. About 6 million HEK 293T cells in 2 mL of culture medium were pelleted by centrifugation at 2000 rpm in a MiniSpin centrifuge (Eppendorf), resuspended in 0.3 mL of the same medium, gently mixed at 42 °C with an equal volume of a solution of 1% low-melt agarose L (LKB) in PBS, and distributed on a mold containing 100-μL wells. The mold was covered with parafilm and placed on ice for 2–5 min. The agarose plugs were then placed in Petri dishes with 5 mL of solution containing 0.5 M EDTA (pH 9.5), 1% sodium lauroylsarcosine, and 1–2 mg of proteinase K solution per mL for 40–48 h at 50 °C, and stored at 4 °C in the same solution. Each DNA-agarose plug usually contained about 15 μg of DNA, corresponding to about 1 million cells.

To test the quality of isolated DNA, fractionation in pulsed-field gels was performed as described previously [2,7]. Portions of the original agarose-DNA plugs (5–50 μL) containing 1–10 μg of DNA were used for electrophoresis without any restriction enzyme digestion. The DNA

http://dx.doi.org/10.1016/j.gdata.2014.10.024

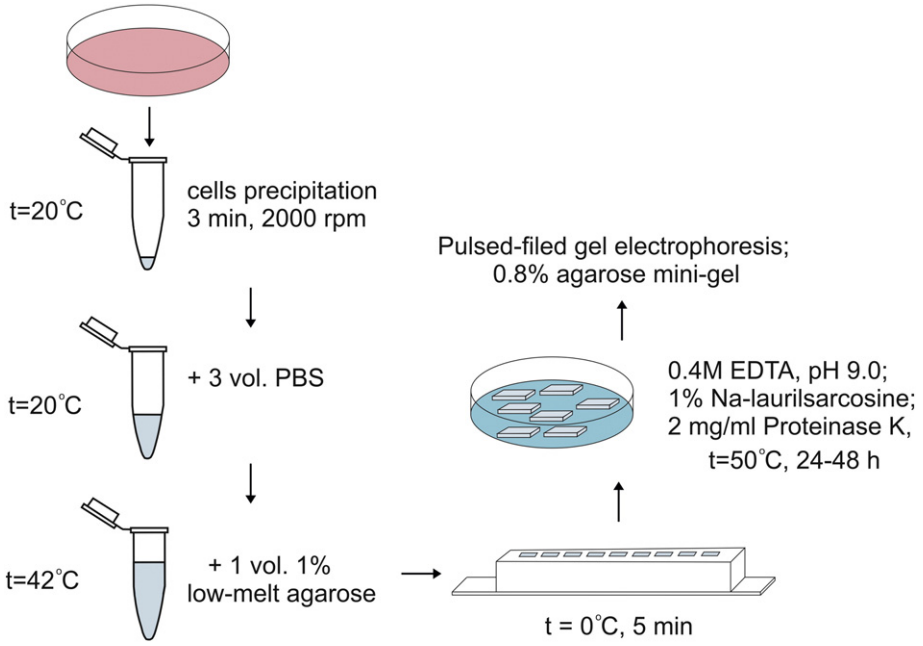

Fig. 1. Schematic representation of the procedure used for isolation of DNA samples inside 0.5% low-melt agarose.

samples were run in 0.8% agarose gels on a Pulsaphor system (LKB) using a hexagonal electrode and switching times of 25 or 450 s.

For elution of DNA preparations, fractionation in a 0.8% agarose conventional mini-gel was performed. One-half of the DNA-agarose plug was washed in 1× TE three times (for 15 min each), followed by washing (three times) in the same solution containing 17.4 μg/mL phenylmethylsulfonyl fluoride (PMSF) in ethanol. After fractionation in the mini-gel, the ethidium bromide-stained DNA band was excised and electroeluted inside the dialysis cellulose-membrane bag. After overnight dialysis without stirring against 1 L of 0.01× TE at 4 °C, the DNA was concentrated with PEG at 4 °C.

Rapid amplification of forum domains termini (RAFT) procedure

The steps of the procedure are shown schematically in Fig. 2A. About 1.5 μg of isolated DNA (see above) was ligated with 70 ng of double-stranded oligonucleotide (25-bp long 5′-phosphorylated 5′ pCCCCTGC AGTATAAGGAGAATTCGGG 3′ oligonucleotide annealed to a 26-bp long 5′ biotinylated 5′ bio-CCGAATTCTCCTTATACTGCAGGGG 3′ oligonucleotide) in 150 μL of solution containing 0.1 M NaCl, 50 mM Tris–HCl (pH 7.4), 8 mM $MgCl_2$, 9 mM 2-mercaptoethanol, 7 μM ATP, 7.5% PEG, and 40 units of T4 DNA ligase at 20 °C for 16 h. After heating at 65 °C for 10 min, the DNA preparation was digested with Sau3A enzyme to shorten the forum domain to the positions of the termini attached to the ligated oligonucleotide. The selection of such termini was performed in 0.5-mL Eppendorf tubes using 300 μL of a suspension containing Streptavidin Magnesphere Paramagnetic Particles, (SA-PMP; Promega) according to the manufacturer's recommendations. After extensive washing with 0.5× SSC to remove DNA fragments corresponding to the internal parts of forum domains, the forum termini (FT) DNA preparation was eluted from the SA-PMP using digestion with EcoRI enzyme in a final volume of 50 μL (double-stranded FT). The FT was then ligated with 100× molar excess of double-stranded Sau3A adaptor (5′-phosphorylated 5′ pGATCGTTTGCGGCCGCTTAAGCTTGGG 3′ oligonucleotide annealed to 5′ CCCAAGCTTAAGCGGCCGCAAAC 3′ oligonucleotide). In some experiments, the DNA preparation was eluted from the SA-PMP using heating by incubation at 100 °C for 3 min in 50 μL of 0.01× TE (single-stranded FT). Before heating, the FT preparation was ligated with 100× molar excess of double-stranded Sau3A adaptor in suspension with SA-PMP (see above). Both final DNA samples (double-stranded FT or single-stranded FT) were used for PCR amplifications. PCR amplification (15–20 cycles) in 30 μL of a solution containing 67 mM Tris–HCl (pH 8.4); 6 mM $MgCl_2$; 10 mM 2-mercaptoethanol; 16.6 mM ammonium sulfate; 6.7 μM EDTA; 5 μg/mL BSA; 1 mM dNTPs; 1 μg of primer corresponding to Sau3A adaptor (5′ CCCAAGCTTAAGCGGCCGCAAAC 3′); 1 μg of primer corresponding to

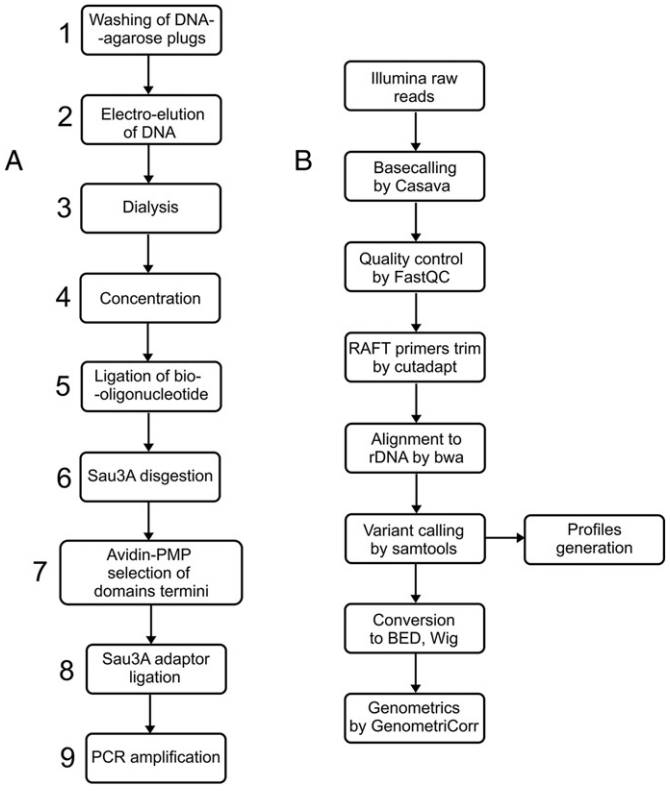

Fig. 2. Experimental and bioinformatics pipelines. (A) The major steps of the RAFT procedure. (B) Bioinformatics analysis pipeline.

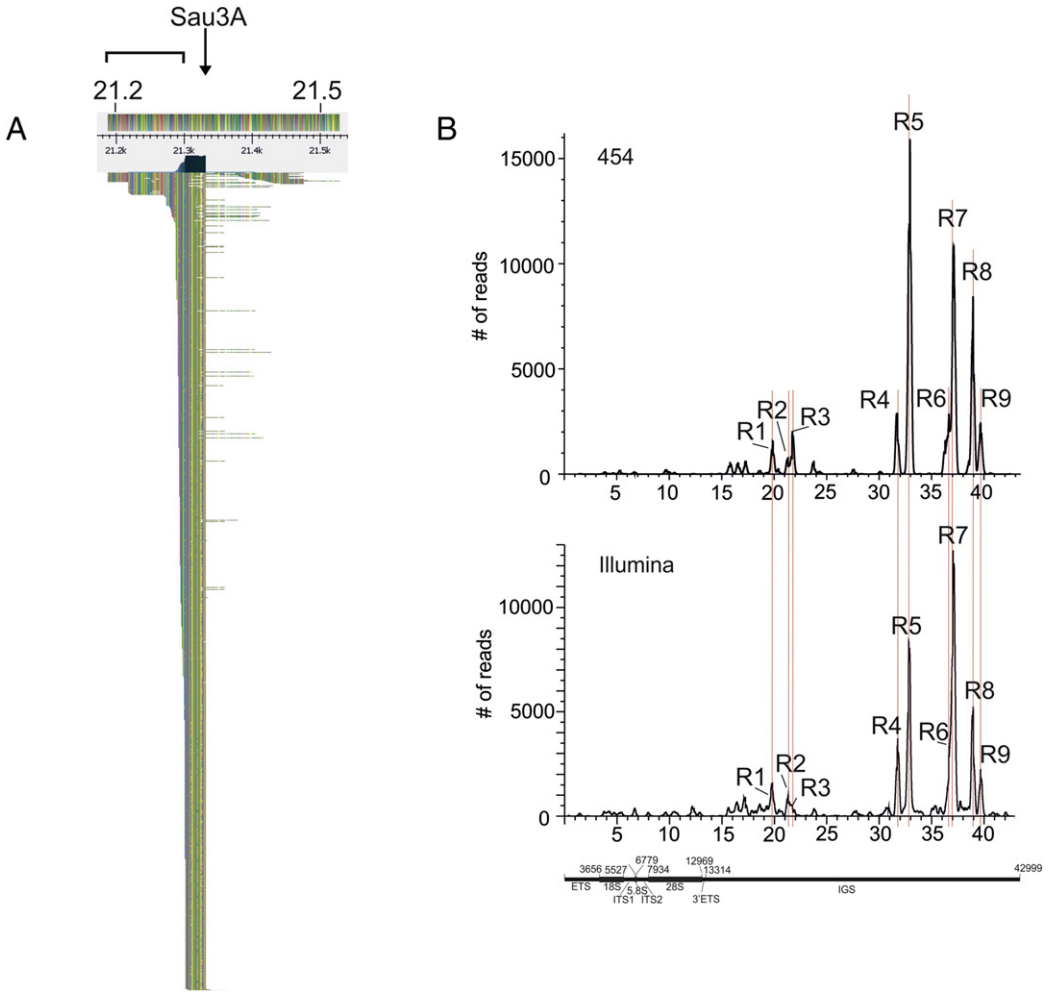

Fig. 3. Analysis of Illumina reads mapped inside rDNA units. (A) The mapping results of Illumina reads inside rDNA units using UGENE software (http://ugene.unipro.ru/). The reads (1197 rows) that mapped the region of rDNA between 21.2 and 21.5 kb coordinates inside the 43-kb rDNA sequence (accession number U13369) are shown schematically. The region is indicated in panel B as R2. There are regions possessing many more mapped Illumina reads (R4–R9 on the panel B). The bracket at the top shows the region of about 100 bp in length where DSBs are scattered. The arrow indicates the position of the Sau3A site that delimits the Illumina reads. (B). Comparison of profiles of DSBs determined in independent RAFT experiments using 454 or Illumina platforms. Hot spots of DSBs are indicated as R1–R9 (Pleiades).

biotinylated oligonucleotide (5′ CCGAATTCTCCTTATACTGCAGGGG 3′); and 1 U of Taq polymerase was performed using an Mastercycler Personal thermal cycler (Eppendorf). Amplification conditions were 90 °C for melting, 65 °C for annealing, and 72 °C for extension, for 1 min each.

Library preparation

Libraries were prepared according to Illumina's instructions accompanying the DNA Sample Kit (Part# 0801-0303). Briefly, DNA was end-repaired using a combination of T4 DNA polymerase, *Escherichia coli* DNA Pol I large fragment (Klenow polymerase), and T4 polynucleotide kinase. The blunt, phosphorylated ends were treated with Klenow fragment and dATP to yield a protruding 3′- A base for ligation of Illumina's adapters, which have a single T-base overhang at the 3′ end. After adapter ligation, DNA was PCR amplified with Illumina primers for 15 cycles, and library fragments of ~200–400 bp (insert plus adaptor and PCR primer sequences) were band isolated from an agarose gel. The purified DNA was captured on an Illumina flow cell for cluster generation. Libraries were sequenced on the Genome Analyzer IIx following the manufacturer's protocols.

Data processing

Fig. 2B shows the bioinformatics pipeline used. Illumina Casava 1.8 software was used for basecalling. All reads were merged in the one file. Next, reads were trimmed for RAFT primer sequences by cutadapt v. 1.2.1 using the following options: –minimum-length = 30 –trimmed-only –quality-base = 33 –quality-cutoff = 3 -n 2 -g CCCAAGCTTAAGCGGCCGCAAAC -g CCGAATTCTCCTTATACTGCAG GGG. Option "–trimmed-only" was used to remove all sequences

Table 1
Correlation of the data on mapping of DSBs in a rDNA unit obtained by 454 or Illumina sequencing.

Tracks	Relative distances by Kolmogorov–Smirnov test	Projection test	Jaccard test	Scaled absolute minimum distances test
Illumina–454	Passed, $p = 0.0047$	Passed, $p = 0$	Passed, $p = 0.001$	Passed, $p = 0.017$

The Illumina's track was used as the reference, and the Roche 454 track was used as the query. GenometriCorr can process only segmented tracks (not profiles) in formats like BED, and so we used this type of data for the calculation of correlations. Results that are presented in the table are consistent with the visual analysis, and support the consistency of DSB mapping using different platforms and independent RAFT experiments.

that do not have RAFT primers. Trimmed reads were mapped to rDNA (GenBank Accession number U13369) and to hg19/GRCh37p10 by bwa [8] 0.7.5a using mem algorithm and SAMtools 0.1.12a-r862 [9]. Variant calling was also performed by SAMtools. Final mappings were converted for further analysis into tables and formats, including BED and WIG, by *ad hoc* Perl scripts. The further genometric analysis was performed using GenometriCorr software package [10].

Profile-like curves were obtained in the following way. First, the density coverage for the each alignment file was obtained by BEDTools [11]: bamToBed -ed. Second, the data were converted by density.bed to the profile data with F-seq [12]: fseq -f 200 density.bed. The resulting WIG files were converted to the common ASCII coordinate format files by our own *ad hoc* Perl script.

Discussion

The RAFT procedure includes several steps of manipulations with very long DNA molecules in solution (Fig. 2A)—from elution of DNA domains to ligation of biotinylated oligonucleotide (steps 2–5 in Fig. 2A). Although only a gentle mixing of solution after addition of ligase was performed, a random fragmentation of forum domains cannot be excluded during these steps. Nevertheless, our data demonstrate that the level of this random hydrodynamic fragmentation of DNA molecules in the conditions used is much lower than the non-random fragmentation detected at hot spots of DSBs (Fig. 3). The outline of mapped reads inside rDNA within one hot spot is shown in Fig. 3A. Nine major hot spots of DSBs, which we denote as Pleiades, were detected (Fig. 3B). We are aware that these data correspond to repeated rDNA units. There are about 300 copies of rDNA genes in the human genome [13]. It follows that to map the hot spot of DSBs with the same robustness as within unique genomic regions, one needs a higher number of original Illumina reads corresponding to the entire genome. Currently, we perform such analyses using HiSeq 2000 reads.

The validation of the approach was performed by comparison of the data obtained in different experiments using both the independent RAFT preparations and the deep-sequencing platforms. In these experiments, the same profiles of DSB hot spots were detected inside human rDNA units (Fig. 3B). The data regarding the correlation between the mapping data using GenometriCorr software package [10] are shown in Table 1.

The data on hot spots of DSBs inside rDNA units strongly suggest that the *in vivo* chromosomal breakage is associated with active transcription in these units, producing up to 80% of total cellular RNA in human cells [13]. Our data indicate that the hot spots of DSBs correspond to sites that are important for both regulation of expression and for inter- and intra-chromosomal interactions with specific regions of chromosomes also possessing hot spots of DSBs [5,14]. These data, taken together with the data on coordinated expression of genes located in forum domains that are delimited by hot spots of DSBs and binding sites of PARP1 and HRNPA2B1 [4], suggest that forum domains share some properties in common with so-called topological or interacting domains [15]. The data on mapping of hot spots of DSBs are important in cancer genomics and for the study of chromosomal translocations, including Robertsonian translocations involving five human chromosomes bearing rDNA clusters.

Acknowledgments

This work was supported by a grant from the Molecular and Cellular Biology Program of the Russian Academy of Sciences (#130220122525) and by grants from the Russian Foundation for Basic Research (#12-04-01416-a, #12-04-01311-a, #14-04-01638-a, and #15-04-00299-a).

References

[1] N.A. Tchurikov, N.A. Ponomarenko, L.G. Airich, Isolation of forum DNA—a specific fraction in human DNA. Dokl. Akad. Nauk USSR 303 (1988) 491–497.
[2] N.A. Tchurikov, N.A. Ponomarenko, Detection of DNA domains in *Drosophila*, human and plant chromosomes possessing mainly 50- to 150-kilobase stretches of DNA. Proc. Natl. Acad. Sci. U. S. A. 89 (1992) 6751–6755.
[3] N.A. Tchurikov, O.V. Kretova, D.V. Sosin, I.A. Zykov, I.F. Zhimulev, Y.V. Kravatsky, Genome-wide profiling of forum domains in *Drosophila melanogaster*. Nucleic Acids Res. 39 (2011) 3667–3685.
[4] N.A. Tchurikov, O.V. Kretova, D.M. Fedoseeva, D.V. Sosin, S.A. Grachev, M.V. Serebraykova, S.A. Romanenko, N.V. Vorobieva, Y.V. Kravatsky, DNA double-strand breaks coupled with PARP1 and HNRNPA2B1 binding sites flank coordinately expressed domains in human chromosomes. PLoS Genet. 9 (4) (2013) e1003429.
[5] N.A. Tchurikov, D.M. Fedoseeva, D.V. Sosin, A.V. Snezhkina, N.V. Melnikova, A.V. Kudryavtseva, Y.V. Kravatsky, O.V. Kretova, Hot spots of DNA double-strand breaks and genomic contacts of human rDNA units are involved in epigenetic regulation. J. Mol. Cell Biol. 0 (0) (2014) 1–17.
[6] N.A. Tchurikov, A.N. Krasnov, N.A. Ponomarenko, Y.B. Golova, B.K. Chernov, Forum domain in *Drosophila melanogaster* cut locus possesses looped domains inside. Nucleic Acids Res. 26 (1998) 3221–3227.
[7] N.A. Tchurikov, O.V. Kretova, B.K. Chernov, Y.B. Golova, I.F. Zhimulev, I.A. Zykov, SuUR protein binds to the boundary regions separating forum domains in *Drosophila melanogaster*. J. Biol. Chem. 279 (2004) 11705–11710.
[8] H. Li, R. Durbin, Fast and accurate long-read alignment with Burrows–Wheeler Transform. Bioinformatics 26 (2010) 589–595.
[9] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, 1000 Genome Project Data Processing Subgroup, The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics 25 (2009) 2078–2209.
[10] A. Favorov, L. Mularoni, L.M. Cope, Y. Medvedeva, A.A. Mironov, V.J. Makeev, S.J. Wheelan, Exploring massive, genome scale datasets with the GenometriCorr Package. PLoS Comput. Biol. 8 (2012) e1002529.
[11] A.R. Quinlan, I.M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26 (2010) 841–842.
[12] A.P. Boyle, J. Guinney, G.E. Crawford, T.S. Furey, F-Seq: a feature density estimator for high-throughput sequence tags. Bioinformatics 24 (2008) 2537–2538.
[13] D.M. Stults, M.W. Killen, E.P. Williamson, J.S. Hourigan, H.D. Vargas, S.M. Arnold, J.A. Moscow, A.J. Pierce, Human rRNA gene clusters are recombinational hotspots in cancer. Cancer Res. 69 (2009) 9096–9104.
[14] D.V. Sosin, O.V. Kretova, Y.V. Kravatsky, N.A. Tchurikov, Analysis of genome-wide contacts of forum terminus in *Drosophila* S2 cells. Dokl. Biochem. Biophys. 452 (2013) 259–263.
[15] J.R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. Liu, B. Ren, Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 11 (2012) 376–380.