Analysis of telomerase target gene expression effects from murine models in patient cohorts by homology translation and random survival forest modeling

Frederik Otzen Bagger [a,1], Claudia Bruedigam [b,1], Steven W. Lane [b,c,*]

[a] The Finsen Laboratory, Bioinformatics Centre at Department of Biology, and Biotech Research and Innovation Center (BRIC), University of Copenhagen, 2200 Copenhagen N, Denmark
[b] Division of Immunology, QIMR Berghofer Medical Research Institute, Brisbane, QLD 4006, Australia
[c] University of Queensland, Brisbane, QLD 4072, Australia

ARTICLE INFO

ABSTRACT

Acute myeloid leukemia (AML) is an aggressive and rapidly fatal blood cancer that affects patients of any age group. Despite an initial response to standard chemotherapy, most patients relapse and this relapse is mediated by leukemia stem cell (LSC) populations. We identified a functional requirement for telomerase in sustaining LSC populations in murine models of AML and validated this requirement using an inhibitor of telomerase in human AML. Here, we describe in detail the contents, quality control and methods of the gene expression analysis used in the published study (Gene Expression Omnibus GSE63242). Additionally, we provide annotated gene lists of telomerase regulated genes in AML and R code snippets to access and analyze the data used in the original manuscript.

© 2016 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Specifications	
Organism/cell line/tissue	*Mus musculus*
Sequencer or array type	Illumina MouseWG-6 v2.0 expression beadchip
Data format	Raw and processed data
Experimental factors	G3 Terc −/− LSC versus WT LSC
Experimental features	Total RNA obtained from G3 Terc −/− LSCs compared to WT LSCs (MLLAF9 + gfp +, Lin −, Kit +, FcgR +) purified from primary recipients at individual disease onset.
Consent	Public available data
Sample source location	N/A

1. Direct link to deposited data

The online data can be accessed at: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63242

2. Experimental design, materials and methods

2.1. Objective

To determine whether the gene expression changes induced by telomerase loss in a mouse model of acute myeloid leukemia have relevance to human disease.

2.2. Preparation of mouse microarray samples

2.2.1. Generation of murine leukemia

Murine AML was generated by isolating purified hematopoietic stem and progenitor cell populations using fluorescent activated cell sorting (FACS on lineage^{negative}Kit$^+$ Sca1$^+$) from wild type C57Bl6 (WT) or 3rd generation Terc −/− mice. Stem cells were transduced with retrovirus pMIG-MLLAF9 [4,10] and injected into irradiated WT recipient mice (5.5Gy radiation) via the lateral tail vein. At disease onset, bone marrow was harvested from the mice and purified leukemia stem cell enriched populations were obtained by FACS (GFP$^+$ lineage^{negative}Kit$^+$ Sca1$^-$ FcgR$^+$).

2.2.2. Preparation of microarray samples

WT and G3 Terc −/− MLL-AF9 LSC were purified from primary recipients at AML onset. RNA was extracted with a QIAGEN RNeasy

Micro Kit, preamplified with the Illumina TotalPrep RNA Amplification Kit, and hybridized on Illumina MouseWG-6 v2.0 BeadChip array.

2.3. Analysis of microarray data

2.3.1. Mouse Terc −/− expression array pre-processing

Illumina MouseWG-6 v2 BeadChip array images were processed with default parameters by Illumina GenomeStudio including trimming and collapsing of beads. The arrays were processed using a single color to determine the expression intensities (green). In R (programming language for statistical computing) [13] we imported the expression intensities from the resulting idat files using IDATreader (http://www.compbio.group.cam.ac.uk/software/idatreader). The IDATreader package imports the binary .idat-files and returns a data frame with values from GenomeStudio summarized over beads, including statistics on the background intensity and the number of good beads used for the trimmed, averaged, and binned final value for each probe that we used for further processing. The Illumina bin codes were used to correctly annotate each bin to probe with information acquired from Illumina webpage (http://support.illumina.com/array/downloads.html).

In order to import the dataset into R and make a standard expression Set class run the following code:

```
#install library if not present, and import
if (!"GEOquery" %in% installed.packages())
{source("http://bioconductor.org/biocLite.R");
biocLite("GEOquery")};
library(GEOquery);

#download the data
geoData <-getGEO('GSE63242')

#extract the expressionSet class
Geset <-geoData[[1]]

# use sample names in the expression matrix
colnames(exprs(Geset))=as.character(pDta(Geset)[['title']])
```

2.3.2. Quality control

In order to test for quality of the arrays we used the Bioconductor [5] package arrayQualityMetrics [9]. Here we found that one array failed (7166151049_F) and the density distribution was slightly more narrow (Fig. 1, higher blue stippled line). In the principle component analysis plot, which allows 2D inspection of the relation between samples using information from all probes (full dimensionality), we see that the sample marked for low quality lies in between all the rest of the samples, and does not look like an outlier. Hence, the biological signal (although having less quantitative intensity) is in line with the rest of the samples. We therefore decided to include it in the further analysis. The signal appears to be weaker, but not diverging from the replicates. For reuse of this data set some attention should be given to whether the signal in the sample is strong enough, if none of the replicates are used. Furthermore, we decided to include a technical replicate (7166151048_F_Grn) in the analysis to be able to better model between-array variance.

The data were background corrected with bgAjust, using the Illumina control probes and normalized using variance stabilization transformation [12] and quantile normalized (quantile bin size = 1, as described by Bolstad et al. [2]) in the multistep function lumiExpresso in the R package lumi [5] (Fig. 2).

It has been previously described that the provided probe annotation from Illumina includes imprecise or erroneous entries [2]. Therefore probes were re-annotated using a multiple sequence alignment based directory as described previously [7] (version mm9_V1.0.0_Aug09). For analysis where a very high specificity is preferred at the cost of the total measured transcripts a filtering for poor or unspecific probes was performed as described previously [1].

For differential expression analysis between Wild type mouse (WT) and Terc −/− groups, we used the standard limma package pipeline using eBayes on the lm Fit object of an expressionSet of all 12 samples. A top 140 probes were selected, corresponding to an unadjusted alpha of 0.001 (Table 1).

2.4. Translation of murine telomerase regulated genes into human AML

In order to translate differentially expressed genes into human homologs we used HomoloGene (downloaded at http://www.ncbi.nlm.nih.gov/homologene, build 67), which is a dictionary of species-specific genes translated into cross species genes identifiers. Currently, a more direct and precise solution for species conversion of genes could be obtained though the R interface to Biomart (http://www.
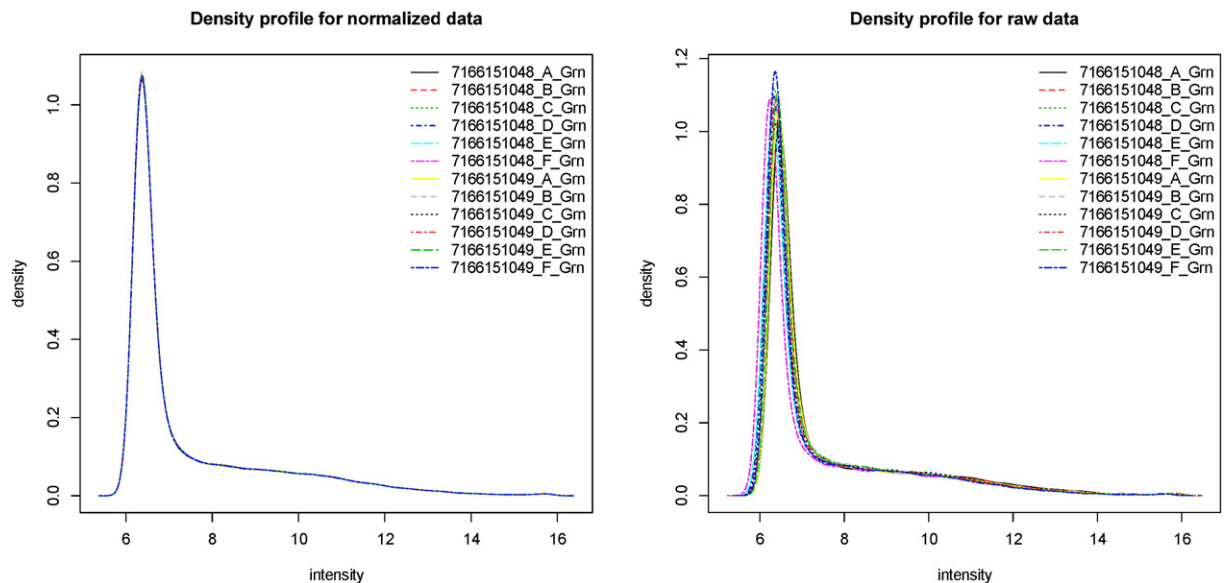

Fig 1. Density distribution of expression intensities of probes for the 12 arrays. One sample was marked outlier by arrayQualityMetrics based on a more narrow density profile (7166151049_F).