Molecular Therapy
Methods & Clinical Development
Original Article

VSeq-Toolkit: Comprehensive Computational Analysis of Viral Vectors in Gene Therapy

Saira Afzal,[1] Raffaele Fronza,[2] and Manfred Schmidt[1,2]

[1]Department of Translational Oncology, German Cancer Research Center (DKFZ), National Center for Tumor Diseases (NCT), Heidelberg, Germany; [2]GeneWerk GmbH, Heidelberg, Germany

Viral vector characterization and analysis are important components for the development of safe gene therapeutic products, elucidating the potential genotoxic and immunogenic effects of vectors and establishing their safety profiles. Here, we present VSeq-Toolkit, which offers varying analysis modes for viral gene therapy data. The first mode determines the undesirable known contaminants and their frequency in viral preparations or other sequencing data. The second mode is designed for the analysis of intra-vector fusion breakpoints and the third mode for unraveling the viral-host fusion events distribution. Analysis modes of our toolkit can be executed independently or together and allow the analysis of multiple viral vectors concurrently. It has been designed and evaluated for the analysis of short read high-throughput sequencing data, including whole-genome or targeted sequencing. VSeq-Toolkit is developed in Perl and Bash programming languages and is available at https://github.com/CompMeth/VSeq-Toolkit.

INTRODUCTION

Advancements in gene therapy and approval of products for retinal dystrophy and lipoprotein lipase deficiency reinforce the promises to treat challenging disorders ranging from hereditary, infectious, metabolic, cardiovascular, and ophthalmologic to various cancer types.[1–6] The use of viruses as carriers of genetic products requires extensive and in-depth understanding of viral vectors starting from viral preparation until clinical employment.[7–10] The viral preparations must be strictly quality assessed[11,12] for exclusion of any viral or cellular contaminants and impurities to exclude any immunogenic effects. Additionally, during the pre-clinical and clinical settings, viral gene therapy should be closely monitored for any insertional mutagenesis effects.[13,14] Analyzing and tracking viral integration events within the host genome helps to estimate the likelihood of viral vector safety in relevance to deregulation of any tumor suppressor or oncogene.[15–17] Certain vectors, for example adeno-associated viruses (AAVs), are susceptible to vector-vector rearranged junction formation.[18–20] Therefore, elucidating vector-vector breakpoint fusions is important in enhancing the understanding of viral vectors.

Here, we present a comprehensive toolkit for the analysis of viral gene therapy data starting from the analysis of viral vector preparations or any contaminant estimations to the pre-clinical and clinical monitoring of vector-vector or vector-host fusions. Various tools are available that deal with the vector-host fusions or integration site (IS) analysis, in the context of viral cancers mainly[21–25] and in gene therapy focusing on PCR-based methods as linear-amplification-mediated (LAM) PCR[26–31] and targeted sequencing.[28] However, here, our aim is to provide an easy-to-use tool suite that can provide multiple analyses, including contaminant distribution within the data, intra-viral vector fusion events that were not previously addressed by available methods, and viral-genome fusion events in whole-genome sequencing (WGS) or targeted sequencing data. In addition, we aimed for the analysis of multiple viral vectors simultaneously within a single sample. The analysis modes of VSeq-Toolkit can be used independently or jointly for reliable and precise characterization and estimation of contaminants and their frequency, vector-vector, and vector-genome fusion distributions. Our method is designed for Illumina short read paired-end (PE) data and can be reliably used for the pre-clinical assessment to clinical monitoring of viral vectors risk and safety profiles.

RESULTS

To evaluate the reliability and accuracy of VSeq-Toolkit (Figure 1), *in silico* datasets generated for each respective analysis mode were analyzed. The 250 bp PE dataset D1 comprising of 6,000 reads was analyzed with contaminant analysis mode, which was able to detect all respective sequences correctly, including 1,550 and 800 contaminant one and two sequences, respectively, 1,750 vector sequences, and 1,900 reads from human genome assembly (hg38) (Figure 2A). Similarly, D2 and D3 *in silico* datasets were analyzed with vector-vector and vector-host fusion analysis modes, respectively. The vector-vector breakpoints of 250 bp PE reads were accurately identified for each vector reference sequence. Vector-host fusion analysis mode also identified all fusion event reads in 250 bp PE data (Figure 2A; Data S1). The datasets comprising of contaminant, vector-vector, and vector-host 150 bp PE reads were also evaluated that showed similarly high precision and recall values. Additionally, D1A, D2A, and D3A datasets with 0.25% error rates were analyzed in a similar manner and showed reliability of toolkit modes (Figure 2A). We

Received 9 March 2020; accepted 25 March 2020;
https://doi.org/10.1016/j.omtm.2020.03.024.

Correspondence: Saira Afzal, Department of Translational Oncology (G100), German Cancer Research Center (DKFZ), National Center for Tumor Diseases (NCT), Im Neuenheimer Feld 581, 69120 Heidelberg, Germany.
E-mail: saira.afzal@nct-heidelberg.de

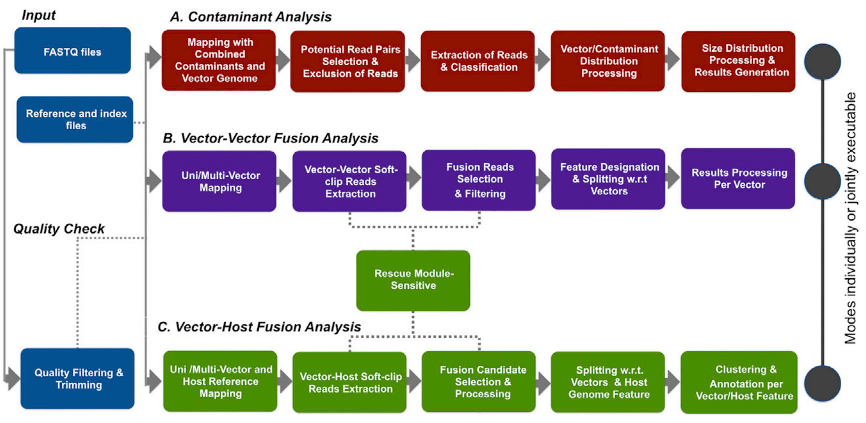
Input
FASTQ files
Reference and index files
Quality Check
Quality Filtering & Trimming
A. Contaminant Analysis
Mapping with Combined Contaminants and Vector Genome
Potential Read Pairs Selection & Exclusion of Reads
Extraction of Reads & Classification
Vector/Contaminant Distribution Processing
Size Distribution Processing & Results Generation
B. Vector-Vector Fusion Analysis
Uni/Multi-Vector Mapping
Vector-Vector Soft-clip Reads Extraction
Fusion Reads Selection & Filtering
Feature Designation & Splitting w.r.t Vectors
Results Processing Per Vector
Rescue Module-Sensitive
C. Vector-Host Fusion Analysis
Uni /Multi-Vector and Host Reference Mapping
Vector-Host Soft-clip Reads Extraction
Fusion Candidate Selection & Processing
Splitting w.r.t. Vectors & Host Genome Feature
Clustering & Annotation per Vector/Host Feature
Modes individually or jointly executable

Figure 1. Schematics of Basic VSeq-Toolkit Workflow

The VSeq-Toolkit is comprised of five modules with three main analysis modes. The input module takes FASTQ paired-end reads, and the pre-processing module performs the quality filtering and trimming. The contaminant analysis mode allows simultaneous detection of contaminants or vector in the sample along with their respective distribution and fragment size statistics. The vector-vector fusion mode provides the analysis of vector rearranged breakpoint events. The vector-genome fusion mode analyzes the distribution of vector integration sites within the host genome.

estimated the percentages of read pairs (contaminant analysis) or fusion events (vector-vector and vector-host) within ± 3 bp of the expected position. In case of contaminant analysis mode (D1), all read-pair positions were detected at the expected position, whereas in case of vector-vector (D2) and vector-host (D3) modes, approximately >76% fusion breakpoint positions were detected at expected positions, 19% within ± 1 bp difference and 3.5%–4.5% in ± 2 bp of expected positions (Figure 2B).

We had additionally analyzed experimental datasets from previously published studies. The S1 experimental dataset,[28] i.e., a control lentiviral vector (LV) sample with three known vector-genome ISs, was evaluated with vector-host fusion mode. Similarly, AAV-based publicly available samples[32] S2 and S3 were analyzed to mainly show the performance of vector-vector fusion analysis mode. In the case of S1 sample of about 37 million reads, the three expected vector-genome integration events were accurately identified with significantly high sequence count numbers (Data S2). In the case of LV sample, as expected, no significant vector-vector fusions were detected, and only one breakpoint was detected, which could be most likely an experimental artifact sequence. On the other hand, the AAV-based samples S2 and S3, comprising of 5,561,416 and 2,540,471 reads, respectively, were analyzed to show the performance of vector-vector fusion analysis mode. The significant numbers of vector-vector breakpoints were detected in both samples, mainly at the inverted terminal repeat (ITR) regions, as shown in Figure 3. Here, the breakpoints were evaluated by considering zero and one as a cutoff value for sequence count.

We have also evaluated the computational time efficiency of each mode of VSeq-Toolkit on *in silico* datasets DS1.1, DS2.1, and DS3.1 for contaminant, vector-vector, and vector-host fusion analysis modes, respectively (Figure 4A). Each of these datasets consists of 100k reads (250 bp PE). Additionally, we have also evaluated the time consumption for analyzing experimental datasets S1 (for vector-host fusions), and S2 and S3 (for vector-vector fusions). The S1 dataset of about 37 million reads was analyzed for vector-host fusions in 71 min. The S2 and S3 (about 5.5 and 2.5 million reads) were analyzed for vector-vector fusions within less than 12 and 6 min, respectively (Figure 4B). The S1 dataset was 100 bp PE, whereas S2 and S3 were 250 bp PE.

Additionally, we measured the time consumption per each main module of the toolkit for vector-host (S1 dataset) and vector-vector (S2 and S3 datasets) analysis modes. In case of vector-host analysis for S1, the highest time is consumed by mapping step, followed by

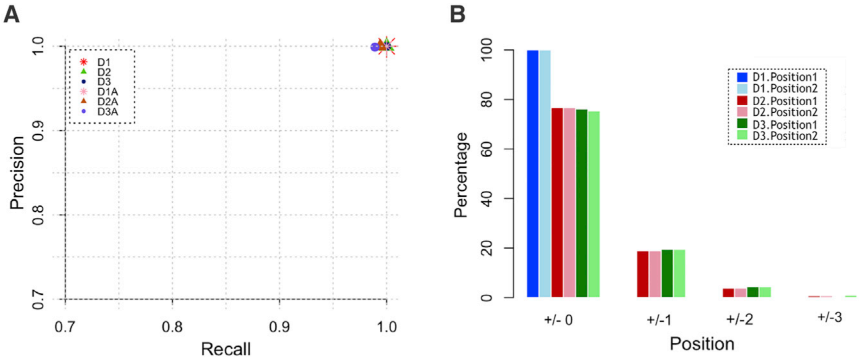

Figure 2. Performance and Reliability Evaluation of VSeq-Toolkit Modes on *In Silico* Datasets

(A) Statistical measures, recall and precision estimation on *in silico* datasets are depicted here. Each mode of VSeq-Toolkit; contaminant analysis, vector-vector fusion, and vector-host fusion is evaluated with respective *in silico* datasets without errors D1, D2, and D3 and with 0.25% error rate D1A, D2A, and D3A. (B) The histogram represents the percentages of detected read pair positions (in contaminant analysis mode on D1) or fusion positions (in vector-vector analysis mode on D2 and vector-host analysis mode on D3) within a ± 3 bp range of the expected positions. For each dataset, both the positions were evaluated for each read pair (in contaminant analysis mode) or each fusion event region (in vector-vector and vector-host analysis modes).

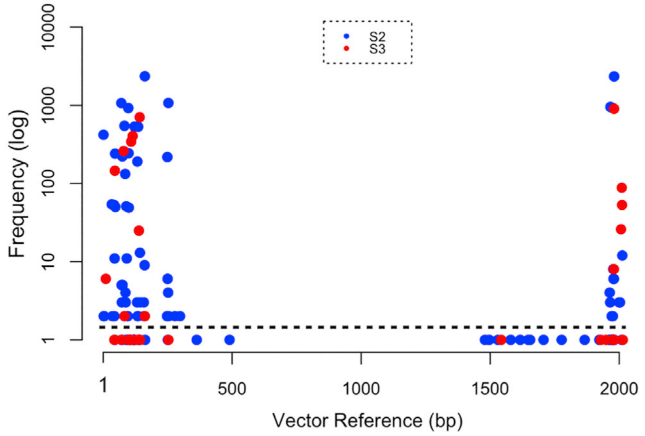

Figure 3. Distribution of Vector-Vector Breakpoint Junctions

The vector-vector fusion breakpoints analyzed in the two experimental datasets S2 and S3 are presented here. The x axis represents the vector reference, and the y axis represents the frequency of breakpoint positions in a logarithmic scale. The dashed line represents the cutoff value of one for fusion sequence count. The breakpoints are accumulated mainly in the ITR regions.

quality check (Figure 4C). In vector-vector analysis the major time consumption was in quality check module followed by mapping (Figure 4D).

DISCUSSION

We have described VSeq-Toolkit, which combines the functionalities to cover a range of viral gene therapy data analysis requirements at one platform in a comprehensive and computationally efficient manner. It provides specific modes for analyzing contaminants distribution, vector-vector rearranged junctions, and vector-genome breakpoint distribution in high-throughput sequencing data. The toolkit allows characterization of the contaminants in viral preparations along with determination of their respective frequencies and fragment sizes. The vector-vector fusion analysis mode allows characterization of vector breakpoint events. Currently, to the best of our knowledge, no tools are available so far that are specifically designed and evaluated for vector breakpoint distribution profiling for gene therapy data.

The viral-genome fusion mode of the toolkit helps to unravel the distribution of ISs of vectors within the respective genomes. Although a wide range of tools are available for IS analysis specifically suitable for LAM-PCR[26–31] and targeted or WGS with focus on viral cancers[21–25] and gene therapy.[28] Here, we provide the analysis of different aspects of gene therapy data with a single toolkit for WGS or targeted sequencing data. VSeq-Toolkit provides an added advantage that both vector-vector fusions and vector-genome integration events can be investigated together. At the first stage, vector-vector breakpoint reads are analyzed, followed by vector-genome fusion events analysis. The exclusion of vector-vector fusions at the first step increases the specificity of subsequent vector-genome distribution analysis, along with providing a useful and detailed vector breakpoint profile. In contrast to the available viral IS tools, the vector-fusion mode of our toolkit investigates in depth not only the genomic part of the fusion read, but also the vector region that provides a more transparent view of vector positions and their distribution. Furthermore, as compared to the other available methods,[21–25,28] the toolkit modes allow analysis of multiple viral references within a sample concurrently. This feature has broader implications for analysis of viral cancers as well.

We have evaluated each analysis mode of toolkit with *in silico* datasets and have shown the reliability and accuracy of our method. Additionally, we depicted the performance of analysis modes on LV and AAV experimental datasets. The toolkit is easy to use and allows multiple adjustable parameters that can be tailored according to analysis requirements. Additionally, the specificity and sensitivity levels for selection of reads are user adjustable. A range of other configurable parameters are also provided, including minimum vector or genome region length, maximum number of non-mapped or overlapped bases between the fusion events, the minimum identity percentage of each fusion event region, etc. In the case of vector-host IS analysis, the clustering of reads based on genomic positions can be performed within the required range. The results are reported in an easy-to-read format for the end-user with detailed information by each respective mode. The toolkit can be employed with any host genome or viral vector sequence as reference.

In-depth analysis of various dimensions of gene therapy data is an important step to assess safety and efficacy of viral-vector based gene therapy. VSeq-Toolkit provides a compact workflow to analyze contaminant distribution, viral vector breakpoint profiles, and viral-genome ISs in a reliable and highly time-efficient way. It additionally has broader implications in analyzing next-generation sequencing data for the presence of viral and non-viral contaminants. Moreover, it allows investigation of data from insertional mutagenesis screens, viral cancers, and infectious diseases.

MATERIALS AND METHODS

Toolkit

The input module of VSeq-Toolkit accepts FASTQ PE data that is processed initially with the quality-control module for quality filtering of reads and trimming of sequencing adapters with skewer.[33] In the next step, data is processed, based on user requirements, through one or all of the main analysis modes: contaminant analysis, vector-vector fusion analysis, and vector-host fusion analysis (Figure 1). The contaminant analysis mode characterizes the undesired known contaminants and estimates their abundance. The filtered and trimmed dataset is first mapped with the reference genome by the Burrows-Wheeler Aligner (BWA) MEM aligner.[34] Multiple potential contaminants can be provided as a concatenated single reference file, if desired, along with vector or host genome sequences. In the next steps, the sequence alignment/map (SAM)[35] file is processed for duplicates removal with Samtools, and correctly mapped read pairs are selected for further processing. Subsequently, the distribution of each contaminant sequence, vector, and host is analyzed,

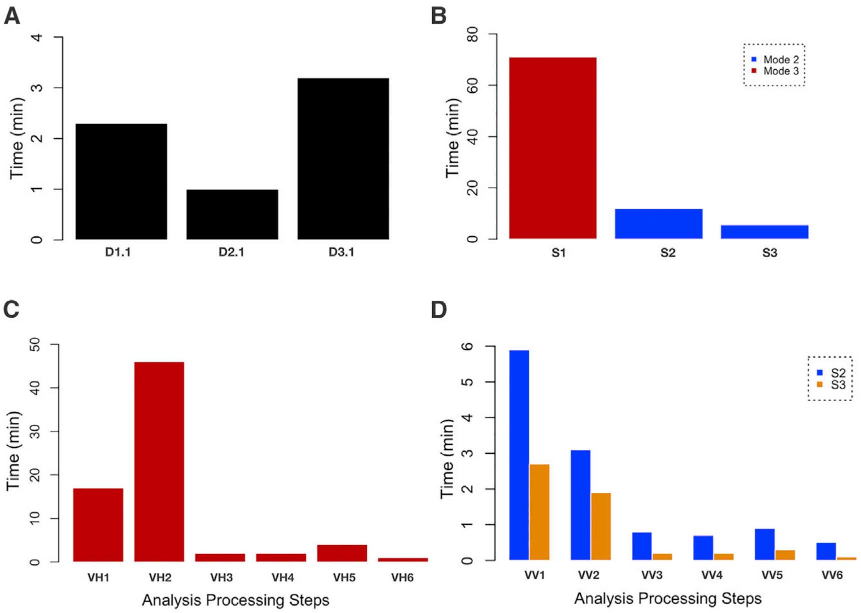

Figure 4. Computational Performance of VSeq-Toolkit Modes

(A) Representation of run time for contaminant analysis, vector-vector, and vector-host fusion analysis modes of VSeq-Toolkit on respective *in silico* datasets D1.1, D2.1, and D3.1, each comprising of 100k reads. (B) Time consumption for the S1 experimental dataset, comprising of approximately 37 million reads analyzed by vector-host fusion (mode 3), and for the S2 and S3 experimental datasets, consisting of about 5.5 and 2.5 million reads, respectively, with vector-vector fusion analysis (mode 2). (C) Time taken by different processing steps of vector-host analysis mode for the S1 dataset (VH1, quality control; VH2, vector-genome mapping; VH3, extraction of vector-genome candidates; VH4, selection and processing; VH5, parameter based estimation and filtering; VH6, clustering, annotation, and results generation). (D) Time consumed per each main module of vector-vector fusion analysis mode for S2 and S3 datasets (VV1, quality control; VV2, vector mapping; VH3, extraction of potential vector-vector reads; VV4, selection, filtering, and processing; VV5, parameter estimation and feature designation; and VV6, results generation).

and frequency values are calculated. The fragment size distribution for the contaminants and vector or host genome references is estimated from the SAM file for the final processed subset of reads. The read number are estimated within six categories of fragment sizes, ranging from less than 50 to greater than 1,000. At the final step, the contaminant sequences are removed from the actual raw data files, and cleaned FASTQ files are generated for further investigation either by external tools or VSeq-Toolkit.

The second mode of vector-vector fusion analysis is designed to detect the rearranged vector sequence breakpoints. The first step of this analysis mode performs mapping of reads with single or multiple combined vector references. In the subsequent steps, the soft-clip reads are extracted and potential vector-vector fusion reads within the same vector reference are selected. These are filtered further based on the user configurable stringency levels. The candidate reads are then processed in the next steps to determine individual vector-vector breakpoint positions that undergo a next round of final selection based on user-adjustable parameters for minimum length, identity, and maximum overlap or distance among fused regions. Finally, the detailed information about each fusion breakpoint is generated in the result files.

In the vector-host fusion analysis mode, the first module performs mapping of data with the combined host and vector reference sequences. The soft-clipped read pairs with alternative mapping with either vector or genome are extracted, and genome-genome rearranged reads are excluded at this stage. This subset of reads is subsequently processed for selection of correctly mapped read pairs, followed by a number of steps leading to the determination of the exact fusion junction positions for vector and genome regions. Similar to vector-vector fusion mode, the vector-genome final fusion reads are also filtered based on user-customizable parameters. Each of the viral vector-genome files is then processed for position-based clustering followed by annotation with custom scripts and bedtools.[36] Finally, the resulted IS output per each vector type is concatenated and provided as a result file.

In sensitive mode for vector-vector and vector-host fusion analysis, the sequencing reads that do not show alternative mapping are processed with a more sensitive remapping module to rescue candidate fusions. In all three modes, each sequence is tagged with a unique and non-unique label, based on either a read pair (first mode) or the regions of a read that contribute to afusion event (second and third modes) is mapped uniquely or non-uniquely in the respective reference.

Datasets

We generated *in silico* datasets for analyzing the performance of all three modes of VSeq-Toolkit. For contaminant analysis mode, the *in silico* dataset was designed by extracting random sequences from two reference sequences regarded as contaminants, a vector reference and hg38. A random region of 500 bp length was extracted from each of the aforementioned references. The 250 bp PE dataset comprising of a total of 6,000 reads (D1) was designed that includes 1,550 and 800 read pairs for first and second contaminant sequences, respectively. In addition, the dataset contains 1,750 vector reads and 1,900 reads from hg38. For vector-vector fusion analysis, the *in silico* dataset was created by randomly extracting the sequences of 500 bp from the two different vectors to create 250 bp PE end data, then we excluded 50 bp region from these reads and inserted a known vector region to simulate vector-vector fusion reads. This dataset is comprised of 215 reads (D2) from each of the first and second vectors. For vector-host fusion analysis, 19,550 random sequences were extracted from the hg38 genome and 250 PE reads were created. For each vector type, different regions of the vector were selected and introduced within the reads, resulting in final dataset of 19,550 reads (D3). Furthermore, these datasets were simulated with 0.25% errors to generate D1A, D2A, and D3A datasets, respectively, for each mode. The vector-vector and vector-host datasets were generated with 150 bp PE read lengths as well. In addition, we also generated in a similar manner DS1.1 (comprising reads from two contaminants and one vector sequence, excluding hg38), DS2.1, and DS3.1 for contaminant analysis, vector-vector, and vector-host fusion modes, respectively. Each of these datasets was comprised of 100k reads to evaluate the time-efficiency for each VSeq-Toolkit analysis mode.

The analysis was performed on *in silico* datasets with respective modules, and basic statistical measures, including precision and recall, were estimated. An event is considered true positive if it lies within ± 3 bp of the expected position, false negative if any event is not detected by the respective module of the toolkit, and false positive if other than the expected positions are reported. In the case of vector-vector and vector-genome fusion analysis, the statistical-measures estimation takes into account both positions of the fusion event, i.e., the position of both vector regions that makes one fusion event in the case of vector-vector analysis, and in the case of the vector-genome, the fusion positions of vector as well as genomic regions are taken into account.

Additionally, we have analyzed the experimental datasets to depict the reliability and performance of vector-vector and vector-host fusion modes of the toolkit. We have analyzed a control sample of LV-transduced HeLa cells with three integration events comprising of approximately 37 million PE reads reported in a previous study.[28] In this study, we have regarded this sample as S1. Furthermore, we have analyzed two publicly available datasets (SRR7085814 and SRR7085812)[32] with AAV generated by targeted enrichment sequencing (https://www.ncbi.nlm.nih.gov/Traces/study/?acc=SRP144114). Here, in this work, we refer to these samples as S2 and S3, respectively.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.omtm.2020.03.024.

AUTHOR CONTRIBUTIONS

S.A. and M.S. conceived the project. S.A. designed and developed the methods, performed the analysis, and wrote the manuscript. M.S. and R.F. revised the manuscript. M.S. provided the resources.

CONFLICTS OF INTEREST

M.S. is co-founder and CEO of GeneWerk GmbH.

ACKNOWLEDGMENTS

We are grateful to our colleagues at the Translational Oncology group for helpful feedback and discussions.

REFERENCES

1. Biffi, A., Montini, E., Lorioli, L., Cesani, M., Fumagalli, F., Plati, T., Baldoli, C., Martino, S., Calabria, A., Canale, S., et al. (2013). Lentiviral hematopoietic stem cell gene therapy benefits metachromatic leukodystrophy. Science *341*, 1233158.
2. Linette, G.P., Stadtmauer, E.A., Maus, M.V., Rapoport, A.P., Levine, B.L., Emery, L., Litzky, L., Bagg, A., Carreno, B.M., Cimino, P.J., et al. (2013). Cardiovascular toxicity and titin cross-reactivity of affinity-enhanced T cells in myeloma and melanoma. Blood *122*, 863–871.
3. MacLaren, R.E., Groppe, M., Barnard, A.R., Cottriall, C.L., Tolmachova, T., Seymour, L., Clark, K.R., During, M.J., Cremers, F.P., Black, G.C., et al. (2014). Retinal gene therapy in patients with choroideremia: initial findings from a phase 1/2 clinical trial. Lancet *383*, 1129–1137.
4. Naldini, L. (2015). Gene therapy returns to centre stage. Nature *526*, 351–360.
5. Kassner, U., Hollstein, T., Grenkowitz, T., Wühle-Demuth, M., Salewsky, B., Demuth, I., Dippel, M., and Steinhagen-Thiessen, E. (2018). Gene Therapy in Lipoprotein Lipase Deficiency: Case Report on the First Patient Treated with Alipogene Tiparvovec Under Daily Practice Conditions. Hum. Gene Ther. *29*, 520–527.
6. Ludwig, P.E., Freeman, S.C., and Janot, A.C. (2019). Novel stem cell and gene therapy in diabetic retinopathy, age related macular degeneration, and retinitis pigmentosa. Int. J. Retina Vitreous *5*, 7.
7. Connolly, J.B. (2002). Lentiviruses in gene therapy clinical research. Gene Ther. *9*, 1730–1734.
8. Zhang, X., and Godbey, W.T. (2006). Viral vectors for gene delivery in tissue engineering. Adv. Drug Deliv. Rev. *58*, 515–534.
9. Lukashev, A.N., and Zamyatnin, A.A., Jr. (2016). Viral vectors for gene therapy: Current state and clinical perspectives. Biochemistry (Mosc.) *81*, 700–708.
10. Finer, M., and Glorioso, J. (2017). A brief account of viral vectors and their promise for gene therapy. Gene Ther. *24*, 1–2.
11. Merten, O.-W., and Wright, J.F. (2016). Towards routine manufacturing of gene therapy drugs. Mol. Ther. Methods Clin. Dev. *3*, 16021.
12. van der Loo, J.C.M., and Wright, J.F. (2016). Progress and challenges in viral vector manufacturing. Hum. Mol. Genet. *25* (R1), R42–R52.
13. Howe, S.J., Mansour, M.R., Schwarzwaelder, K., Bartholomae, C., Hubank, M., Kempski, H., Brugman, M.H., Pike-Overzet, K., Chatters, S.J., de Ridder, D., et al. (2008). Insertional mutagenesis combined with acquired somatic mutations causes leukemogenesis following gene therapy of SCID-X1 patients. J. Clin. Invest. *118*, 3143–3150.
14. Wu, C., and Dunbar, C.E. (2011). Stem cell gene therapy: the risks of insertional mutagenesis and approaches to minimize genotoxicity. Front. Med. *5*, 356–371.
15. Schmidt, M., Schwarzwaelder, K., Bartholomae, C., Zaoui, K., Ball, C., Pilz, I., Braun, S., Glimm, H., and von Kalle, C. (2007). High-resolution insertion-site analysis by linear amplification-mediated PCR (LAM-PCR). Nat. Methods *4*, 1051–1057.
16. Hacein-Bey-Abina, S., Garrigue, A., Wang, G.P., Soulier, J., Lim, A., Morillon, E., Clappier, E., Caccavelli, L., Delabesse, E., Beldjord, K., et al. (2008). Insertional oncogenesis in 4 patients after retrovirus-mediated gene therapy of SCID-X1. J. Clin. Invest. *118*, 3132–3142.
17. Negre, O., Bartholomae, C., Beuzard, Y., Cavazzana, M., Christiansen, L., Courne, C., Deichmann, A., Denaro, M., de Dreuzy, E., Finer, M., et al. (2015). Preclinical evaluation of efficacy and safety of an improved lentiviral vector for the treatment of β-thalassemia and sickle cell disease. Curr. Gene Ther. *15*, 64–81.
18. Nowrouzi, A., Penaud-Budloo, M., Kaeppel, C., Appelt, U., Le Guiner, C., Moullier, P., von Kalle, C., Snyder, R.O., and Schmidt, M. (2012). Integration frequency and intermolecular recombination of rAAV vectors in non-human primate skeletal muscle and liver. Mol. Ther. *20*, 1177–1186.
19. Colella, P., Ronzitti, G., and Mingozzi, F. (2017). Emerging Issues in AAV-Mediated *In Vivo* Gene Therapy. Mol. Ther. Methods Clin. Dev. *8*, 87–104.
20. Hanlon, K.S., Kleinstiver, B.P., Garcia, S.P., Zaborowski, M.P., Volak, A., Spirig, S.E., Muller, A., Sousa, A.A., Tsai, S.Q., Bengtsson, N.E., et al. (2019). High levels of AAV vector integration into CRISPR-induced DNA breaks. Nat. Commun. *10*, 4439.

21. Chen, Y., Yao, H., Thompson, E.J., Tannir, N.M., Weinstein, J.N., and Su, X. (2013). VirusSeq: software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue. Bioinformatics *29*, 266–267.

22. Wang, Q., Jia, P., and Zhao, Z. (2013). VirusFinder: software for efficient and accurate detection of viruses and their integration sites in host genomes through next generation sequencing data. PLoS ONE *8*, e64465.

23. Ho, D.W.H., Sze, K.M.F., and Ng, I.O.L. (2015). Virus-Clip: a fast and memory-efficient viral integration site detection tool at single-base resolution with annotation capability. Oncotarget *6*, 20959–20963.

24. Forster, M., Szymczak, S., Ellinghaus, D., Hemmrich, G., Rühlemann, M., Kraemer, L., Mucha, S., Wienbrandt, L., Stanulla, M., and Franke, A.; UFO Sequencing Consortium within I-BFM Study Group (2015). Vy-PER: eliminating false positive detection of virus integration events in next generation sequencing data. Sci. Rep. *5*, 11534.

25. Shieh, F.S., Jongeneel, P., Steffen, J.D., Lin, S., Jain, S., Song, W., and Su, Y.H. (2017). ChimericSeq: An open-source, user-friendly interface for analyzing NGS data to identify and characterize viral-host chimeric sequences. PLoS ONE *12*, e0182843.

26. Arens, A., Appelt, J.-U., Bartholomae, C.C., Gabriel, R., Paruzynski, A., Gustafson, D., Cartier, N., Aubourg, P., Deichmann, A., Glimm, H., et al. (2012). Bioinformatic clonality analysis of next-generation sequencing-derived viral vector integration sites. Hum. Gene Ther. Methods *23*, 111–118.

27. Hocum, J.D., Battrell, L.R., Maynard, R., Adair, J.E., Beard, B.C., Rawlings, D.J., Kiem, H.P., Miller, D.G., and Trobridge, G.D. (2015). VISA–Vector Integration Site Analysis server: a web-based server to rapidly identify retroviral integration sites from next-generation sequencing. BMC Bioinformatics *16*, 212.

28. Afzal, S., Wilkening, S., von Kalle, C., Schmidt, M., and Fronza, R. (2017). GENE-IS: Time-Efficient and Accurate Analysis of Viral Integration Events in Large-Scale Gene Therapy Data. Mol. Ther. Nucleic Acids *6*, 133–139.

29. Spinozzi, G., Calabria, A., Brasca, S., Beretta, S., Merelli, I., Milanesi, L., and Montini, E. (2017). VISPA2: a scalable pipeline for high-throughput identification and annotation of vector integration sites. BMC Bioinformatics *18*, 520.

30. Berry, C.C., Nobles, C., Six, E., Wu, Y., Malani, N., Sherman, E., Dryga, A., Everett, J.K., Male, F., Bailey, A., et al. (2016). INSPIIRED: Quantification and Visualization Tools for Analyzing Integration Site Distributions. Mol. Ther. Methods Clin. Dev. *4*, 17–26.

31. Calabria, A., Beretta, S., Merelli, I., Spinozzi, G., Brasca, S., Pirola, Y., Benedicenti, F., Tenderini, E., Bonizzoni, P., Milanesi, L., and Montini, E. (2020). γ-TRIS: a graph-algorithm for comprehensive identification of vector genomic insertion sites. Bioinformatics *36*, 1622–1624.

32. Senís, E., Mosteiro, L., Wilkening, S., Wiedtke, E., Nowrouzi, A., Afzal, S., Fronza, R., Landerer, H., Abad, M., Niopek, D., et al. (2018). AAVvector-mediated in vivo reprogramming into pluripotency. Nat. Commun. *9*, 2651.

33. Jiang, H., Lei, R., Ding, S.-W., and Zhu, S. (2014). Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. BMC Bioinformatics *15*, 182.

34. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

35. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

36. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.