Genomics

Molecular Cancer Research

Next-Generation Sequencing Analysis and Algorithms for PDX and CDX Models

Garima Khandelwal[1], María Romina Girotti[2], Christopher Smowton[3], Sam Taylor[4], Christopher Wirth[3], Marek Dynowski[3], Kristopher K. Frese[5], Ged Brady[5], Caroline Dive[5], Richard Marais[2], and Crispin Miller[1]

Abstract

Patient-derived xenograft (PDX) and circulating tumor cell–derived explant (CDX) models are powerful methods for the study of human disease. In cancer research, these methods have been applied to multiple questions, including the study of metastatic progression, genetic evolution, and therapeutic drug responses. As PDX and CDX models can recapitulate the highly heterogeneous characteristics of a patient tumor, as well as their response to chemotherapy, there is considerable interest in combining them with next-generation sequencing to monitor the genomic, transcriptional, and epigenetic changes that accompany oncogenesis. When used for this purpose, their reliability is highly dependent on being able to accurately distinguish between sequencing reads that originate from the host, and those that arise from the xenograft itself. Here, we demonstrate that failure to correctly identify contaminating host reads when analyzing DNA- and RNA-sequencing (DNA-Seq and RNA-Seq) data from PDX and CDX models is a major confounding factor that can lead to incorrect mutation calls and a failure to identify canonical mutation signatures associated with tumorigenicity. In addition, a highly sensitive algorithm and open source software tool for identifying and removing contaminating host sequences is described. Importantly, when applied to PDX and CDX models of melanoma, these data demonstrate its utility as a sensitive and selective tool for the correction of PDX- and CDX-derived whole-exome and RNA-Seq data.

Implications: This study describes a sensitive method to identify contaminating host reads in xenograft and explant DNA- and RNA-Seq data and is applicable to other forms of deep sequencing. *Mol Cancer Res; 15(8); 1012–6. ©2017 AACR.*

Introduction

Human xenograft models have been widely used to study cancer. They provide an excellent tool with which to investigate the dynamics of oncogenesis, tumor heterogeneity, evolution, and responses to therapy (1–8). This has led to considerable interest in combining them with next-generation sequencing (NGS). This is challenging because downstream analyses are highly dependent on the quality and purity of the samples (9), leading to poor mutation calling accuracy and poor estimates of gene expression. Although efforts can be made to mitigate these effects experimentally, high levels of infiltrating stromal cells often render this impractical. Consequently, levels of contamination as high as 73% have been observed in pancreatic cancer patient-derived xenograft (PDX) models (10), and data are often variable (11). Instead, studies have typically addressed read-heterogeneity *in silico* (9, 12). Although the precise filtering strategy differs between studies, these studies all compare reads with both the mouse and human genomes and then eliminate those that match strongly to the mouse genome.

Despite the importance of reliable read filtering, only one method, Xenome, is implemented and readily available as a software tool (13). It is a computationally efficient approach that works by identifying 25-mer matches between the experimental data and the two candidate genomes, and using these to partition the data into host, graft, and ambiguous sets.

Here, we describe a new algorithm for deconvolving host and graft reads. Unlike Xenome, our algorithm makes use of full-length alignments and their scores and can use values extracted from the CIGAR string or mapping quality scores when alignment scores are unavailable. With paired end data, in which two reads are generated for each DNA or RNA fragment, corresponding to its 5′ and 3′ ends, the algorithm resolves conflicts at the individual read level, not the fragment level, allowing more data to be retained. These approaches allow a weak but significant match to one organism to be ignored in favor of a stronger match to the other. Together, these enhance its discriminatory power. We demonstrate its utility for the analysis of human melanoma CDX models. The algorithm is freely available and released as an open source tool at https://github.com/CRUKMI-ComputationalBiology/bamcmp.git.

[1]RNA Biology Group, Cancer Research UK Manchester Institute, The University of Manchester, Manchester, United Kingdom. [2]Molecular Oncology Group, Cancer Research UK Manchester Institute, The University of Manchester, Manchester, United Kingdom. [3]Scientific Computing Team, Cancer Research UK Manchester Institute, The University of Manchester, Manchester, United Kingdom. [4]Computational Biology Support Team, Cancer Research UK Manchester Institute, The University of Manchester, Manchester, United Kingdom. [5]Clinical and Experimental Pharmacology Group, Cancer Research UK Manchester Institute, The University of Manchester, Manchester, United Kingdom.

Note: Supplementary data for this article are available at Molecular Cancer Research Online (http://mcr.aacrjournals.org/).

Corresponding Author: Crispin Miller, Cancer Research UK Manchester Institute, The University of Manchester, Wilmslow Road, Manchester M204BX, United Kingdom. Phone: 4416-1446-3176; Fax: 4416-1446-3109; E-mail: crispin.miller@cruk.manchester.ac.uk

doi: 10.1158/1541-7786.MCR-16-0431

AACR

Materials and Methods

DNA/RNA from xenografts is always contaminated, and although an assay has been published (10) to quantify the proportion of human/mouse DNA in the samples from pancreatic cancer, a generalized method is still lacking. Various studies have reported the need to preprocess xenograft data before performing downstream analysis (9, 13).

This method was developed to address the issues involving the analysis of both DNA/RNA xenograft data with a high accuracy. The model was designed so as to be generally applicable to any type of genomic data and also to a subset of common aligners. The method is based on filtering the host reads from the graft reads after aligning the reads to both host and graft genome using preexisting alignment methods, such as Burrows–Wheeler Aligner (BWA; ref. 14), Bowtie2 (15) for DNA-Seq or MapSplice2.0 (16), Tophat2 (17), STAR (18), and others, for RNA sequencing (RNA-Seq) data.

Full experimental details for the $BRAF^{V600E}$-mutated cutaneous melanoma sample, patient information, ethics approval, and animal procedures are available in ref. 3.

As the alignment to both the host and graft is performed using the same aligner, the alignment scores from the aligner can be easily utilized to differentiate the origin of the reads. The reads are filtered on the basis of any of the four different parameters described below, ordered on the basis of stringency (parameter names are in parentheses). In each case, reads are assigned to the genome with the highest score. Consequently, no explicit thresholds are required:

(i) Alignment scores if generated by the software (as).
(ii) CIGAR string values along with NM and MD tags discerned by the aligner (match).
(iii) Mapping Quality (MAPQ) scores of the alignments (mapq).
(iv) Remove everything that matches the host genome (balwayswins).

The reads are categorized into human only, mouse only, and both. The latter category is further categorized into align better to human or align better to mouse after filtering. The reads that align only to human as well as those that align better to human (from the both categories) are merged and returned as human reads; those that remain are assigned to mouse. The method can be used as a standalone application for filtering the contaminated reads or incorporated in the pipelines of routine NGS analysis. It has been implemented in C++ utilizing the htslib library from SAMtools (19).

Filtering process

(i) Align the fastq files to both human and mouse genomes.
(ii) Filter the mouse reads on any of the four filtering parameters.
(iii) Downstream processing as applicable (mutation calling/read count generation/peak calling).

Usage

bamcmp -n -1 ABC_human.bam -2 ABC_mouse.bam -a ABC_humanOnly.bam -A ABC_humanBetter.bam -b ABC_mouseOnly.bam -B ABC_mouseBetter.bam –C ABC_humanLoss.bam -D ABC_mouseLoss.bam -s [as/match/mapq/balwayswins].

All analyses for this study were performed with default parameters for MapSplice2 (version 2.1.6), BWA-mem (version 0.7.11), Picard (version 1.96), GATK (version 3.3), Samtools (version1.3.1), and Mutect (version 1.1.7). Output files from Xenome required minor additional processing to format them correctly for subsequent use by BWA and MapSplice; results from Xenome were calculated from the graft reads only.

Results and Discussion

The data utilized in this study were derived from a cutaneous melanoma (20) patient (patient 10) with a $BRAF^{V600E}$ mutation (3). The patient presented with primary melanoma on the back and bilateral axillary nodal metastasis. A PDX was derived from the bilateral axillary nodal metastasis. The patient relapsed after 3 months with liver, spleen, and lymph node metastases. A CDX (CDXF1) was established from the patient's circulating tumor cells taken at the time of relapse and grown in subsequent passage (CDXF2) that developed macrometastases in the liver, lymph nodes, kidneys, lungs, brain, and distant subcutaneous tissue (3). Whole-exome sequencing (WES) and RNA-Seq were used to profile the lymph node tumor (tumor/primary tumor), PDX, CDXF1, CDXF2, CDXF2 liver metastasis, and CDXF2 lymph node metastasis. WES was also performed for patient whole blood (germline) and mouse kidney.

Here and throughout, all data were processed through the same pipeline (Supplementary Fig. S1), with summary statistics computed in R (21) and Bioconductor (22). WES data were first aligned to human (hg19) and mouse (mm10) genomes separately using BWA (14) with default parameters. Although 99.81% human germline (i.e., never in mouse) reads aligned to the human genome, 42.68% of these mapped also to mouse; similar patterns were also observed for the mouse germline data (99.66%; 40.08%). Similar proportions of cross-species matches were also observed for the mouse xenograft material (Fig. 1A). Together, these data illustrate how a naïve filtering strategy that simply discards reads that map significantly to the mouse genome will be driven largely by orthology between human and mouse and will thus discard substantial proportions of the data. We therefore sought to develop a filtering strategy better able to distinguish between host and graft reads.

WES data were processed using the default GATK framework (23) with mutations called using Mutect (24). Following filtering, using our new algorithm, a minimum of 99.5% of human and mouse germline reads were correctly assigned to the right organism, while at most 0.20% reads could not be reliably mapped to either genome (Fig. 1A). Similar improvements were observed for the xenograft material.

We next asked what effect the software had on mutation calling. Somatic mutations were called relative to the human germline control using Mutect. Without filtering, data were highly variable (631–8,465 SNVs/sample), and concordance poor, despite the fact that all samples were derived from the same patient (Fig. 1B). The number of single-nucleotide variants (SNV) predicted for each sample was also correlated with the level of host read contamination ($r = 0.98$) in the xenograft samples (Fig. 1C). Consistency increased dramatically after filtering (Fig. 1A, B, D, and E). Importantly, this was achieved with minimal effect on sensitivity: only one SNV called in the human primary tumor was lost, and no false positives were obtained when the data were passed through the filtering pipeline (Fig. 1F). On performing the same analysis using Xenome, fewer SNVs were detected; 4 SNVs were lost after filtering and an additional 6 false positives were obtained in the primary tumor (Fig. 1G). We also calculated the

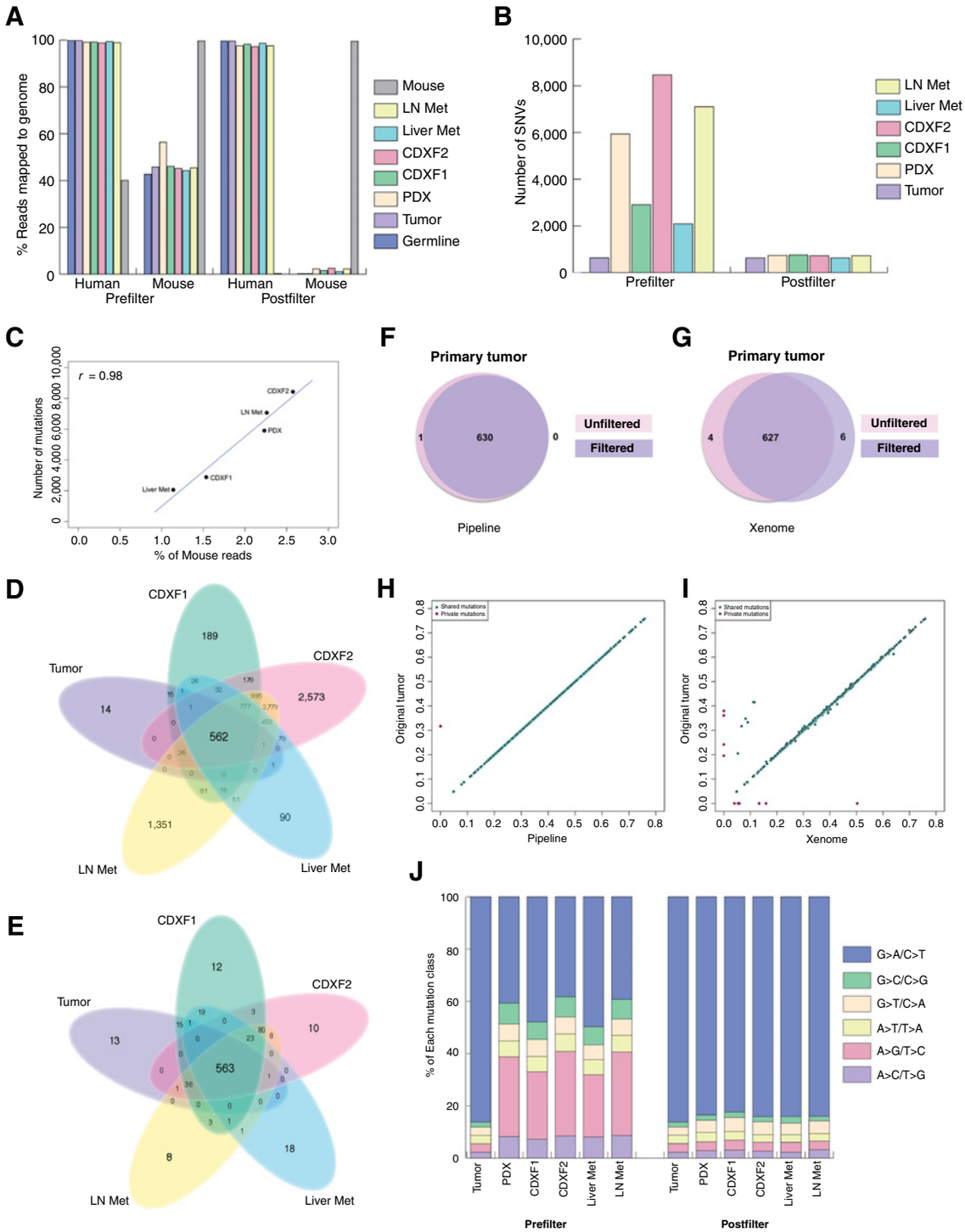

Figure 1.
Filtering reads of mouse origin improves sensitivity and selectivity of mutation calling from CDX models. **A,** Proportion of reads mapping to human and mouse genomes before and after filtering. Mouse, mouse germline sequenced from kidney; LN Met, lymph node metastasis; Liver Met, liver metastasis; CDXF1, CDXF2, circulating tumor cell–derived xenografts; tumor, patient primary tumor; germline, patient whole blood. **B,** Number of SNVs called relative to human germline sequence, before and after filtering. **C,** Number of SNVs called increases linearly with the number of mouse reads detected. **D,** Correspondence in SNVs before filtering. **E,** Correspondence in SNVs after filtering. **F,** Filtering patient primary tumor against mouse removes only one SNV erroneously, and does not lead to others being detected. **G,** As **F**, but filtering using Xenome. **H,** Comparison of VAF before and after filtering for the primary tumor data. **I,** As **H**, but filtering using Xenome. **J,** The canonical C>T transition signature of UV damage is only detectable with correct read processing.

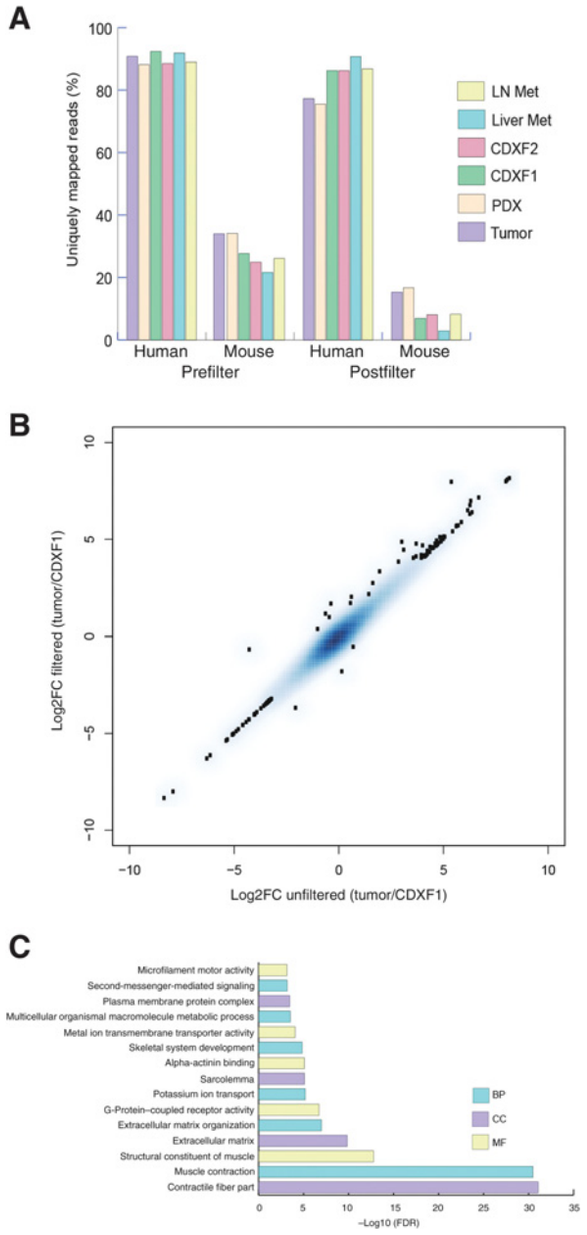

Figure 2.

Filtering removes mouse reads from RNA-Seq data without systematically disrupting expression levels. **A,** Substantial reduction in cross-species mappings following filtering of RNA-Seq data**. B,** High correspondence in fold changes (FC) between human primary and CDX model before and after filtering. Loci no longer detected following filtering have been removed from the figure. **C,** Overrepresentation analysis of loci no longer detected in xenograft data following filtering (BP, biological process; CC, cellular component; MF, molecular function).

variant allele frequency (VAF) using primary tumor before and after filtering. As in the primary tumor data, if no reads are removed, optimal performance would result in no changes to the VAF. This analysis revealed higher agreement before and after filtering using our algorithm (Fig. 1H), than with Xenome (Fig. 1I). Similar analysis of CDX data reveals a similar trend, as expected (Supplementary Fig. S2).

UV-related melanoma is strongly associated with a UV mutation signature comprising a disproportionate number of G>A/C>T transitions (25). Although detected in the primary tumor, this signature was not evident in the xenograft samples prior to filtering. After filtering, it emerged strongly (Fig. 1J).

RNA-Seq data from the same study were then aligned using MapSplice2.0 (16) and filtered using values extracted from the CIGAR string to provide mapping scores. As with the WES data, cross-species mappings were substantially reduced following filtering (Fig. 2A), with levels of mouse contamination concordant to those of the WES data, but at an overall higher level (∼15%). To investigate the effect of filtering on expression changes, we calculated fold changes between the human primary and the mouse CDX model CDXF1 and compared those with the fold changes calculated after filtering. Fold changes for majority of the loci remained consistent, with 17 protein-coding genes differing more than 2-fold between filtered and unfiltered sets (Fig. 2B). When mouse filtering was applied to the human tumor data, 202 protein-coding genes were removed due to sequence homology. A total of 1,394 protein-coding genes exhibited greater than 4-fold difference between the unfiltered CDX and the filtered CDX data (values were computed for the mean of CDXF1 and CDXF2). Over-enrichment analysis of this set using gProfiler (26) found significant enrichment for genes associated with the extracellular matrix (Fig. 2C), indicating that the reads filtered from the dataset are of mouse stromal cell origin. Broadly, similar results were obtained with Xenome (Supplementary Table S1), although a small portion of reads that mapped better to the mouse genome remained, even after filtering. Twenty protein-coding genes exhibited more than 2-fold change between filtered and unfiltered sets (Supplementary Fig. S3), and a similar number of protein-coding genes (1,405) displayed greater than 4-fold difference between the unfiltered CDX and the filtered CDX data. However, 988 protein-coding genes were absent from the filtered tumor dataset, versus 202 with our algorithm, again confirming the improved selectivity of the bamcmp-based pipeline.

Conclusions

In conclusion, we present a sensitive and selective tool for identifying contaminating host reads in deep sequencing data from xenograft and explant models. Although the results we present here focus on WES and RNA-Seq data, the approach is equally applicable to other deep sequencing analyses including ChIP-seq and WGS.

Disclosure of Potential Conflicts of Interest

R. Marais has ownership interest (including patents) in The Institute of Cancer Research. No potential conflicts of interest were disclosed by the other authors.

Authors' Contributions

Conception and design: G. Khandelwal, R. Marais, C. Miller
Development of methodology: G. Khandelwal, C. Smowton, C. Miller
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): M.R. Girotti, K.K. Frese, C. Dive, R. Marais
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): G. Khandelwal, G. Brady

Writing, review, and/or revision of the manuscript: G. Khandelwal, K.K. Frese, G. Brady, C. Dive, C. Miller
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): G. Khandelwal, S. Taylor, C. Wirth, M. Dynowski
Study supervision: C. Miller
Other (programming): M. Dynowski

Acknowledgments

The authors would like to thank Nathalie Dhomen for assistance on this project.

Grant Support

This work was funded by Cancer Research UK C5759/A20971 (to all authors).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received November 29, 2016; revised February 14, 2017; accepted April 20, 2017; published OnlineFirst April 25, 2017.

References

1. Morton CL, Houghton PJ. Establishment of human tumor xenografts in immunodeficient mice. Nat Protoc 2007;2:247–50.
2. Hodgkinson CL, Morrow CJ, Li Y, Metcalf RL, Rothwell DG, Trapani F, et al. Tumorigenicity and genetic profiling of circulating tumor cells in small-cell lung cancer. Nat Med 2014;20:897–903.
3. Girotti MR, Gremel G, Lee R, Galvani E, Rothwell D, Viros A, et al. Application of sequencing, liquid biopsies, and patient-derived xenografts for personalized medicine in melanoma. Cancer Discov 2016;6: 286–99.
4. Fidler IJ. Rationale and methods for the use of nude mice to study the biology and therapy of human cancer metastasis. Cancer Metastasis Rev 1986;5:29–49.
5. Tentler JJ, Tan AC, Weekes CD, Jimeno A, Leong S, Pitts TM, et al. Patient-derived tumour xenografts as models for oncology drug development. Nat Rev Clin Oncol 2012;9:338–50.
6. DeRose YS, Wang G, Lin Y-C, Bernard PS, Buys SS, Ebbert MTW, et al. Tumor grafts derived from women with breast cancer authentically reflect tumor pathology, growth, metastasis and disease outcomes. Nat Med 2011;17:1514–20.
7. Daniel VC, Marchionni L, Hierman JS, Rhodes JT, Devereux WL, Rudin CM, et al. A primary xenograft model of small-cell lung cancer reveals irreversible changes in gene expression imposed by culture *in vitro*. Cancer Res 2009;69:3364–73.
8. Day C-P, Merlino G, Van Dyke T. Preclinical mouse cancer models: a maze of opportunities and challenges. Cell 2015;163:39–53.
9. Rossello FJ, Tothill RW, Britt K, Marini KD, Falzon J, Thomas DM, et al. Next-generation sequence analysis of cancer xenograft models. PLoS One 2013;8:e74432.
10. Lin M-T, Tseng L-H, Kamiyama H, Kamiyama M, Lim P, Hidalgo M, et al. Quantifying the relative amount of mouse and human DNA in cancer xenografts using species-specific variation in gene length. Biotechniques 2010;48:211–8.
11. Pathak S, Nemeth MA, Multani AS. Human tumor xenografts in nude mice are not always of human origin. Cancer 1998;83:1891–3.
12. Tso K-Y, Lee SD, Lo K-W, Yip KY. Are special read alignment strategies necessary and cost-effective when handling sequencing reads from patient-derived tumor xenografts? BMC Genomics 2014;15:1172.
13. Conway T, Wazny J, Bromage A, Tymms M, Sooraj D, Williams ED, et al. Xenome—a tool for classifying reads from xenograft samples. Bioinformatics 2012;28:i172–8.
14. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 2009;25:1754–60.
15. Ben Langmead, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012;9:357–9.
16. Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, et al. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res 2010;38:e178–8.
17. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 2013;14:R36.
18. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013;29:15–21.
19. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics 2009;25:2078–9.
20. Shannan B, Perego M, Somasundaram R, Herlyn M. Heterogeneity in melanoma. In: Kaufman HL, Mehnert JM, editors. Melanoma. Cham, Switzerland: Springer International Publishing; 2016. p. 1–15.
21. R Development Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2015. Available from: https://www.R-project.org/.
22. Gentleman RC, Carey VJ, Bates DM, Ben Bolstad, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 2004;5:R80.
23. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: a mapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;20:1297–303.
24. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol 2013;31:213–9.
25. Miller JH. Mutagenic specificity of ultraviolet light. J Mol Biol 1985;182:45–65.
26. Reimand J, Arak T, Adler P, Kolberg L, Reisberg S, Peterson H, et al. g:Profiler—a web server for functional interpretation of gene lists (2016 update). Nucleic Acids Res 2016;44:W83–9.