Forensic Science International: Genetics 17 (2015) 35–42

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Transcriptomic analysis of degraded forensic body fluids

Meng-Han Lin [a], Daniel F. Jones [b], Rachel Fleming [a,*]

[a] Forensic Research and Development, Institute of Environmental Science and Research Limited, Auckland, New Zealand
[b] New Zealand Genomics Ltd./Bioinformatics Institute, School of Biological Sciences, University of Auckland, Auckland, New Zealand

ARTICLE INFO

Article history:
Received 22 December 2014
Received in revised form 6 March 2015
Accepted 10 March 2015

Keywords:
RNASeq
Transcriptomics
Forensic
Body fluid identification
Degraded samples

ABSTRACT

Massively parallel sequencing (MPS) has facilitated a significant increase in transcriptomic studies in all biological disciplines. However, the analysis of degraded RNA remains a genuine challenge in practice. In forensic science the biological samples encountered are often extensively degraded and of low abundance. RNA from these compromised samples is used for body fluid identification through the detection of body fluid-specific transcripts. Here we demonstrate the sequencing of four forensically relevant body fluids: oral mucosa/saliva (buccal), circulatory blood, menstrual blood and vaginal fluid. RNA was extracted from fresh, two and six week aged samples. Despite the extensive degradation of most body fluids, significant high quality sequencing output (>80% sequence above Q30) was generated. An average of over 80% of reads from all but one sample aligned successfully to the reference human genome. Furthermore, FPKMs (fragments per kilobase of exon per million fragments mapped) generated indicate the accurate detection of known body fluid markers in respective body fluids. Assessment of global gene expression levels over degradation time enabled the characterisation of differential RNA degradation in different body fluids. This study demonstrates the practical application of MPS technology for the accurate analysis of degraded RNA from minimal samples.

1. Introduction

The rapid advance of massively parallel sequencing (MPS) technologies has accelerated molecular research and diagnostic capabilities over the past decade [1]. Although the first human genome project was completed using shotgun sequencing with Sanger sequencing chemistry [2], whole transcriptome sequencing was only made possible by the availability of RNA sequencing (RNAseq) through MPS.

Transcriptomic analyses before MPS largely involved the targeted, deductive quantification of the abundance of known transcripts using PCR- or hybridisation-based methods [1]. The availability of RNAseq using MPS has enabled the rapid discovery and characterisation of novel transcripts and other RNA species such as, among others, non-coding RNA, small RNA and microRNA [3,4]. Furthermore, RNAseq has facilitated the high throughput identification of transcript isoforms [5].

The intricate nature of RNA analysis means that input RNA integrity and abundance are highly influential on transcript quantification [6,7]. In spite of this, there is no clear consensus on the level of degradation that renders samples unusable, although a range of arbitrary sample quality thresholds based mostly on RNA integrity scores (RIN) have been suggested [8–10]. RIN is now also known to have a significant impact on RNAseq analysis [11], with manufacturers applying recommended sample quality requirements in standard methodology. Although sample quality thresholds may be easily met for standard laboratory applications, this is often not possible for realistic forensic samples that are obtained *in situ*.

A novel application of RNA analysis is in the discipline of forensic body fluid identification. Although DNA profiling is routinely used worldwide to identify the donor of biological traces from crime samples [12], it does not determine the types of cells from which the biological material was deposited or transferred. This contextual information may have significant implications for the investigation of the crime. For example, semen identified in a sample would indicate a different activity had occurred than if saliva was identified.

Current forensic casework body fluid identification using RNA based technology involves the detection of known body fluid-specific RNA markers using RT-PCR and qPCR [13–19]. This is often challenging as forensic samples are usually of limited quantity and have been subjected to environmental insults before sampling. Recent transcriptomic analyses of forensic body fluids through RNAseq further characterise the body fluid transcriptomes in an effort to identify more novel, specific and more readily detectable

* Corresponding author. Tel.: +64 9845 1736; fax: +64 98496046.
E-mail address: rachel.fleming@esr.cri.nz (R. Fleming).

http://dx.doi.org/10.1016/j.fsigen.2015.03.005
1872-4973/

targets [20,21]. However, a potential limitation of these studies is that the samples sequenced may have been relatively intact (RIN not assessed) and abundant, which is unlike most samples encountered in a forensic setting.

RNAseq of compromised samples has been a focus of recent research [11,22–24]. Here we demonstrate whole transcriptome sequencing of realistic, degraded forensically relevant body fluid samples. We believe that by using this method body fluid specific RNA markers are more likely to be detectable in a realistic sample and that this approach will have far reaching applications to similar disciplines where the analysis of degraded samples is the only option.

2. Materials and methods

2.1. Body fluid sampling and aging (RNA degradation)

Fresh body fluid samples (oral mucosa/saliva (buccal), circulatory blood, vaginal fluid and menstrual blood) were collected on sterile Cultiplast® rayon swabs and aged at room temperature with exposure to ambient laboratory conditions, for $t = 0$, two and six weeks. Samples were collected from two individuals for circulatory blood and buccal and from one individual for menstrual blood and vaginal fluid. Triplicate samples (2 swabs per replicate) were collected on the same day from each individual, for each body fluid at each time point. Oral mucosa/saliva, vaginal fluid and menstrual blood samples were obtained by swabbing by the participants themselves while 50 μL of fresh circulatory blood was drawn using a sterile ACCU-CHEK® Safe-T-Pro Plus lancet (Roche Diagnostics USA, Indianapolis, IN, USA) and deposited onto each swab.

2.2. RNA extraction

Total RNA for all samples was extracted using the Promega® ReliaPrep™ RNA Cell Miniprep System (Promega Corporation, Madison, WI, USA) following the manufacturer's instructions. DNA was removed from extracted RNA using on-column DNase I treatment during the RNA extraction process. RNA was eluted in 50 uL elution buffer. Complete removal of human DNA was verified using the Quantifiler® Human DNA quantification kit (Life Technologies Corp., Carlsbad, CA, USA) using 1 uL of sample in a 12.5 uL reaction.

2.3. RNA integrity analysis and quantification

RNA integrity for each sample was determined using the Agilent RNA 6000 pico kit (Agilent Technologies, Santa Clara, CA, USA) and the 2100 Bioanalyzer instrument (Agilent Technologies, Santa Clara, CA, USA). RNA concentration was determined using the Qubit® RNA HS assay kit (Life Technologies Corp., Carlsbad, CA, USA) and the Qubit® 2.0 fluorometer (Life Technologies Corp., Carlsbad, CA, USA).

2.4. Library preparation and sequencing

cDNA libraries for RNAseq were prepared using Bioo Scientific NEXTFlex directional RNA-seq Kit (dUTP-Based) v2 48 (Bioo Scientific, Austin, TX, USA). Total RNA was not subjected to ribosomal RNA depletion. Due to the low concentration and degraded nature of some samples, 13 μL total RNA input was used for library preparation irrespective of concentration. ERCC (External RNA Controls Consortium) spike-in controls are a set of 92 artificial RNAs, at a range of known concentrations and lengths [25,26]. One microlitre of ERCC controls (Life Technologies Corp., Carlsbad, CA, USA) diluted 1000 fold was added to each sample. Barcodes (1–16) were added to each library using the NEXTflex RNA-Seq barcodes kit (Bioo Scientific, Austin, TX, USA).

Barcoded libraries were sequenced across three lanes on an Illumina HiSeq2500 sequencer, with 2×100 bp paired-end chemistry.

2.5. Bioinformatics analysis

Read quality for all samples were analysed using SolexaQA [27]. Data were preprocessed using DynamicTrim v1.9 using default settings [27]. Data were length-sorted and unpaired reads discarded using Lengthsort v1.9 using default settings [27]. Subsequent processed data consisted entirely of reads with <5% probability of error (or a Q score of >13), with pairs, and length >25 bp.

Reads were aligned to the human genome hg19 (GRCh37) [28]. The "UCSC genes" annotation track of known genes was downloaded from the UCSC genome browser as hg19_UCSC_genes.gtf [28].

FASTA and gtf format files of ERCC spike-in controls were obtained from the manufacturer's website (http://www.lifetechnologies.com/order/catalog/product/4456739). These ERCC annotations were concatenated onto the end of the hg19 FASTA and gtf annotation tracks. ERCC controls were analysed in the same way as the other genes in subsequent analyses. Therefore, abundance estimation values for the 92 artificial ERCC RNAs are directly comparable to other transcripts in the sample.

Processed reads were mapped to the combined human genome (hg19)/ERCC controls using Tophat2 v2.0.12 [29] with the following switches: `—library-type fr-firststrand -M $leftread $rightread`.

Transcripts were reconstructed from splice-aware mapping results from Tophat2, using Cufflinks v2.2.1 [30] with the following switches: `-g -b -u —library-type fr-firststrand —library-norm-method geometric`.

The reconstructed transcripts from each sample were merged into a single .gtf file using Cuffmerge v2.2.1 [31] with the following switches: `-g -s`.

Abundance values are expressed as Fragments Per Kilobase of exon per Million fragments mapped (FPKM = [# of fragments]/([length of transcript]/1000)/([total reads]/10^6); [32])). Library size normalised expression (FPKM) for each sample was generated using Cuffnorm v2.2.1 [30] with the following switches: `—library-type fr-firststrand -library-norm-method geometric -output-format cuffdiff`. FPKM values for all 92 ERCC RNAs were obtained for all samples, and plotted against their known concentrations using a custom R script [29].

3. Results

3.1. RNA integrity analysis

RNA integrity of the samples were measured on the basis of the RNA Integrity (RIN) score using the Agilent 2100 Bioanalyzer system [33]. RIN scores were used to determine RNA degradation over time for respective body fluid types. RIN scores for all samples are listed in Supplementary Table 1. Circulatory blood samples exhibit a consistent and proportional decrease in RIN score throughout the degradation time course. Menstrual blood, buccal and vaginal fluid consistently exhibit low RIN scores (average RIN ~1–4) throughout the degradation time course. With the exception of fresh ($t = 0$) circulatory blood, the RIN scores of all other samples are below the minimum recommended (RIN > 8) for gene expression/transcript abundance analysis (qPCR, microarray, RNAseq) [11,34,35].