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

Bagger, Frederik Otzen; Bruedigam, Claudia; Lane, Steven W.

Published in:
Genomics Data

DOI:
10.1016/j.gdata.2016.01.014

Publication date:
2016

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY-NC-ND

Citation for published version (APA):
Analysis of telomerase target gene expression effects from murine models in patient cohorts by homology translation and random survival forest modeling

Frederik Otzen Baggera,1, Claudia Bruedigamb,1, Steven W. Laneb,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

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/).

Keywords: AML, Leukemia, Stem cells, Telomere, Telomerase

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−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−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 Genomics Data 7 (2016) 275–280

⁎ Corresponding author at: Division of Immunology, QIMR Berghofer Medical Research Institute, Brisbane, QLD 4006, Australia.
1 Denotes equal contribution.

http://dx.doi.org/10.1016/j.gdata.2016.01.014
2213-5960/© 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/).
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) 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:

```r
#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]]

colnames(exprs(Geset))=as.character(pData(Geset)[[‘title’]])
```

2.3.2. Quality control

In order to test for quality of the arrays we used the Bioconductor 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 lumiExpert 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.
To assess whether differentially expressed genes from our mouse model could separate patient data we used unsupervised Hierarchical clustering with the 112 genes in R with the Hartigan–Wong algorithm [7] with ten random starts for robust clustering. The patient data was publicly available data from GSE15210, where the survival information can also be found. We did indeed see that the 5 groups of patients with distinct expression patterns of these genes displayed significantly different survival patterns.

In order to pinpoint the TERC related genes that were driving the difference in patient survival we used Random survival forest (RSF) models [3]. We built the RSF using censored (known date of death) and uncensored survival observations in a survival model as implemented by Ishwaran and Kogalur, 2010 [8]. We doubled the number of deregulated gene homologs from the Terc−/− experiment as training set for the model, in order to gain some depth of the trees, while still retaining only TERC associated genes. Our RSF consisted of 20,000 decision trees, each trained on a subset of the data (bootstrapping), in such a way that it best explains the survival of the patients; lastly a majority vote between the trees gives the final prediction for new samples. RSF analysis for survival models is a powerful way to assess driving genes since many genes are known to be co-regulated or otherwise show correlating expression patterns. Since RSF is an ensemble model (results are summarized over multiple models) made with bootstrapping (only some of data is used to make each model) it is not sensitive to collinearity between the covariates, that is, the highly correlating genes A and B will not be featuring together in all models, and hence correct assessment of the impact of the individual contribution of A will be less dependent on B. Current standards like Cox proportional hazard regression assume no collinearity. In order to assess which genes are driving survival for the Terc−/− signature we used random permutation of a random selection of the labels, and thus the importance measure was the increase in error rate when a label was permuted, summarized over all the trees. In this way we are able to find which genes in our signature drive the importance of survival.

In order to work with the validation data from Metzeler et al. use GSE15210.

Random survival forest can be built on an expression matrix (here “training set”) that should include two columns providing information on overall survival (“os”) and censorship (“status”).

RSF=rsf (Surv (os, status) ~ ., ntree=20000, importance="permute", proximity=TRUE, data=trainingset)
IMPORTANT GENES=varSel (Surv (os, stat) ~ ., ntree=20000, data=trainingset)

3. Discussion

The healthy human hematopoietic system and the transformation to leukemia provide excellent, accessible and tractable models of normal cellular development and cancer progression. A number of high quality and well-annotated datasets from human donors with acute myeloid leukemia are publicly available for data driven cohort studies in a highly aggressive cancer. Here we describe a protocol for using these public datasets for hypothesis driven research, where they make findings from a knockout experiment in a model organism directly relevant in a clinical context. We translate a murine genetic signature of a 140 genes into 112 human homologs based on sequence similarity on protein and level and can show that the homolog signature impacts on patient survival, as expected from the mouse phenotype, and further we identify key driver genes in the signature.

Interspecies translations have given rise to a number of disappointments in the drug development industry and recently the species effect on gene expression was estimated higher than the tissue effect in the ENCODE mRNA expression studies [15], which was later confirmed [11]. However, careful reanalysis has greatly questioned this notion [6], also in line with previous studies [14]. The method proposed in this study, where a small binary list of genes, translated by evolutionary protein family, is used for investigating the clinical effect only is more conservative in both the means and conclusions. Furthermore, the effect of single genes, potentially misclassified, is greatly reduced by the following RSF model analysis for important contribution to survival. Rather than a species comparison we utilize a patient cohort study to enrich our data, support our findings from the animal model system, and further provide direct clinical relevance.

Fig. 2. Principle component analysis of microarrays, showing clustering of wild type and G3 Terc−/− samples. Furthermore, one sample marked for low quality (sample 12, triangle) lies in center of the plot and does thus not drag the axis of the PCA, having a signal which is weaker than the rest of the samples, but not diverging from them. Right plot shows the normalized data (lumi package for R) and left is the un-normalized data. Here, the technical replicates are closely associated, as expected.

From this conversion 112 genes could be translated into human counterparts (see Table 1). To summarize over all the trees. In this way we are able to sure was the increase in error rate when a label was permuted, a proportion of a random selection of the labels, and thus the importance measure between the covariates, that is, the highly correlating genes A and B would be less dependent on B. Current standards like Cox proportional hazard regression assume no collinearity. In order to pinpoint the TERC related genes that were driving the difference in patient survival we used Random survival forest (RSF) models [3]. We built the RSF using censored (known date of death) and uncensored survival observations in a survival model as implemented by Ishwaran and Kogalur, 2010 [8]. We doubled the number of deregulated gene homologs from the Terc−/− experiment as training set for the model, in order to gain some depth of the trees, while still retaining only TERC associated genes. Our RSF consisted of 20,000 decision trees, each trained on a subset of the data (bootstrapping), in such a way that it best explains the survival of the patients; lastly a majority vote between the trees gives the final prediction for new samples. RSF analysis for survival models is a powerful way to assess driving genes since many genes are known to be co-regulated or otherwise show correlating expression patterns. Since RSF is an ensemble model (results are summarized over multiple models) made with bootstrapping (only some of data is used to make each model) it is not sensitive to collinearity between the covariates, that is, the highly correlating genes A and B will not be featuring together in all models, and hence correct assessment of the impact of the individual contribution of A will be less dependent on B. Current standards like Cox proportional hazard regression assume no collinearity. In order to assess which genes are driving survival for the Terc−/− signature we used random permutation of a random selection of the labels, and thus the importance measure was the increase in error rate when a label was permuted, summarized over all the trees. In this way we are able to find which genes in our signature drive the importance of survival.

In order to work with the validation data from Metzeler et al. use GSE15210.

biomart.org). From this conversion 112 genes could be translated into human counterparts (see Table 1).
<table>
<thead>
<tr>
<th>Probe number</th>
<th>Illumina probe ID</th>
<th>mgi Symbol</th>
<th>Gene group</th>
<th>HUGO homolog</th>
<th>Terc −/− vs WT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,770,767</td>
<td>ILMN_284389</td>
<td>abParts</td>
<td>NA</td>
<td>MYOM1</td>
<td>DOWN</td>
</tr>
<tr>
<td>3,120,619</td>
<td>ILMN_2815138</td>
<td>Mym1</td>
<td>31,196</td>
<td>KIAA0930</td>
<td>DOWN</td>
</tr>
<tr>
<td>70,431</td>
<td>ILMN_2428798</td>
<td>Cdk5r1</td>
<td>31,200</td>
<td>CDK5R1</td>
<td>DOWN</td>
</tr>
<tr>
<td>7,210,458</td>
<td>ILMN_2815138</td>
<td>Pxxcn1</td>
<td>4211</td>
<td>PXXCN1</td>
<td>DOWN</td>
</tr>
<tr>
<td>130,437</td>
<td>ILMN_2624571</td>
<td>Msx3</td>
<td>32,333</td>
<td>MSX3</td>
<td>DOWN</td>
</tr>
<tr>
<td>3,710,544</td>
<td>ILMN_1238479</td>
<td>Mgst3</td>
<td>3327</td>
<td>MGST3</td>
<td>DOWN</td>
</tr>
<tr>
<td>7,160,133</td>
<td>ILMN_2507232</td>
<td>Gass2/3</td>
<td>18,386</td>
<td>GAS2L3</td>
<td>DOWN</td>
</tr>
<tr>
<td>1,710,377</td>
<td>ILMN_1251616</td>
<td>Skp2</td>
<td>55,942</td>
<td>SKP2</td>
<td>DOWN</td>
</tr>
<tr>
<td>4,640,414</td>
<td>ILMN_2541675</td>
<td>583018K08Rik</td>
<td>27,936</td>
<td>KIAA1731</td>
<td>DOWN</td>
</tr>
<tr>
<td>2,140,052</td>
<td>ILMN_2856861</td>
<td>Nduc-ps1</td>
<td>NA</td>
<td>SCG1</td>
<td>DOWN</td>
</tr>
<tr>
<td>3,120,672</td>
<td>ILMN_2542231</td>
<td>Ppig</td>
<td>3520</td>
<td>PPG1</td>
<td>DOWN</td>
</tr>
<tr>
<td>1,300,725</td>
<td>ILMN_2900216</td>
<td>Ndufb10</td>
<td>3343</td>
<td>NDUFB10</td>
<td>DOWN</td>
</tr>
<tr>
<td>5,050,072</td>
<td>ILMN_2853714</td>
<td>Apte1</td>
<td>66,004</td>
<td>APET1</td>
<td>DOWN</td>
</tr>
<tr>
<td>1,590,593</td>
<td>ILMN_1218592</td>
<td>Tes</td>
<td>41,051</td>
<td>TES</td>
<td>DOWN</td>
</tr>
<tr>
<td>7,610,450</td>
<td>ILMN_3155180</td>
<td>Itpr2</td>
<td>37,593</td>
<td>ITPR2</td>
<td>DOWN</td>
</tr>
<tr>
<td>7,550,121</td>
<td>ILMN_3084954</td>
<td>Tes</td>
<td>41,051</td>
<td>TES</td>
<td>DOWN</td>
</tr>
<tr>
<td>510,673</td>
<td>ILMN_2488125</td>
<td>Vrk1</td>
<td>2541</td>
<td>VRK1</td>
<td>DOWN</td>
</tr>
<tr>
<td>6,900,762</td>
<td>ILMN_1218128</td>
<td>Tatdn1</td>
<td>57,158</td>
<td>TATDN1</td>
<td>DOWN</td>
</tr>
</tbody>
</table>
In conclusion, we devise a protocol for analyzing gene expression effects from model organisms in patient cohorts, by means of homology translation and RSF models. We present our pipeline for analyzing a two-condition Illumina expression array study, with considerations on probe re-annotation, outlier detection and batch effects. All the presented data are available in raw and processed on GSE63242.

Acknowledgments

This work was funded by the Leukaemia Foundation (2013_GIA_Lane), Cancer Australia (1059292_2014), Royal Brisbane and Women’s Hospital Foundation: CIA_Lane and the Rhys Pengelly Fellowship in Leukaemia Research. SWL is a Career Development Fellow of the NHMRC Australia.
References


