Journal of Genetics and Genomics xxx (2018) 1–6

Contents lists available at ScienceDirect

Journal of Genetics and Genomics

Journal homepage: www.journals.elsevier.com/journal-of-genetics-and-genomics/

Review

A comparison of next-generation sequencing analysis methods for cancer xenograft samples

Wentao Dai [a, b, d], Jixiang Liu [a, b, d], Quanxue Li [a, c], Wei Liu [a, b, d], Yi-Xue Li [a, b, c, d, **], Yuan-Yuan Li [a, b, d, *]

[a] Shanghai Center for Bioinformation Technology, Shanghai 201203, China
[b] Shanghai Engineering Research Center of Pharmaceutical Translation & Shanghai Industrial Technology Institute, Shanghai 201203, China
[c] School of Biotechnology, East China University of Science and Technology, Shanghai 200237, China
[d] Shanghai Industrial Technology Institute, Shanghai 201203, China

ARTICLE INFO

Article history:
Received 8 January 2018
Received in revised form
15 June 2018
Accepted 9 July 2018
Available online xxx

Keywords:
Patient-derived xenograft
Next-generation sequencing
Host contamination control
Alignment

ABSTRACT

The application of next-generation sequencing (NGS) technology in cancer is influenced by the quality and purity of tissue samples. This issue is especially critical for patient-derived xenograft (PDX) models, which have proven to be by far the best preclinical tool for investigating human tumor biology, because the sensitivity and specificity of NGS analysis in xenograft samples would be compromised by the contamination of mouse DNA and RNA. This definitely affects downstream analyses by causing inaccurate mutation calling and gene expression estimates. The reliability of NGS data analysis for cancer xenograft samples is therefore highly dependent on whether the sequencing reads derived from the xenograft could be distinguished from those originated from the host. That is, each sequence read needs to be accurately assigned to its original species. Here, we review currently available methodologies in this field, including Xenome, Disambiguate, bamcmp and pdxBlacklist, and provide guidelines for users.

1. Introduction

The use of preclinical models is essential in translational cancer research, especially when identifying biomarkers and developing therapeutic agents (Venditti et al., 1984; Teicher, 2013; Pauli et al., 2017). Patient-derived xenograft (PDX) models, involving grafting fresh tumor tissues into immunodeficient mice (Tentler et al., 2012; Gao et al., 2015; Pauli et al., 2017), recapitulate complex tumor heterogeneity and drug responses observed in the patients more faithfully than traditional cell line-derived xenograft (CDX) models (Fidler, 1986; Morton and Houghton, 2007; DeRose et al., 2011a; Julien et al., 2012; Tentler et al., 2012; Hodgkinson et al., 2014; Alizadeh et al., 2015; Aparicio et al., 2015; Bernardo et al., 2015; Cassidy et al., 2015; Day et al., 2015; Gao et al., 2015; Nunes et al., 2015; Girotti et al., 2016; Wang et al., 2017). PDX models have proven to be by far the best preclinical tool for investigating the dynamics of oncogenesis, tumor heterogeneity, evolution, and responses to therapy (Fidler, 1986; Morton and Houghton, 2007; DeRose et al., 2011b; Calles et al., 2013; Siolas and Hannon, 2013; Hidalgo et al., 2014; Hodgkinson et al., 2014; Day et al., 2015; Girotti et al., 2016; Ledford, 2016). This has naturally led to considerable interest in applying next-generation sequencing (NGS) technology to PDX models, by which the genomic, transcriptomic and epigenetic profiles during oncogenesis could be monitored (Rossello et al., 2013; Li et al., 2014; Girotti et al., 2016). However, the sensitivity and specificity of NGS analysis tend to be compromised by the contamination of mouse DNA and RNA (Lin et al., 2010; Rossello et al., 2013; Khandelwal et al., 2017), which would inevitably affect the downstream analyses by causing inaccurate mutation calling and gene expression estimates (Tso et al., 2014; Li et al., 2015; Khandelwal et al., 2017). Thus, each sequence read needs to be accurately assigned to its original species.

In order to address this issue, quite a few of algorithms aiming to disambiguate the host and tumor xenograft sequences have been designed (Conway et al., 2012; Ahdesmäki et al., 2016; Khandelwal et al., 2017; Salm et al., 2017; Callari et al., 2018), and the reliability

* Corresponding author. Shanghai Center for Bioinformation Technology, Shanghai 201203, China.
** Corresponding author. Shanghai Center for Bioinformation Technology, Shanghai 201203, China.
E-mail addresses: yxli@scbit.org (Y.-X. Li), yyli@scbit.org (Y.-Y. Li).

https://doi.org/10.1016/j.jgg.2018.07.001

of NGS data analysis for cancer xenograft samples is therefore highly dependent on the performance of the algorithms. Here, we review currently available methodologies in this field, including Xenome, Disambiguate, bamcmp and pdxBlacklist, and provide guidelines for users.

2. Characteristics of cancer xenograft sample sequencing

Xenograft models are important for biomedical research, such as oncology and immunology (Cvetkovich et al., 1992; Cook and Tyor, 2006; Han et al., 2013; Buckingham et al., 2015; Cusinato et al., 2016). When the sequence information of xenograft samples is needed, tumor samples are first collected from the animal models (e.g., immunodeficient mice in this case), and then sequenced (Bruna et al., 2016; Khandelwal et al., 2017; Pauli et al., 2017). In this way, the xenografts are unavoidably contaminated by the host cells (DeRose et al., 2011a; Moro et al., 2012; Pearson et al., 2016; Stewart et al., 2017). Although increasing attention is being paid to this issue, it is far from adequate, which is mainly attributed to the assumption that if a sufficiently careful dissection of tumor tissue is taken, the level of host contamination is low enough to be ignored (Cvetkovich et al., 1992; Cook and Tyor, 2006; Conway et al., 2012; Han et al., 2013; Rossello et al., 2013; Hu et al., 2017). It has been more and more clear that a considerable amount of the host (mouse) DNA/RNA will commingle with the graft (human) DNA/RNA since at least part of the xenograft stroma is originated from the host (Conway et al., 2012; Rossello et al., 2013; Hidalgo et al., 2014; Gao et al., 2015; Bruna et al., 2016; Pearson et al., 2016). As Conway et al. (2012) concluded in their literature, when the host contamination takes up to 10% of overall sequencing data, for a certain gene, it may still be the case that its host homologue accounts for most or even all of the expression of this gene.

Compared with other contamination issues, for example, host contamination in sequencing data of parasites and infectious microbes (Cook and Tyor, 2006; Han et al., 2013), host contamination in cancer xenograft samples has several unique characteristics. First, both hosts and grafts are mammals, so their genomes are quite similar (Hidalgo et al., 2014; Aparicio et al., 2015; Day et al., 2015). Second, the genome sizes of both hosts and grafts are large, thereby demanding intensive computational resources when performing disambiguation (Cvetkovich et al., 1992; Cook and Tyor, 2006; Li et al., 2015; Hu et al., 2017; Khandelwal et al., 2017). Third, the heterogeneity of tumor samples is far more pronounced than that of other commonly analyzed biosamples (Ni et al., 2013; Alizadeh et al., 2015; Ling et al., 2015; Martincorena et al., 2017; Wang et al., 2017; Wei et al., 2017). The last but not the least, considering that xenograft models are used for clinical purposes, data reliability is crucial, and distinguishing the sequencing reads derived from the xenograft from those originated from the host cells is the first essential step toward the final goal (Tso et al., 2014; Aparicio et al., 2015; Girotti et al., 2016; Khandelwal et al., 2017; Wang et al., 2017). All these points make it much more challenging than previously thought to accurately disambiguate the host and tumor xenograft sequences. Fortunately, there has been growing interest in developing NGS data analysis methods tailored for cancer xenograft samples.

3. Analysis methods for sequencing data from cancer xenograft samples

In recent years, a series of algorithms and tools aiming to separate the components in mixed human-mouse samples have been designed, including Xenome (Conway et al., 2012), Disambiguate (Ahdesmäki et al., 2016), bamcmp (Khandelwal et al., 2017) and pdxBlacklist (Salm et al., 2017). This ensures efficient recovery of relevant sequencing data for more accurate variant calling and gene expression quantification (Ahdesmäki et al., 2016). It is noted that the alignment algorithms, such as Bowtie2 (Langmead and Salzberg, 2012), BWA (Li and Durbin, 2009), SOAP2 (Li et al., 2009) and MAQ (Li et al., 2008) for DNA-Seq data, and MapSplice2 (Wang et al., 2010), TopHat2 (Kim et al., 2013), STAR (Dobin et al., 2013) and HISAT (Kim et al., 2015) for RNA-Seq data, are the basis of the above analysis methods.

In this section, we first summarize the main strategy and framework of these methods, and then describe their technical details.

3.1. Main strategy and framework

Generally speaking, Xenome (Conway et al., 2012), Disambiguate (Ahdesmäki et al., 2016), bamcmp (Khandelwal et al., 2017) and pdxBlacklist (Salm et al., 2017) share similar strategy and framework, which mainly involve alignment and disambiguation. First, the raw sequencing reads from cancer xenograft samples are aligned to graft and host genomes, respectively; subsequently, the alignment results including alignment files (i.e., *.bam) and scores are used for the following disambiguation process (Conway et al., 2012; Ahdesmäki et al., 2016; Khandelwal et al., 2017; Salm et al., 2017). Additionally, all these processes and methods could be applied to both DNA-Seq and RNA-Seq analyses, where BWA and TopHat are the most popular alignment tools corresponding to DNA-Seq and RNA-Seq, respectively. Specifically, hg19 and mm10 are used as the reference genomes of human and mouse, respectively, which could be updated by *.fa file (Ahdesmäki et al., 2016; Khandelwal et al., 2017; Salm et al., 2017).

It is apparent that the key point of the pipeline is to identify and remove the host-originated sequence information in an efficient and affordable way. According to the design of this point, currently available methods could be roughly classified into three types: 1) multi-alignment strategy, 2) unique classification model, and 3) preset-blacklist strategy. The first one, multi-alignment strategy, involves at least three rounds of alignment including graft alignment, host alignment, and graft-specific re-alignment, and therefore is extremely computational resource and time consuming (Rossello et al., 2013). Since there are no mature softwares corresponding to this strategy, we focus on the other two types of strategies in the following part.

3.1.1. Unique classification model

For the first step, all sequencing reads from cancer xenograft samples are aligned to host and graft genomes respectively with current mainstream alignment algorithms. Then, the aligned reads are classified and screened according to the alignment scores and files, which is the key part of the whole process. Finally, the chosen xenograft-derived reads are applied to downstream processing including mutation calling, read count generation, and peak calling. The Xenome algorithm developed by Conway et al. (2012) pioneered this strategy, and some other approaches were subsequently designed in subtly varied ways (Ahdesmaki et al., 2016; Khandelwal et al., 2017), eventually making this strategy into a mainstream methodology in this field.

Xenome adopts TopHat-based and k-mer-based methods for alignment; the aligned reads are partitioned into four classes: host, graft, both and neither. The first two classes, host and graft, are directly input into subsequent analyses, while the other two classes, both and neither, could be re-aligned to decide their usability or be directly filtered out (Conway et al., 2012).

The follow-up algorithm, Disambiguate, developed by Ahdesmaki et al. (2016), has the following two features: 1) it combines not only alignment scores but also name sorting to

Please cite this article in press as: Dai, W., et al., A comparison of next-generation sequencing analysis methods for cancer xenograft samples, Journal of Genetics and Genomics (2018), https://doi.org/10.1016/j.jgg.2018.07.001