Genomics Data 11 (2017) 43–45

Contents lists available at ScienceDirect

Genomics Data

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

Single nucleotide-level mapping of DNA double-strand breaks in human HEK293T cells

Bernard J. Pope [a], Khalid Mahmood [a], Chol-hee Jung [a], Peter Georgeson [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 17 October 2016
Received in revised form 7 November 2016
Accepted 9 November 2016
Available online 11 November 2016

Keywords:
Double-strand breaks
Fragile sites
Human genome
Forum domains
HEK293T

ABSTRACT

Constitutional biological processes involve the generation of DNA double-strand breaks (DSBs). The production of such breaks and their subsequent resolution are also highly relevant to neurodegenerative diseases and cancer, in which extensive DNA fragmentation has been described Stephens et al. (2011), Blondet et al. (2001). Tchurikov et al. Tchurikov et al. (2011, 2013) have reported previously that frequent sites of DSBs occur in chromosomal domains involved in the co-ordinated expression of genes. This group report that hot spots of DSBs in human HEK293T cells often coincide with H3K4me3 marks, associated with active transcription Kravatsky et al. (2015) and that frequent sites of DNA double-strand breakage are likely to be relevant to cancer genomics Tchurikov et al. (2013, 2016) . Recently, they applied a RAFT (rapid amplification of forum termini) protocol that selects for blunt-ended DSB sites and mapped these to the human genome within defined co-ordinate 'windows'. In this paper, we re-analyse public RAFT data to derive sites of DSBs at the single-nucleotide level across the built genome for human HEK293T cells (https://figshare.com/s/35220b2b79eaaaf64ed8). This refined mapping, combined with accessory ENCODE data tracks and ribosomal DNA-related sequence annotations, will likely be of value for the design of clinically relevant targeted assays such as those for cancer susceptibility, diagnosis, treatment-matching and prognostication.

1. Direct link to deposited data

https://figshare.com/s/35220b2b79eaaaf64ed8

2. Experimental design, materials and methods

2.1. Sequencing data

The FASTQ file for Illumina Genome Analyzer IIx (GAIIx) run accession SRR944107 (single-end reads) 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 origins of these data have been reported previously [12].

Briefly, HEK293T cells were suspended in 1% 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, eluted via *Eco*RI restriction endonuclease digestion and then subjected to *Sau*3AI site adapter ligation and PCR amplification. PCR products were ligated to Illumina adapters, allowing them to be represented in either orientation. Library fragments of ~200–400 bp (insert plus adapter and PCR primer sequences) were band isolated from agarose gels and the purified libraries were sequenced in single-ended fashion using the Illumina Genome Analyzer IIx sequencing platform.

2.2. Data processing

Fig. 1 provides a schematic representation of our bioinformatic analysis pipeline. Specifications are summarised in Table 1. In the first step, we used our custom software `[raft_fastq_2sites_parse.py]` to produce a modified representation of `SRR944107.fastq`, `SRR944107_2sites.fastq`.. This tool is available at https://github.com/djpark1974/raft_hotspots_se. Briefly, it filters reads based on the observation of expected arrangements of adapter sequences, with the stringent requirement that both adapters be evident in a given read. Reads exhibiting evidence of ligation artefacts or insufficient evidence of expected adapter sequences were removed. Accepted reads were processed to trim adapter sequences, and those with library inserts greater than or equal to 25 nucleotides in length were retained and transformed to orient the DSB site at the start.

* Corresponding author at: Victorian Life Sciences Computation Initiative, The University of Melbourne, Australia.
E-mail address: djp@unimelb.edu.au (D.J. Park).

http://dx.doi.org/10.1016/j.gdata.2016.11.007

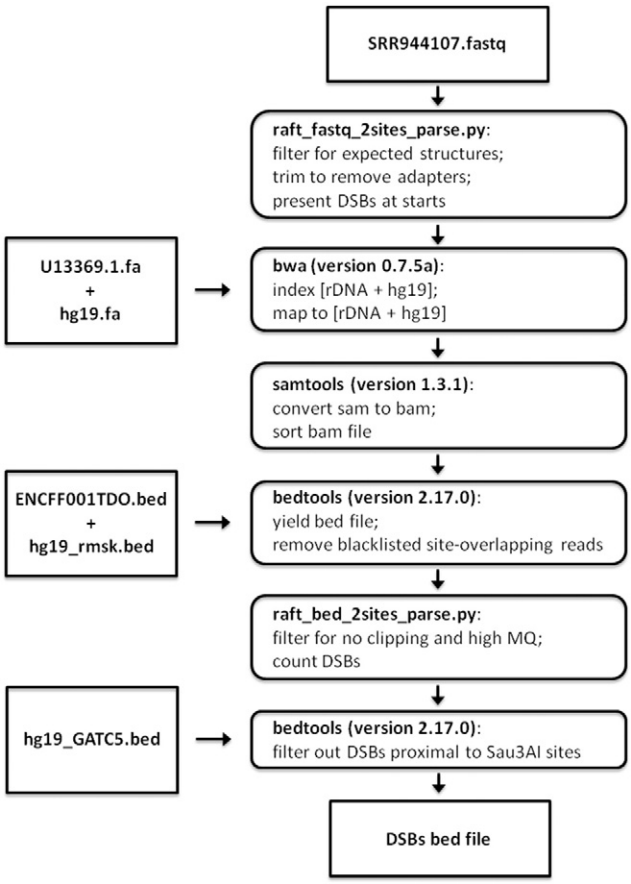

Fig. 1. Schematic illustration of our bioinformatic analysis pipeline to derive counts of DSBs by co-ordinate across genome-build hg19 concatenated with rDNA contiguous sequence U13369.1.

The concatenated sequences of `U13369.1.fa` plus human reference genome build `hg19.fa`,, represented as `rdnahg19.fa`,, were indexed using BWA (version 0.7.5a) [4] using the command:

```
bwa index -a is rdnahg19.fa rdnahg19
```

Reads of the transformed FASTQ file were then mapped to `rdnahg19.fa`, using BWA, thus:

```
bwa mem rdnahg19.fa SRR944107_2sites.fastq \
> SRR944107_rdnahg19.sam
```

Table 1
Materials, data, tools and resources employed in the present study.

Systems and resources	Specifications
Sequencing platform	GAIIx single-read (SRR944107.fastq)
Cell line	Human HEK293T cells
Sequencing library	RAFT-seq
Reference files	hg19.fa; U13369.1.fa; ENCFF001TDO.bed; hg19_rmsk.bed; hg19_GATC5.bed
Data processing software	raft_fastq_2sites_parse.py; bwa (0.7.5a); samtools (1.3.1); bedtools (2.17.0); raft_bed_2sites_parse.py

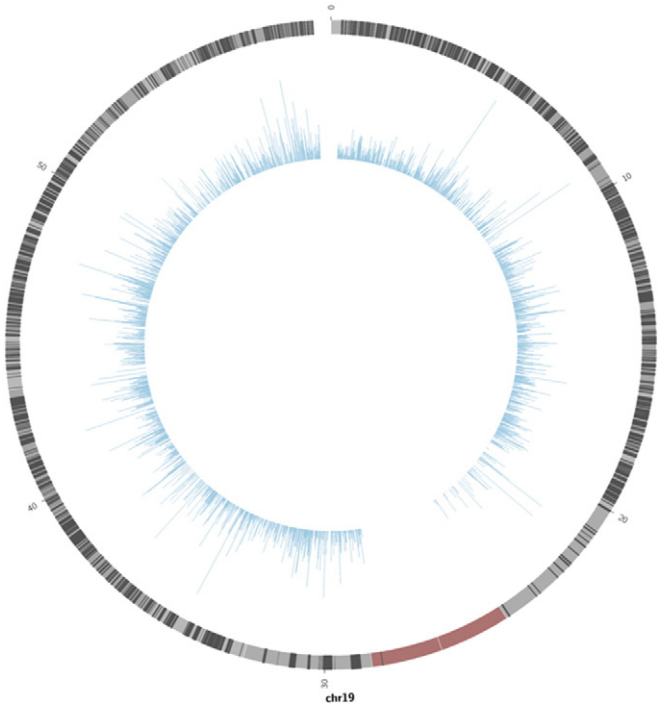

Fig. 2. Circos plot depicting relative DSB counts by co-ordinate for chromosome 19 of human genome-build hg19. The outer numbers indicate co-ordinates in megabases along the chromosome. Black bars indicate gene regions. The red portion indicates centromeric DNA.

SAMtools (version 1.3.1) [5] 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_rdnahg19.sam \
| samtools sort -@ 4 -o SRR944107.rdnahg19.sort.bam
```

BEDtools (version 2.17.0) [7] 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.rdnahg19.sort.bam \
> SRR944107.rdnahg19.bed
```

To reduce false positives resulting from mapping artefacts, we filtered out reads that overlapped with ENCODE project [3] blacklist regions and RepeatMasker-derived repetitive regions as follows (`hg19blacklist.bed` represents a file created by sorting a concatenation of the hg19 co-ordinate-associated files `ENCFF001TDO.bed` and `hg19_rmsk.bed`):

```
bedtools subtract -A -a SRR944107.rdnahg19.bed -b \
hg19blacklist.bed > SRR944107.rdnahg19.blf.bed
```

We then used our custom software [`raft_bed_2sites_parse.py`] (available at https://github.com/djpark1974/raft_hotspots_se) to further filter the data and to count the number of observations of DSBs at co-ordinates in

`-b hg19_GATC5.bed \` (yielding `SRR944107.rdnahg19.blf.counts.bed`). Briefly, this tool assesses the orientation of mapping for each read. Since 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 either orientation were treated as likely to be erroneous if the CIGAR string showed evidence of clipping at either terminus. Additionally, we required reads to exhibit mapping qualities (MQs) of greater than 40 for them to be included in our DSB site counting.

For increased specificity, we removed DSB sites located within 5 base pairs of a *Sau*3AI consensus site (GATC), thus (`hg19_GATC5.bed`was derived via a custom python script):

```
bedtools subtract -A \
-a SRR944107.rdnahg19.blf.counts.bed \
-b hg19_GATC5.bed \

> SRR944107.rdnahg19.blf.sau3ai.counts.bed
```

To enable detailed downstream analyses, we have supplemented the co-ordinate-DSB-count data with annotation derived from the ENCODE project and BLAST alignment scores derived from aligning U13369.1 to the human genome. ENCODE annotations are recorded with the identity and proximity of respective ENCODE elements and BLAST alignment scores provide the highest scoring sequence similarity match for a contiguous sequence spanning a given co-ordinate. Fig. 2 illustrates the frequency of DSBs at single nucleotide resolution sites across the hg19 reference human genome.

3. Discussion

Here, we present the relative frequencies of DSBs across the human reference genome for HEK293T cells, at single nucleotide resolution. Since DNA strand breakage and genomic rearrangements are highly relevant to cancer and other diseases [2,8,14], it is probable that our new data will have utility for the development of clinically important diagnostic tests.

The highest ranking DSB regions reported by Tchurikov et al. [12] for the SRR944107.fastq dataset predominantly relate to regions that would be likely to present problems to short read mapping software, such as satellite sequences. In an attempt to reduce mapping-related artefacts, we have elected to remove regions known to result in low-confidence mapping from our analysis. Top ranking single nucleotide-resolved DSB sites resulting from our analysis relate to regions listed previously as enriched for DSBs, albeit to a lesser extent than reported for numerous low-complexity (and low confidence) sequence regions.

Our data relate to a particular subgroup of DSBs. The RAFT protocol from which our data are derived is theoretically enriched for blunt-ended forum domain termini, previously shown to be associated with transcriptional control [6,10,11]. They will be biased towards termini that occur within a particular range of genomic distances from a *Sau*3AI restriction endonuclease site. Future protocols that make use of multiple restriction endonucleases for cleavage following the initial ligation step, as alternatives to (and as well as) *Sau*3AI, should mitigate this to a large extent. Our data are further biased towards genomic regions that can be mapped unambiguously. We have applied high-stringency thresholding on mapping quality as part of our algorithm and discarded library elements that could not be uniquely assigned to a single genomic location with high confidence and, as such, repetitive genomic elements harbouring DSB sites will not be represented.

It should be noted that the data we present relate to human HEK293T cells. Other cell-types will likely exhibit differences in their RAFT-detectable DSB profiles due to variations in higher-order chromosomal architecture and DNA cleavage-inducing enzyme activity. These differences will be elucidated with the expansion of studies to a range of cell and tissue types.

We have supplemented the profiling of the relative frequency of DSB sites in HEK293T cells with ENCODE-derived annotation [3], including regional information pertaining to important transcription factor binding sites and other marks of gene regulation, regions of DNaseI hypersensitivity and repetitive elements. Further, we provide annotation in the form of sequence similarity scores, derived from BLAST analysis [1], for sites that occur in regions with high similarity to human ribosomal DNA, since such sequences are known to include hot spots for DSBs and present particular mapping challenges due to their representation at high copy number at multiple sites in the genome. This information should assist with the selection of suitable targets for diagnostic test design, allowing the user optionally to avoid sites that present excessive mapping difficulties or to focus on regions associated with particular genomic marks, for example.

The refined characterisation of the propensity for particular types of DSBs, such as those identified by the RAFT procedure, across the human genome will likely allow more efficient assessment of genomic 'scarring' for an individual. This should be highly relevant to clinical management approaches such as risk stratification for particular types of cancer and treatment response prediction. As such, the use of these data has the potential to be beneficial to the reduction of disease associated mortality and morbidity.

Acknowledgements

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

References

[1] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool. J. Mol. Biol. 215 (1990) 403–410.
[2] 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. Lett. 305 (2001) 202–206.
[3] The ENCODE Project Consortium, A User's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol. 9 (2011), e1001046.
[4] H. Li, R. Durbin, Fast and accurate long-read alignment with burrows-wheeler transform. Bioinformatics 26 (2010) 589–595.
[5] 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.
[6] Y.V. Kravatsky, V.R. Chechetkin, N.A. Tchurikov, G.I. Kravatskaya, Genome-wide study of correlations between genomic features and their relationship with the regulation of gene expression. DNA Res. 22 (2015) 109–119.
[7] A.R. Quinlan, I.M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26 (2010) 841–842.
[8] 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.
[10] 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.
[11] 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 (2013), e1003429.
[12] 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. Genom. Data 3 (2015) 15–18.
[14] 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.