Genomics Data 10 (2016) 19–21

Contents lists available at ScienceDirect

Genomics Data

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

Data in Brief

Fine resolution mapping of double-strand break sites for human ribosomal DNA units

Bernard J. Pope [a], Khalid Mahmood [a], Chol-hee Jung [a], Daniel J. Park [a,b,*]

[a] Victorian Life Sciences Computation Initiative, The University of Melbourne, Australia
[b] Genomic Technologies Group, Genetic Epidemiology Laboratory, Department of Pathology, The University of Melbourne, Australia

ARTICLE INFO

Article history:
Received 11 August 2016
Accepted 22 August 2016
Available online 24 August 2016

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

ABSTRACT

DNA breakage arises during a variety of biological processes, including transcription, replication and genome rearrangements. In the context of disease, extensive fragmentation of DNA has been described in cancer cells and during early stages of neurodegeneration (Stephens et al., 2011 Stephens et al. (2011) [5]; Blondet et al., 2001 Blondet et al. (2001) [1]). Stults et al. (2009) Stults et al. (2009) [6] reported that human rDNA gene clusters are hotspots for recombination and that rDNA restructuring is among the most common chromosomal alterations in adult solid tumours. As such, analysis of rDNA regions is likely to have significant prognostic and predictive value, clinically. Tchurikov et al. (2015a, 2016) Tchurikov et al. (2015a, 2016) [7,9] have made major advances in this direction, reporting that sites of human genome double-strand breaks (DSBs) occur frequently at sites in rDNA that are tightly linked with active transcription - the authors used a RAFT (rapid amplification of forum termini) protocol that selects for blunt-ended sites. They reported the relative frequency of these rDNA DSBs within defined co-ordinate 'windows' of varying size and made these data (as well as the relevant 'raw' sequencing information) available to the public (Tchurikov et al., 2015b). Assay designs targeting rDNA DSB hotspots will benefit greatly from the publication of break sites at greater resolution. Here, we re-analyse public RAFT data and make available rDNA DSB co-ordinates to the single-nucleotide level.

1. Direct link to deposited data

https://figshare.com/s/1d21827bb891461845cc

2. Experimental design, materials and methods

2.1. Sequencing data

The FASTQ file for Illumina Genome Analyzer IIx run accession SRR944107 was downloaded from http://www.ebi.ac.uk/ena/data/view/SRR944107, having sourced the accession code via http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE49302. The origin of these data has been reported previously [8]. Briefly, HEK 293T cells were suspended in 1% (w/v) low-melt agarose prior to lysis. DNA was then fractionated by gel electrophoresis and collected by electroelution. Free DNA ends (sites of DSBs) were ligated to a double-stranded biotinylated adapter oligonucleotide before digestion with the restriction endonuclease *Sau*3AI. DSB site-containing termini were phase-purified using streptavidin paramagnetic particles prior to Sau3AI site adapter ligation and PCR amplification. PCR products were ligated to Illumina adapters such that they had the potential to be represented in either orientation. Library fragments of ~200–400 bp (insert plus adapter and PCR primer sequences) were band isolated from an agarose gel and the purified DNA was subjected to Illumina Genome Analyzer IIx sequencing.

2.2. Data processing

Fig. 1 provides a schematic representation of our bioinformatic analysis pipeline. In the first step, we produced custom software [raft_fastq_parse.py] to produce a modified representation of SRR944107.fastq, SRR944107_cleaned.fastq. We will describe this tool in more detail and make it available to the public via a forthcoming publication. Briefly, it filters reads based on the observation of expected arrangements of adapter sequences. Reads exhibiting evidence of ligation artefacts or insufficient evidence of expected adapter sequences were

* Corresponding author at: Genomic Technologies Group, Genetic Epidemiology Laboratory, Department of Pathology, The University of Melbourne, Australia.
E-mail address: djp@unimelb.edu.au (D.J. Park).

http://dx.doi.org/10.1016/j.gdata.2016.08.012

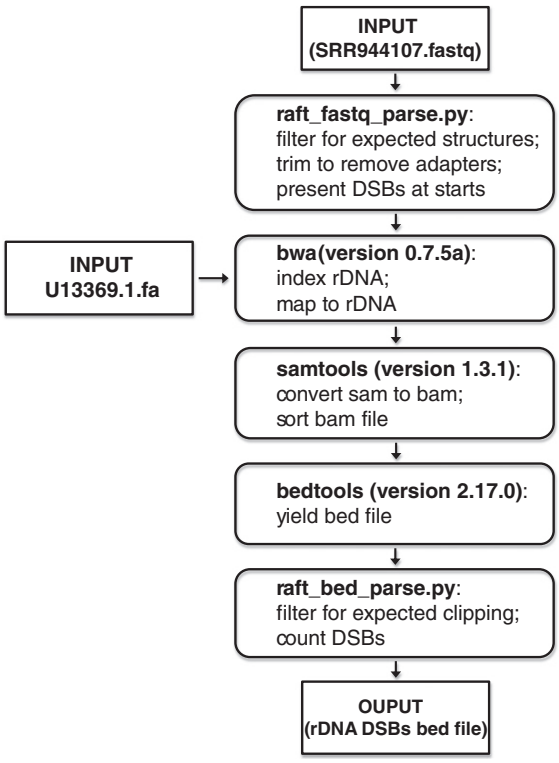

Fig. 1. Schematic illustration of our bioinformatic analysis pipeline to derive counts of DSBs in human rDNA by co-ordinate.

removed. Reads oriented with the DSB site towards the start were processed to remove adapter sequence(s). Reads with the DSB site oriented distally were also processed to remove adapter sequence(s) and were transformed to be represented with the DSB site at the start. Reads harbouring the *Sau*3AI site towards the start but without both adapters being evident were removed because the DSB site could not be defined. Reads were retained if the library insert was greater than or equal to 25 nucleotides in length.

The human rDNA sequence U13369.1.fa was indexed using BWA (version 0.7.5a) [2] using the command:

```
bwa index -a is U13369.1.fa U13369.1
```

Reads of the transformed FASTQ file were then mapped to U13369.1 using BWA, thus:

```
bwa mem U13369.1.fa SRR944107_cleaned.fastq
SRR944107_cleaned_rDNA.sam
```

Samtools (version 1.3.1) [3] was used to convert from SAM file format to BAM file format and to sort the resulting BAM file with the following command:

```
samtools view -u SRR944107_cleaned_rDNA.sam | samtools
sort -@ 4 -o SRR944107.rDNA.sort.bam
```

Bedtools (version 2.17.0) [4] was then employed to produce a BED file representing the mapping, including CIGAR string information and mapping orientation, with the following command:

```
bedtools bamtobed -cigar -i SRR944107.rDNA.sort.bam >
SRR944107.rDNA.bed
```

We then produced custom software [raft_bed_parse.py] to further filter the data and to count the number of times DSBs were observed at co-ordinates in rDNA U13369.1. We will include further details of this and make it available to the public along with [raft_fastq_parse.py]. Briefly, this tool assesses the orientation of mapping for each read. Because we presented the DSB at the beginning of each read prior to mapping, we can determine the exact location of the DSB at the single nucleotide level for each read. This tool also performs additional filtering steps. Reads that mapped in the '+' orientation (sequence was compared directly to the 'sense' strand of U13369.1 to result in successful mapping) were tested for evidence of 'clipping' at the start of the cigar string. If such clipping was evident, the read did not contribute to DSB recording - instead, the would-be DSB signal was treated as likely artefactual in nature. Similarly, reads that mapped in the '-' orientation (reverse complemented sequence mapped successfully to the 'sense' strand of U13369.1) were treated as likely artefactual if the cigar string showed clipping at the end.

The most frequent single nucleotide-resolved sites of rDNA DSBs derived from this analysis fall within the hotspot windows reported previously [8]. Fig. 2 illustrates the frequency of DSBs at single nucleotide resolution in the hotspot-enriched rDNA region between co-ordinates 30,000 and 40,000 of rDNA. Fig. 3 depicts such sites across the entire length of human rDNA U13369.1.

3. Discussion

Here, we present sites contributing to DSBs in human rDNA, among the most fragile regions of the genome, at single nucleotide resolution. The most frequent sites for rDNA DSBs concur with previously reported genomic windows [7,8]. Given the relevance of these sites to cancer and other diseases, it is likely that these new data will prove to be useful for the development of predictive, diagnostic and prognostic assays of clinical importance [1,5,6,9]. The fine resolution of such sites will allow better targeted and, therefore, cost-effective approaches to assessing an individual's level of genomic 'scarring'. This will likely be informative in stratifying risks for particular types of cancer and determining likely responses to particular cancer treatments, for example. Such information will be extremely important in informing preventative screening programs and best matching patients to the most likely effective treatments, with enormous potential to reduce morbidity and mortality.

Our methods will be applicable to future RAFT datasets derived from the use of a variety of restriction endonucleases, alone or in combinations (instead of the sole use of *Sau*3AI). Analysis of such data will improve accuracy incrementally by reducing bias in the form of restriction endonuclease site positioning effects – the locations of these sites relative to DSB sites determines the library insert size distribution, which, in turn, affects relative amplification efficiency and the ability to map reads accurately. In future experiments, the use of

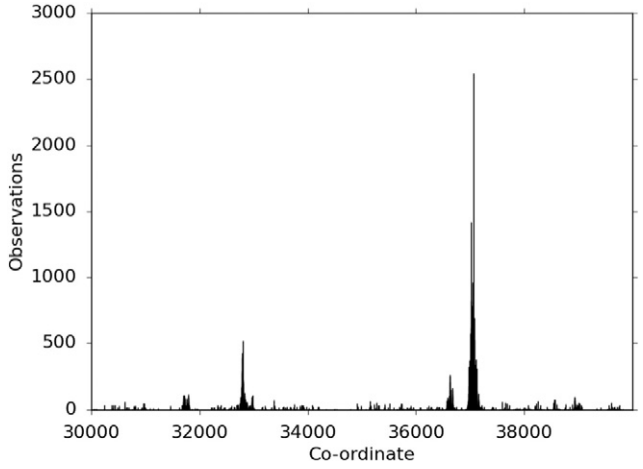

Fig. 2. Histogram depicting counts of DSBs by co-ordinate based on mapping to human rDNA U13369.1.fa. Results are shown for co-ordinates 30,000 to 40,000, in which regions the highest counts of DSBs for rDNA occur.

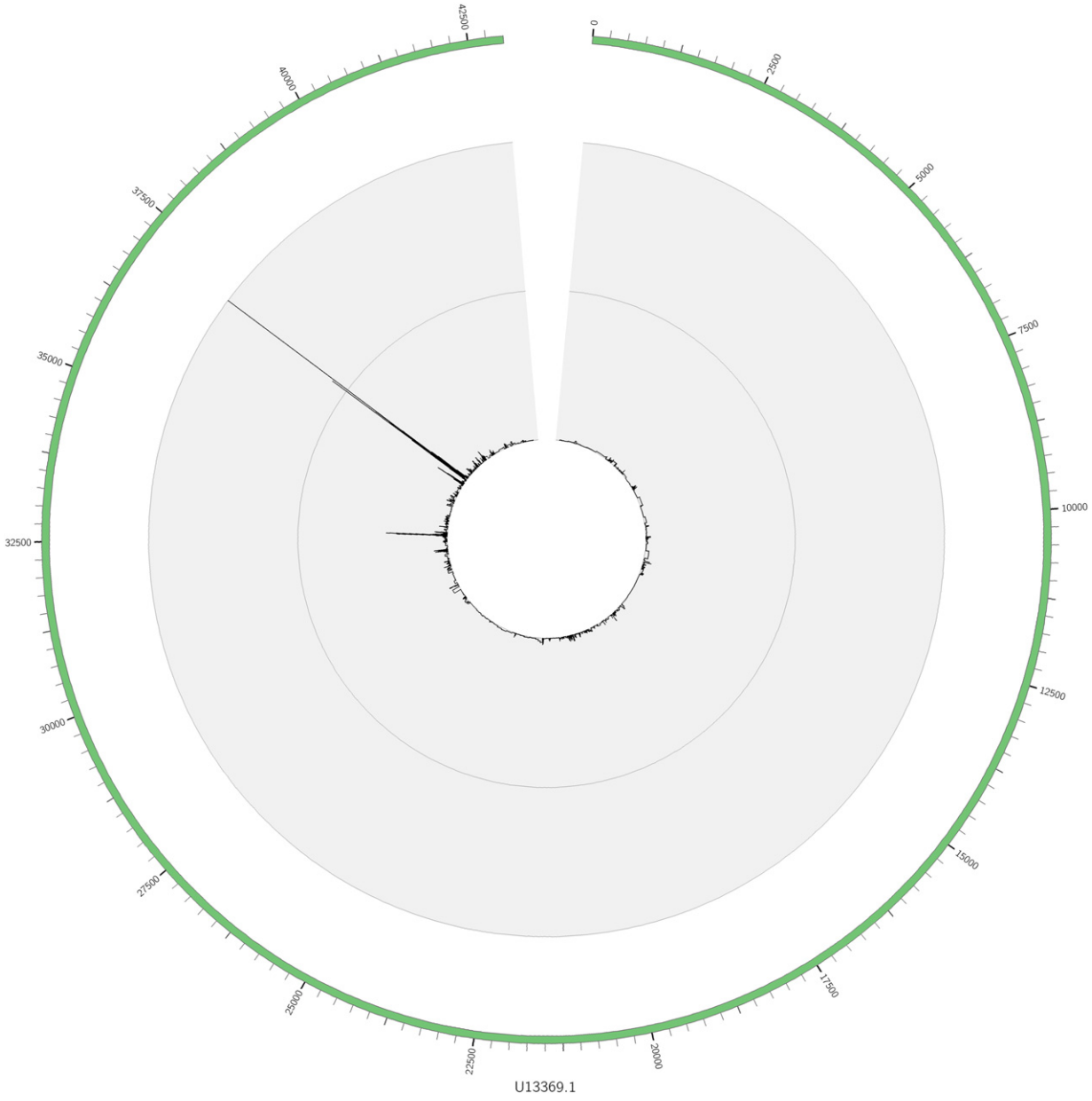

Fig. 3. Circos plot showing counts of DSBs by co-ordinate based on mapping to human rDNA U13369.1.fa.

paired-end sequencing will assist in locating DSB sites regardless of the orientation of the insert for a given library element.

Acknowledgements

This work was supported by NHMRC (Australia) project grant 1108179 and VLSCI research allocation VR0182.

References

[1] B. Blondet, A. Ait-Ikhlef, M. Murawsky, F. Rieger, Transient massive DNA fragmentation in nervous system during the early course of a murine neurodegenerative disease. Neurosci. L. 305 (2001) 202–206.
[2] H. Li, R. Durbin, Fast and accurate long-read alignment with burrows-wheeler transform. Bioinformatics 26 (2010) 589–595.
[3] 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) 2078–2079.
[4] A.R. Quinlan, I.M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26 (2010) 841–842.
[5] P.J. Stephens, C.D. Greenman, B. Fu, F. Yang, G.R. Bignell, L.J. Mudie, E.D. Pleasance, K.W. Lau, D. Beare, L.A. Stebbings, S. McLaren, M.L. Lin, D.J. McBride, I. Varela, S. Nik-Zainal, C. Leroy, M. Jia, A. Menzies, A.P. Butler, J.W. Teague, M.A. Quail, J. Burton, H. Swerdlow, N.P. Carter, L.A. Morsberger, C. Iacobuzio-Donahue, G.A. Follows, A.R. Green, A.M. Flanagan, M.R. Stratton, P.A. Futreal, P.J. Campbell, Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell 144 (2011) 27–40.
[6] 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.
[7] N.A. Tchurikov, O.V. Kretova, D.M. Fedoseeva, V.R. Chechetkin, M.A. Gorbacheva, A.A. Karnaukhov, G.I. Kravatskaya, Y.V. Kravatsky, Hot spots of DNA double-strand breaks and genomic contacts of human rDNA units are involved in epigenetic regulation. J. Mol. Cell Biol. 7 (2015) 366–382.
[8] N.A. Tchurikov, O.V. Kretova, D.M. Fedoseeva, V.R. Chechetkin, M.A. Gorbacheva, A.A. Karnaukhov, G.I. Kravatskaya, Y.V. Kravatsky, Mapping of genomic double-strand breaks by ligation of biotinylated oligonucleotides to forum domains: analysis of the data obtained for human rDNA units. Genomics Data 3 (2015) 15–18.
[9] N.A. Tchurikov, D.V. Yudkin, M.A. Gorbacheva, A.I. Kulemzina, I.V. Grischenko, D.M. Fedoseeva, D.V. Sosin, Y.V. Kravatsky, O.V. Kretova, Hot spots of DNA double-strand breaks in human rDNA units are produced *in vivo*. Sci. Rep. 6 (2016) 25866.