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

Genomics Data 7 (2016) 275–280

Contents lists available at ScienceDirect

Genomics Data

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

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

Article history:
Received 11 January 2016
Received in revised form 18 January 2016
Accepted 27 January 2016
Available online 28 January 2016

Keywords:
AML
Leukemia
Stem cells
Telomere
Telomerase

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.

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

* 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

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.

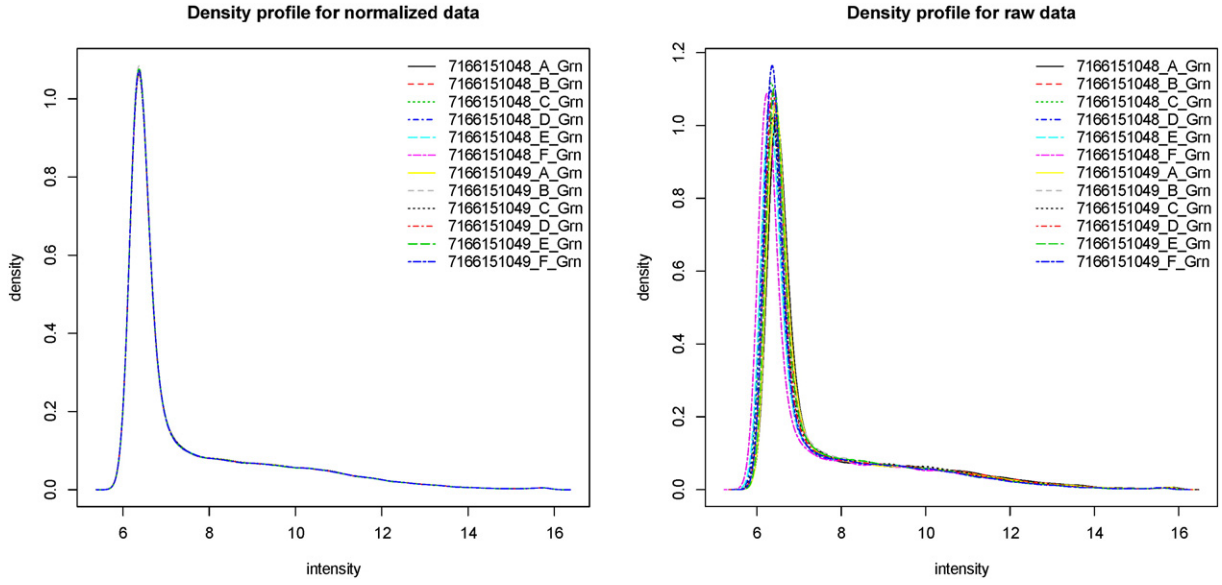

Fig 1. Density distribution of expression intensities of probes for the 12 arrays. One sample was marked outlier by arrayQualityMetrics based on a more narrow density profile (7166151049_F).

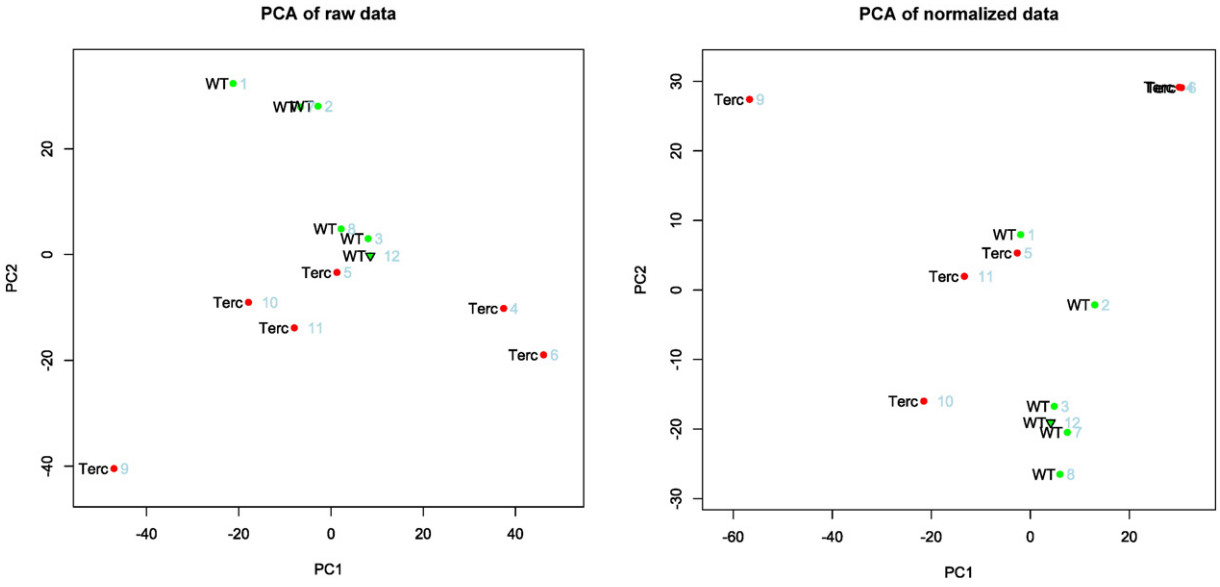

Fig. 2. Principle component analysis of microarrays, showing clustering of wild type and G3 Terc −/− samples. Furthermore, one sample marked for low quality (sample 12, triangle) lies in center of the plot and does thus not drag the axis of the PCA, having a signal which is weaker than the rest of the 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.

biomart.org). From this conversion 112 genes could be translated into human counterparts (see Table 1).

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 patters. 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.

Table 1
Designated Terc −/− regulated AML gene set (Table 1).

Probe number	Illumina probe ID	mgi Symbol	Gene group	HUGO homolog	Terc −/− vs WT
1,770,767	ILMN_2483493	abParts	NA		DOWN
3,120,619	ILMN_2815138	Myom1	31,196	MYOM1	DOWN
70,431	ILMN_2428798	5031439G07Rik	15,140	KIAA0930	DOWN
7,210,458	ILMN_1259339	Cdk5r1	31,200	CDK5R1	DOWN
5,050,072	ILMN_2481902	Plxnc1	4211	PLXNC1	DOWN
130,437	ILMN_2642571	Mxd3	32,333	MXD3	DOWN
3,710,544	ILMN_1238479	Mgst3	3327	MGST3	DOWN
7,160,133	ILMN_2507232	Gas2l3	18,386	GAS2L3	DOWN
1,710,377	ILMN_1251616	Skp2	55,942	SKP2	DOWN
4,640,414	ILMN_2541675	5830418K08Rik	27,936	KIAA1731	DOWN
2,070,242	ILMN_2643883	Brip1	32,766	BRIP1	DOWN
6,980,315	ILMN_1241320	Clspn	11,138	CLSPN	DOWN
6,420,215	ILMN_1235363	Gsg2	49,236	GSG2	DOWN
4,210,619	ILMN_2615035	Mgst3	3327	MGST3	DOWN
110,039	ILMN_2785454	Hist2h2ab	111,318	HIST2H2AB	DOWN
6,290,689	ILMN_1237886	Enc1	2694	ENC1	DOWN
1,510,132	ILMN_1243663	G2e3	32,362	G2E3	DOWN
5,860,139	ILMN_2817151	Chchd8	9567	COA4	DOWN
50,446	ILMN_2589960	Gins3	41,496	GINS3	DOWN
2,320,102	ILMN_2524519	Rasgef1a	17,067	RASGEF1A	DOWN
6,760,088	ILMN_1258300	Ifngr2	4041	IFNGR2	DOWN
4,050,711	ILMN_2627660	Lig1	197	LIG1	DOWN
70,546	ILMN_2981801	Hist1h2ag	69,326	HIST1H2AG	DOWN
730,743	ILMN_2537961	Mcm7	4323	MCM7	DOWN
160,253	ILMN_2760244	Snx7	22,941	SNX7	DOWN
5,090,332	ILMN_1224268	Mrps15	32,636	MRPS15	DOWN
7,210,470	ILMN_2517171	Tuba4a	68,496	TUBA4A	DOWN
7,050,605	ILMN_2511401	Upf3a	23,395	UPF3A	DOWN
3,360,400	ILMN_2633492	Chek2	38,289	CHEK2	DOWN
5,490,767	ILMN_1255902	Smc4	4015	SMC4	DOWN
6,220,609	ILMN_2461345	Zfp41	65,280	ZFP41	DOWN
4,290,524	ILMN_3080371	Fert2	74,300	FER	DOWN
1,980,431	ILMN_1231587	BC030867	69,368	C17orf53	DOWN
2,070,458	ILMN_2989480	Dsn1	49,806	DSN1	DOWN
7,040,612	ILMN_2525289	C330018D20Rik	35,412	C5orf63	DOWN
1,470,050	ILMN_2664593	Hist1h1b	110,910	HIST1H1B	DOWN
2,340,403	ILMN_3137980	Zfp41	65,280	ZFP41	DOWN
4,850,059	ILMN_2483253	Dicer1	13,251	DICER1	DOWN
450,678	ILMN_2826161	Taf12	68,477	TAF12	DOWN
5,700,646	ILMN_2658153	Zcchc17	32,319	ZCCHC17	DOWN
3,180,170	ILMN_1248181	Zbtb7a	7820	ZBTB7A	DOWN
3,890,519	ILMN_3055904	Cbx5	7257	CBX5	DOWN
5,360,368	ILMN_2511868	2310002B06Rik	NA		DOWN
4,610,129	ILMN_1248830	Hist1h2an	69,326	HIST1H2AN	DOWN
5,810,176	ILMN_1214664	Glrx2	41,098	GLRX2	DOWN
4,890,341	ILMN_2683414	Snf8	5239	SNF8	DOWN
5,490,035	ILMN_2416488	Usp37	10,858	USP37	DOWN
670,739	ILMN_1246108	Hist1h2ah	130,520	HIST1H2AI	DOWN
540,037	ILMN_1216285	Creb3	31,375	CREB3	DOWN
3,520,221	ILMN_2694275	Lxn	36,361	LXN	DOWN
5,560,451	ILMN_1251771	Cyc1	55,617	CYC1	DOWN
6,250,446	ILMN_1248184	Senp1	8731	SENP1	DOWN
4,830,291	ILMN_2718861	1600012H06Rik	57,051	C6orf120	DOWN
5,960,491	ILMN_2593872	Mrps15	32,636	MRPS15	DOWN
6,840,170	ILMN_2707291	Prdm2	40,822	PRDM2	DOWN
3,990,360	ILMN_3137920	Sel1l	31,286	SEL1L	DOWN
4,220,504	ILMN_1256203	L3mbtl2	12,882	L3MBTL2	DOWN
2,340,494	ILMN_2665625	Fadd	2836	FADD	DOWN
2,350,221	ILMN_1246502	E330016A19Rik	NA	E330016A19RIK	DOWN
1,990,731	ILMN_2596297	Ddt	1038	DDT	DOWN
5,810,564	ILMN_1245139	Scamp3	4164	SCAMP3	DOWN
3,400,754	ILMN_2919433	Cdc45l	NA		DOWN
7,050,612	ILMN_2822131	Hmgcl	159	HMGCL	DOWN
1,300,358	ILMN_2971481	Znrd1	40,960	ZNRD1	DOWN
2,370,474	ILMN_3026137	Dbndd2	12,276	DBNDD2	DOWN
2,140,092	ILMN_2856861	Nudc-ps1	NA		DOWN
940,427	ILMN_2686509	Sgol1	23,642	SGOL1	DOWN
3,120,672	ILMN_2542231	Ppig	3520	PPIG	DOWN
1,300,725	ILMN_2900216	Ndufb10	3343	NDUFB10	DOWN
2,570,398	ILMN_2983714	Apitd1	66,004	APITD1	DOWN
5,690,593	ILMN_1218592	Tes	41,051	TES	DOWN
7,610,450	ILMN_3155180	Itpr2	37,593	ITPR2	DOWN
7,550,121	ILMN_3084954	Tes	41,051	TES	DOWN
510,673	ILMN_2488125	Vrk1	2541	VRK1	DOWN
6,900,762	ILMN_1218128	Tatdn1	57,158	TATDN1	DOWN

Table 1 (*continued*)

Probe number	Illumina probe ID	mgi Symbol	Gene group	HUGO homolog	Terc −/− vs WT
4,540,619	ILMN_1259294	Tmem126b	10,222	TMEM126B	DOWN
5,130,497	ILMN_1221592	Sec11c	8624	SEC11C	DOWN
2,320,367	ILMN_2798803	Vps4a	69,132	VPS4A	DOWN
4,390,075	ILMN_2802103	Mfap1b	4332	MFAP1	DOWN
1,450,731	ILMN_1240146	Tor1b	56,677	TOR1B	DOWN
6,020,047	ILMN_1214907	Stk10	38,122	STK10	DOWN
4,780,398	ILMN_2462791	Zfp322a	23,460	ZNF322	DOWN
5,080,632	ILMN_2742498	Nup88	1901	NUP88	DOWN
3,400,619	ILMN_2952114	Smc1a	4597	SMC1A	DOWN
5,290,215	ILMN_2963412	Ikzf5	23,363	IKZF5	DOWN
110,327	ILMN_2613904	Hspb2	68,189	HSPB2	DOWN
6,020,161	ILMN_3046362	Traf5	27,079	TRAF5	DOWN
7,380,193	ILMN_2450147	Zfp238	21,276	ZNF238	DOWN
3,120,438	ILMN_2536365	Nrbf2	41,473	NRBF2	DOWN
2,450,612	ILMN_2429469	Otud7b	10,624	OTUD7B	DOWN
2,680,630	ILMN_2659762	Mrpl18	8566	MRPL18	DOWN
6,370,746	ILMN_2466190	Cyb5rl	45,506	CYB5RL	DOWN
3,400,632	ILMN_1253970	Sirt7	56,152	SIRT7	DOWN
5,860,398	ILMN_2620061	Tbcb	981	TBCB	DOWN
6,480,184	ILMN_2830060	Wfdc10	86,879	WFDC10	DOWN
4,260,019	ILMN_2998548	Pycr2	8343	PYCR2	UP
4,120,039	ILMN_2691135	Itpkb	1672	ITPKB	UP
6,860,398	ILMN_1251074	Rffl	12,116	RFFL	UP
4,180,192	ILMN_1242872	9230110K08Rik	NA		UP
3,190,427	ILMN_2814333	Lgtn	NA		UP
1,030,703	ILMN_2790636	Sar1a	90,897	SAR1A	UP
2,970,196	ILMN_2698606	Tmed4	5308	TMED4	UP
3,520,634	ILMN_2700505	Tmem43	11,532	TMEM43	UP
6,110,044	ILMN_2698699	Farsa	3280	FARSA	UP
4,810,347	ILMN_3112185	Ensa	37,924	ENSA	UP
2,100,528	ILMN_2649101	Ncf2	374	NCF2	UP
6,550,474	ILMN_1230339	Slc9a8	75,041	SLC9A8	UP
5,560,408	ILMN_2759335	Rnmt	2816	RNMT	UP
1,780,725	ILMN_2580895	Gns	1568	GNS	UP
70,608	ILMN_2744121	Tmem181	44,787	TMEM181	UP
7,050,370	ILMN_2880536	Uck2	40,850	UCK2	UP
4,780,328	ILMN_1231573	Serpinb1a	69,399	SERPINB1	UP
5,360,474	ILMN_2627217	Pi4k2b	32,405	PI4K2B	UP
2,470,309	ILMN_2452717	AK011460	NA		UP
4,850,133	ILMN_1248389	Inpp5k	75,059	INPP5K	UP
5,810,070	ILMN_1242013	Uck2	40,850	UCK2	UP
4,490,639	ILMN_1254631	Uck2	40,850	UCK2	UP
1,990,524	ILMN_2557957	Eprs	5870	EPRS	UP
6,980,601	ILMN_2564872	Ddhd2	66,646	DDHD2	UP
770,445	ILMN_1224840	Bcl9	3191	BCL9	UP
5,090,156	ILMN_2766253	Mbnl1	23,186	MBNL1	UP
1,850,402	ILMN_2568028	Il2rg	172	IL2RG	UP
6,020,400	ILMN_2646322	Samsn1	11,148	SAMSN1	UP
770,575	ILMN_1254736	Myo5a	20,100	MYO5A	UP
150,019	ILMN_2665490	Litaf	37,974	LITAF	UP
3,310,291	ILMN_1225045	1700109H08Rik	130,776	1700109H08RIK	UP
1,450,735	ILMN_3122845	H1fx	4397	H1FX	UP
3,840,600	ILMN_2681601	Slc44a2	10,711	SLC44A2	UP
4,670,504	ILMN_1245263	Tmem181	44,787	TMEM181	UP
1,980,021	ILMN_2628567	Phlda3	8233	PHLDA3	UP
4,150,370	ILMN_1233117	Anxa2	20,857	ANXA2	UP
2,630,605	ILMN_2603976	Cass4	75,128	CASS4	UP
5,560,315	ILMN_2466453	AK005145	NA		UP
5,090,204	ILMN_2805207	B020018G12Rik	NA		UP
1,240,338	ILMN_1239814	AK011411	NA		UP
290,328	ILMN_2923607	Phlda3	8233	PHLDA3	UP
2,570,037	ILMN_2813484	Per1	1966	PER1	UP
2,690,348	ILMN_2700354	Dennd5b	44,911	DENND5B	UP
2,570,053	ILMN_2702102	D630023F18Rik	129,674	C2orf80	UP
160,463	ILMN_1258526	Lgals3bp	4067	LGALS3BP	UP

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

[1] N.L. Barbosa-Morais, M.J. Dunning, S.A. Samarajiwa, J.F.J. Darot, M.E. Ritchie, A.G. Lynch, S. Tavaré, A re-annotation pipeline for Illumina Bead Arrays: improving the interpretation of gene expression data. Nucleic Acids Res. 38 (3) (2010), e17http://dx.doi.org/10.1093/nar/gkp942.

[2] B.M. Bolstad, R. Irizarry, M. Astrand, T.P. Speed, A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19 (2) (2003) 185–193, http://dx.doi.org/10.1093/bioinformatics/19.2.185.

[3] L. Breiman, Random forests. Mach. Learn. 45 (1) (2001) 5–32, http://dx.doi.org/10.1023/A:1010933404324.

[4] C. Bruedigam, F.O. Bagger, F.H. Heidel, C. Paine Kuhn, S. Guignes, A. Song, ... S.W. Lane, Telomerase inhibition effectively targets mouse and human AML stem cells and delays relapse following chemotherapy. Cell Stem Cell 15 (6) (2014) 775–790, http://dx.doi.org/10.1016/j.stem.2014.11.010.

[5] P. Du, W.A. Kibbe, S.M. Lin, Lumi: a pipeline for processing Illumina microarray. Bioinformatics 24 (13) (2008) 1547–1548, http://dx.doi.org/10.1093/bioinformatics/btn224.

[6] Y. Gilad, O. Mizrahi-Man, A reanalysis of mouse ENCODE comparative gene expression data. F1000Research 121 (May) (2015) 1–24, http://dx.doi.org/10.12688/f1000research.6536.1.

[7] J.A. Hartigan, M.A. Wong, Algorithm AS 136: a K-means clustering algorithm. Appl. Stat. 28 (1) (1979) 100, http://dx.doi.org/10.2307/2346830.

[8] H. Ishwaran, U.B. Kogalur, E.H. Blackstone, M.S. Lauer, Random survival forests. Ann. Appl. Stat. 2 (3) (2008) 841–860, http://dx.doi.org/10.1214/08-AOAS169.

[9] A. Kauffmann, R. Gentleman, W. Huber, arrayQualityMetrics—a bioconductor package for quality assessment of microarray data. Bioinformatics 25 (3) (2009) 415–416, http://dx.doi.org/10.1093/bioinformatics/btn647.

[10] S.W. Lane, Y.J. Wang, C. Lo Celso, C. Ragu, L. Bullinger, S.M. Sykes, ... D.A. Williams, Differential niche and wnt requirements during acute myeloid leukemia progression. Blood 118 (10) (2011) 2849–2856, http://dx.doi.org/10.1182/blood-2011-03-345165.

[11] S. Lin, Y. Lin, J.R. Nery, M.A. Urich, A. Breschi, C.A. Davis, ... M.P. Snyder, Comparison of the transcriptional landscapes between human and mouse tissues. Proc. Natl. Acad. Sci. 111 (48) (2014) 201413624, http://dx.doi.org/10.1073/pnas.1413624111.

[12] S.M. Lin, P. Du, W. Huber, W.a. Kibbe, Model-based variance-stabilizing transformation for Illumina microarray data. Nucleic Acids Res. 36 (2) (2008) 1–9, http://dx.doi.org/10.1093/nar/gkm1075.

[13] R Development Core Team, R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2013 (URL http://www.R-project.org/).

[14] K.D. Rasmussen, G. Jia, J.V. Johansen, M.T. Pedersen, N. Rapin, F.O. Bagger, ... K. Helin, Loss of TET2 in hematopoietic cells leads to DNA hypermethylation of active enhancers and induction of leukemogenesis. Genes Dev. 29 (2015) 1–13, http://dx.doi.org/10.1101/gad.260174.115.stood.

[15] F. Yue, Y. Cheng, A. Breschi, J. Vierstra, W. Wu, T. Ryba, ... E.C. Mouse, A comparative encyclopedia of DNA elements in the mouse genome. Nature 515 (7527) (2014) 355–364, http://dx.doi.org/10.1038/nature13992.