Bioinformatics, 37(22), 2021, 4230–4232
doi: 10.1093/bioinformatics/btab366
Advance Access Publication Date: 12 May 2021
Applications Note

Genome analysis

DetectIS: a pipeline to rapidly detect exogenous DNA integration sites using DNA or RNA paired-end sequencing data

Luigi Grassi [1,*], Claire Harris [1], Jie Zhu[2], Colin Hardman [3] and Diane Hatton[1]

[1]Biopharmaceutical Development, BioPharmaceuticals R&D, AstraZeneca, Cambridge, CB21 6GH, UK, [2]Biopharmaceutical Development, BioPharmaceuticals R&D, AstraZeneca, Gaithersburg, MD 20878, USA and [3]Data Science & Artificial Intelligence, BioPharmaceuticals R&D, AstraZeneca, Cambridge, CB21 6GH, UK

*To whom correspondence should be addressed.
Associate Editor: Jan Gorodkin

Received on October 16, 2020; revised on May 7, 2021; editorial decision on May 10, 2021; accepted on May 11, 2021

Abstract

Motivation: Recombinant DNA technology is widely used for different applications in biology, medicine and biotechnology. Viral transduction and plasmid transfection are among the most frequently used techniques to generate recombinant cell lines. Many of these methods result in the random integration of the plasmid into the host genome. Rapid identification of the integration sites is highly desirable in order to characterize these engineered cell lines.

Results: We developed detectIS: a pipeline specifically designed to identify genomic integration sites of exogenous DNA, either a plasmid containing one or more transgenes or a virus. The pipeline is based on a Nextflow workflow combined with a Singularity image containing all the necessary software, ensuring high reproducibility and scalability of the analysis. We tested it on simulated datasets and RNA-seq data from a human sample infected with Hepatitis B virus. Comparisons with other state of the art tools show that our method can identify the integration site in different recombinant cell lines, with accurate results, lower computational demand and shorter execution times.

Availability and implementation: The Nextflow workflow, the Singularity image and a test dataset are available at https://github.com/AstraZeneca/detectIS.

Contact: luigi.grassi@astrazeneca.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Recombinant DNA technology can be used to generate transgenic animals, plants and cell lines, widely used for different applications in biology, medicine and biotechnology (Ghaderi *et al.*, 2012; Khan *et al.*, 2016). Therapeutic proteins with complex post-translational modifications are normally expressed in mammalian cell lines (Walsh, 2018; Zhu and Hatton, 2018). Viral transduction and plasmid transfection are methods largely used to establish recombinant cell lines (Kim and Eberwine, 2010; Lee *et al.*, 2018) and typically result in random integration of the transgene construct into the host genome. The identification of the transgene integration site (IS) is important for the characterization of stable recombinant cell lines and, can reveal regulatory features relevant for transgene expression. It can also detect aberrant transgene–host fusion proteins, potentially caused by the plasmid integrating in the proximity of protein-coding genes. Understanding ISs can identify integration 'hot spots', i.e. genomic sites conferring high expression of the transgene and data from multiple experiments can be used for the design of targeted ISs. Moreover, as the transgene ISs are unique for an individual transfection event, the IS information can be used to design PCR experiments to assess the clonality of a cell line (Sommeregger et al., 2013). Inverse PCR (Liang et al., 2008; Uemura et al., 2014), splinkerette-PCR (Uren et al., 2009) and targeted locus amplification (de Vree *et al.*, 2014) are techniques specifically designed to localize ISs in host genomes. High-throughput sequencing (HTS) experiments have been successfully used to localize a similar biological event: the viral ISs in host genomes (Chen *et al.*, 2019). Moreover, several studies have proved the usefulness of HTS in localizing plasmid ISs in stable cell lines (Brett *et al.*, 2011; Lambirth et al., 2015; Srivastava et al., 2014). Although pipelines have been developed for detecting viral integration sites, some of them are specifically designed for the human genome reference sequence. Moreover, all the tools require the preparation of indexes specific for each host and exogenous DNA element.

We present detectIS, a pipeline to detect the ISs in paired end (PE) HTS experiments (either DNA or RNA sequencing data). It can be directly used with different host and exogenous DNA references,

without the need of creating a specific index. Consequently, it is suitable for different applications, for example detecting ISs of plasmids in stable cell lines, either clones or pools, as well as locating viruses integrated in any host genome. The speed of execution makes the detectIS pipeline well-suited for quickly screening HTS data from panels of different cell lines generated during the cell line development process for therapeutic protein manufacture, enabling the detection of cell lines with undesirable transgene fusion sequences.

2 Materials and methods

DetectIS (Supplementary Fig. S1) consists of three main steps. PE reads are aligned, in single-end mode onto the exogenous sequence reference (i.e. transgene, plasmid or viral sequences). Reads with any overlap with the exogenous reference sequence are subsequently aligned, in single-end mode, to the host genome reference. The alignment is made by using the Minimap2 program (Li, 2018). Finally, a *Perl* script integrates the four alignment results looking for potential ISs. ISs can be identified by split reads—read pairs in which at least one read has a part mapping to the host genome and the remaining part mapping to the plasmid/transgene, and chimeric reads, read pairs in which one of the two reads is mapped to the host genome and the other one to the plasmid/transgene. The pseudocode of the subroutines used by the *Perl* script is reported in Supplementary Figures S2–S9. Final results are provided as a txt file detailing all the potential ISs and the number of supporting split and chimeric read pairs. The same information is also reported in a markdown file that can be converted to a pdf and/or html file. All the steps of the detectIS pipeline are embedded in a Nextflow (Di Tommaso *et al.*, 2017) workflow that, together with the Singularity (Kurtzer *et al.*, 2017) container ensures reproducibility and scalability from a single PC/workstation to high-performance computational (HPC) environments.

3 Usage

In order to use the workflow, the user has to create a configuration file specifying the reference host genome and exogenous sequence references, the directory containing the raw data and the output directory. The analysis can be executed locally or in an HPC environment, in the latter scenario the user also has to specify the cluster executor. A configuration file is provided to analyze a test dataset and can be used as a template for other analyses.

The recipe of the Singularity image with all the necessary software is also supplied. A bash script is also given to analyze a test dataset without Nextflow and can be used as a template for analysis in local environments.

3.1 Comparison with existing tools for structural variant identification

In order to test the functionality of detectIS and the accuracy of its results, we simulated random integrations of a plasmid in a Chinese hamster ovary (CHO) scaffold, exploring different modalities of transgene size, depth of sequencing coverage and read length. We compared the results of detectIS with the ones derived by other tools for viral detection, that are able to use host references different from human. SeekSV (Liang et al., 2017) is a program designed to identify ISs and other structural variants in RNA-seq and DNA-seq experiments and was one of the best performing tools for identifying viral integrations in a recent study (Chen *et al.*, 2019). BatVI (Tennakoon and Sung, 2017) is a sensitive and fast tool used for the detection of viral integrations that, similarly to detectIS, uses a subtractive strategy where raw reads are aligned to the viral reference genomes in the first instance, and the partially mapped reads are then aligned to the host reference genome to detect viral integrations. SurVirus (Rajaby et al., 2021) is a recently published repeat-aware virus integration caller. The detectIS results are among the ones with highest precision and sensitivity in most of the simulated experiments with sequenced read of lengths 250 and 150 bases (Supplementary Figs S10A–F, S11–AF, Supplementary Tables S1–S3). Minimap2 works with read length of 100 bases or higher (Li, 2018) and, for this reason, 100 bases is the lowest read length compatible with detectIS. In this simulated scenario, the tool is less precise and sensitive than SurVirus and SeekSV for sequence coverages of 5× and 10×, but performs similarly at higher coverage (Supplementary Figs S10–GI, S11G–I, Tables Supplementary S1–S3). The execution times of the analyses are similar for detectIS, SurVirus and BatVI and higher for SeekSV in all the simulated experiments (Supplementary Fig. S12). DetectIS has the lowest computational demands with the lowest CPU times in all the simulated experiments (Supplementary Fig. S13). It is also notable that detectIS can be executed without the reference index generation, a time consuming step required by all the other tools (Supplementary Fig. S14). The integration sites detected by all the used tools have an average discrepancy of a few nucleotides in respect to the original sites (Supplementary Fig. S15). In the simulated integrations, plasmid and host had the same orientation 5′→3′ and this feature was captured by all the tools.

We extended the comparison to publicly available RNA-seq experiments of four hepatitis B virus (HBV) positive hepatocellular carcinoma cell lines with verified chimeric viral-human transcripts (Lau et al., 2014). In this analysis, SurVirus terminated with a segmentation fault error in all the four analyzed experiments and produced an empty final result file in three of them. Analogously, BatVI produced a final result file for only one of the four analyzed experiments, for this reason, we could compare only the results generated by detectIS and seekSV. We defined true positives as ISs that supported the chimeric viral-human transcripts verified in the study of Lau *et al.* (2014), with a tolerance of 50 nucleotides (Supplementary Table S4). The two tools gave similar results in term of precision, sensitivity (Supplementary Fig. S16A, Supplementary Table S5) and difference from the real data (Fig. S16B) with a significantly shorter running time for detectIS (Supplementary Fig. S16C and D). This difference in running times can be justified by the fact that the two pipelines are based on different programs and strategies, with seekSV looking for all potential structural variants while detectIS uses a subtractive strategy and is designed to specifically identify variants affecting the exogenous DNA (plasmid/virus). The results presented in this study demonstrate that detectIS is able to identify integration sites in HTS experiments, in a short time without high demands on computational resources. The benchmark analysis indicates that a longer read length improves detectIS precision and sensitivity in experiments made at a lower coverage. The usage of the Minimap2 program for the alignment gives the possibility of running the analysis without any index preparation step and makes the pipeline unique among all the existing programs for viral integration. Due to its versatility, detectIS can be executed to identify viral integration sites in transcriptome or genome sequencing experiments and identify the ISs of plasmids inserted into stable cell lines from HTS experiments routinely made to exclude the presence of variants in transgenic transcripts during clone selection (Harris *et al.*, 2019; Lin et al., 2019).

Financial Support: none declared.

Conflict of Interest: none declared.

References

Brett,B.T. *et al.* (2011) Novel molecular and computational methods improve the accuracy of insertion site analysis in Sleeping Beauty-induced tumors. *PLoS One*, **6**, e24668.

Chen,X. *et al.* (2019) Comprehensive comparative analysis of methods and software for identifying viral integrations. *Brief. Bioinf.*, **20**, 2088–2097.

de Vree,P.J. *et al.* (2014) Targeted sequencing by proximity ligation for comprehensive variant detection and local haplotyping. *Nat. Biotechnol.*, **32**, 1019–1025.

Di Tommaso,P. *et al.* (2017) Nextflow enables reproducible computational workflows. *Nat. Biotechnol.*, **35**, 316–319.

Ghaderi,D. *et al.* (2012) Production platforms for biotherapeutic glycoproteins. Occurrence, impact, and challenges of non-human sialylation. *Biotechnol. Genet. Eng. Rev.*, **28**, 147–175.

Harris,C. *et al.* (2019) Identification and characterization of an IgG sequence variant with an 11 kDa heavy chain C-terminal extension using a combination of mass spectrometry and high-throughput sequencing analysis. *MAbs*, **11**, 1452–1463.

Khan,S. *et al.* (2016) Role of recombinant DNA technology to improve life. *Int. J. Genomics*, **2016**, 2405954.

Kim,T.K. and Eberwine,J.H. (2010) Mammalian cell transfection: the present and the future. *Anal. Bioanal. Chem.*, **397**, 3173–3178.

Kurtzer,G.M. *et al.* (2017) Singularity: scientific containers for mobility of compute. *PLoS One*, **12**, e0177459.

Lambirth,K.C. *et al.* (2015) CONTRAILS: a tool for rapid identification of transgene integration sites in complex, repetitive genomes using low-coverage paired-end sequencing. *Genome Data*, **6**, 175–181.

Lau,C.C. *et al.* (2014) Viral-human chimeric transcript predisposes risk to liver cancer development and progression. *Cancer Cell*, **25**, 335–349.

Lee,J.S. *et al.* (2018) Revealing key determinants of clonal variation in transgene expression in recombinant CHO cells using targeted genome editing. *ACS Synth. Biol.*, **7**, 2867–2878.

Li,H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**, 3094–3100.

Liang,Y. *et al.* (2017) Seeksv: an accurate tool for somatic structural variation and virus integration detection. *Bioinformatics*, **33**, 184–191.

Liang,Z. *et al.* (2008) Identifying and genotyping transgene integration loci. *Transgenic Res.*, **17**, 979–983.

Lin,T.J. *et al.* (2019) Evolution of a comprehensive, orthogonal approach to sequence variant analysis for biotherapeutics. *MAbs*, **11**, 1–12.

Rajaby,R. *et al.* (2021) SurVirus: a repeat-aware virus integration caller. Nucleic Acids Research, **49**, e33 10.1093/nar/gkaa1237PMC: 33444454

Sommeregger,W. *et al.* (2013) Transgene copy number comparison in recombinant mammalian cell lines: critical reflection of quantitative real-time PCR evaluation. *Cytotechnology*, **65**, 811–818.

Srivastava,A. *et al.* (2014) Discovery of transgene insertion sites by high throughput sequencing of mate pair libraries. *BMC Genomics*, **15**, 367.

Tennakoon,C. and Sung,W.K. (2017) BATVI: fast, sensitive and accurate detection of virus integrations. *BMC Bioinformatics*, **18**, 71.

Uemura,S. *et al.* (2014) A simple and highly efficient method to identify the integration site of a transgene in the animal genome. *Neurosci. Res.*, **80**, 91–94.

Uren,A.G. *et al.* (2009) A high-throughput splinkerette-PCR method for the isolation and sequencing of retroviral insertion sites. *Nat. Protoc.*, **4**, 789–798.

Walsh,G. (2018) Biopharmaceutical benchmarks 2018. *Nat. Biotechnol.*, **36**, 1136–1145.

Zhu,J. and Hatton,D. (2018) New mammalian expression systems. *Adv. Biochem. Eng. Biotechnol.*, **165**, 9–50.