EXPERIMENTAL CELL RESEARCH 322 (2014) 12–20

Available online at www.sciencedirect.com

ScienceDirect

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

Review Article

Library preparation methods for next-generation sequencing: Tone down the bias

Erwin L. van Dijk[a,*], *Yan Jaszczyszyn*[b], *Claude Thermes*[a]

[a]*Centre de Génétique Moléculaire – CNRS, Avenue de la Terrasse, 91198 Gif sur Yvette, France*
[b]*Plateforme Intégrée IMAGIF – CNRS, Avenue de la Terrasse, 91198 Gif sur Yvette, France*

ARTICLE INFORMATION

Article Chronology:
Received 25 October 2013
Received in revised form
7 January 2014
Accepted 8 January 2014
Available online 15 January 2014

Keywords:
Next-generation sequencing
Library preparation
Bias
DNA-seq
RNA-seq

ABSTRACT

Next-generation sequencing (NGS) has caused a revolution in biology. NGS requires the preparation of libraries in which (fragments of) DNA or RNA molecules are fused with adapters followed by PCR amplification and sequencing. It is evident that robust library preparation methods that produce a representative, non-biased source of nucleic acid material from the genome under investigation are of crucial importance. Nevertheless, it has become clear that NGS libraries for all types of applications contain biases that compromise the quality of NGS datasets and can lead to their erroneous interpretation. A detailed knowledge of the nature of these biases will be essential for a careful interpretation of NGS data on the one hand and will help to find ways to improve library quality or to develop bioinformatics tools to compensate for the bias on the other hand. In this review we discuss the literature on bias in the most common NGS library preparation protocols, both for DNA sequencing (DNA-seq) as well as for RNA sequencing (RNA-seq). Strikingly, almost all steps of the various protocols have been reported to introduce bias, especially in the case of RNA-seq, which is technically more challenging than DNA-seq. For each type of bias we discuss methods for improvement with a view to providing some useful advice to the researcher who wishes to convert any kind of raw nucleic acid into an NGS library.

Contents

*Corresponding author.
E-mail address: vandijk@cgm.cnrs-gif.fr (E.L. van Dijk).

0014-4827/$ - see front matter © 2014 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.yexcr.2014.01.008

Introduction

Over the recent years, NGS technology has become an essential tool for nearly all fields of biological research [1]. It enables parallel sequencing of millions of small DNA fragments for low per-base costs in a short time. Besides genome (re)sequencing, NGS provides accurate information on the composition of complex (c)DNA samples, making it the method of choice for most, if not all, genomic applications, such as transcriptome analysis (RNA-Seq), metagenomics, or profiling of methylated DNA (MeDip-seq), or DNA-associated proteins (ChIP-Seq). New applications for NGS appear frequently and furthermore, NGS applications are under intense scrutiny to produce even more and better quality data.

NGS requires the conversion of the source nucleic acid material into standard libraries suitable for loading onto a sequencing instrument. A wide variety of NGS library preparation protocols exist, but they all have in common that (fragments of) DNA or RNA molecules are fused with adapters that contain the necessary elements for immobilization on a solid surface and sequencing. In addition, size selection steps are often performed and libraries are usually amplified by PCR (Fig. 1). It is well understood that the quality of sequencing data depends highly upon the quality of the sequenced material. Therefore, the library construction process should guarantee a high molecular recovery of the original fragments (low bias and high complexity) in order to achieve the most genomic coverage with the least amount of sequencing. Importantly however, most popular library preparation protocols being used today may introduce serious biases in sample composition, which poses technical challenges and may lead to the misinterpretation of NGS data.

General methods of NGS sample preparation have been discussed previously in an excellent review [2]. Important new insights on biases in library preparation have appeared since, and here we review the current literature on biases in NGS libraries for 'DNA-seq' applications (e.g. genomic DNA-seq, ChIP-seq, exome sequencing) or 'RNA-seq' applications (e.g. transcriptome analysis, small RNA-seq). The focus will be on solutions to reduce bias and to ameliorate library quality for the Illumina platform, but most of the general principles will also apply to the other systems.

DNA-sequencing

The starting material for DNA-sequencing (DNA-seq) is generally double-stranded DNA in the form of isolated genomic DNA or chromatin (ChIP-seq). This DNA or chromatin is fragmented, followed by immunoprecipitation and removal of DNA-bound proteins in the case of ChIP-seq. These steps are followed by end-repair and adapter ligation, and usually a size selection step to remove free adapters and to select molecules in the desired size range (Fig. 1). Next, PCR amplification is often performed to generate sufficient quantities of template DNA to allow accurate quantification and to enrich for successfully adaptered fragments. PCR can also serve to add additional adapter sequence using tailed primers, resulting in template molecules that contain all the necessary elements for bridge amplification on the flowcell surface and for sequencing. Below we will discuss the different steps of the workflow that have been implicated in bias introduction. These are the steps of fragmentation, size selection and especially PCR; end-repair and adapter ligation do not appear to introduce detectable bias [3]. A summary with suggestions for improvement is shown in Table 1.

Fragmentation

DNA shearing is typically achieved either by mechanical force through nebulization or sonication, or by enzymatic digestion. Whereas fragmentation of naked DNA has not been considered a major source of bias, chromatin sonication for ChIP-seq has been shown to be non-random, with euchromatin being sheared more efficiently than heterochromatin [4]. As a result, DNA fragments of the selected size for library preparation (~200 base pairs) will preferentially contain euchromatin DNA, while heterochromatin is underrepresented. To solve this problem, Mokry and colleagues developed a double-fragmentation ChIP-seq protocol [5]. After conventional crosslinking and immunoprecipitation, chromatin is de-crosslinked and sheared a second time to concentrate fragments in the optimal size range for NGS. This approach not only reduces bias against heterochromatin DNA, but also increases the yield of material.

Size selection

Many current protocols use solid-phase reversible immobilization (SPRI) beads, sold by Beckman Coulter as AMPure beads, for size selection. The use of SPRI beads provides a fast and efficient method to enrich for DNA molecules of a selected size range. However, gel extraction is still commonly used as it allow for a more precise size selection. Quail and colleagues identified that melting agarose gel slices by heating to 50 °C in chaotropic salt buffer decreased the representation of AT-rich sequences. This possibly reflects a higher affinity of spin columns for double-stranded DNA, as strands with a high AT content are most likely to become denatured during this step and may not re-anneal [6]. A simple solution to this problem is to melt the gel slices in the