Genomics Data 11 (2017) 77–80

Contents lists available at ScienceDirect

Genomics Data

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

Data in Brief

Gene expression profiling of U2AF2 dependent RNA-protein interactions during CD4 + T cell activation

Thomas C Whisenant

Department of Molecular and Experimental Medicine, The Scripps Research Institute, La Jolla, CA, United States

ARTICLE INFO

Article history:
Received 7 December 2016
Accepted 11 December 2016
Available online 14 December 2016

Keywords:
RNA-seq
RIP-seq
RIP-chip
T-cell activation

ABSTRACT

CD4 T cell activation is a central component of the mammalian adaptive immune response and is underscored by a dramatic change in the gene expression profile in these cells. The changes in gene expression that occur during T cell activation are regulated in multiple ways including post-transcriptionally by complexes of RNA-binding proteins. Recently, our study explored the role of the RNA-binding protein U2AF2 and its interacting proteins in mediating posttranscriptional changes in constitutive and alternative splicing during T cell activation. First, we used RNA-seq to identify the global changes in gene expression and splicing that occur with T cell activation. Next, we used RIP-seq to identify the specific genes bound to U2AF2 during T cell activation. After identification of the protein interacting partners of U2AF2, we used splicing sensitive microarrays to measure the effects on global gene expression of using siRNAs to knock down a sampling of these proteins. Finally, we used RIP-chip to measure the effects of the same siRNA knockdown on the transcripts specifically bound to U2AF2. Here we provide the experimental details and analysis of the gene expression data for each of these techniques, which have been deposited into Gene Expression Omnibus (GEO) with the Superseries ID: GSE62923.

Specifications	
Organism/cell line/tissue Genome or genomic data origin	Human
Sex Male or female	Female
Sequencer or array type Type of sequencer	Illumina HiSeq 2000 (RNA-seq), Affymetrix Human Transcriptome Array 2.0 GeneChip (Microarray)
Data format Raw or analyzed	Raw data: fastq files (RNA-seq), CEL files (Microarray); Processed data: bam files (RNA-seq), CHP files (Microarray)
Experimental factors Tumor vs. normal, any pretreatment of samples	Resting CD4 T cell culture vs Activated CD4 T cell culture (Total RNA, U2AF2 RIP RNA); U2AF1 siRNA vs Control siRNA (Total RNA, U2AF2 RIP RNA); SYNCRIP siRNA vs Control siRNA (Total RNA, U2AF2 RIP RNA); ILF2 siRNA vs Control siRNA (Total RNA); SRRM2 siRNA vs Control siRNA (Total RNA)
Experimental features Very brief experimental description	RNA-seq and U2AF2 RIP-seq profiling of Resting and Activated CD4 T cell culture to identify differentially expressed/spliced and U2AF2 differentially bound genes, respectively, upon T cell Activation. Microarray gene expression profiling of Activated CD4 T cell culture to identify differentially expressed/spliced transcripts upon siRNA knockdown of RNA-binding proteins (U2AF1, SYNCRIP, ILF2, and SRRM2) relative to Control siRNA. U2AF2 RIP-chip (microarray) profiling of Activated CD4 T cell culture to identify transcripts differentially bound to U2AF2 upon siRNA knockdown of RNA-binding proteins (U2AF1 and SYNCRIP) relative to Control siRNA.
Consent Level of consent allowed for reuse	Data are publically available
Sample source location City, country of model organism	La Jolla, CA, USA

1. Direct link to deposited data

The deposited data can be found at: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE62923.

2. Experimental design, materials and methods

2.1. Ethics, consent and permissions

All the studies in this manuscript were covered by Human Subjects Research Protocols approved by the Institutional Review Board of The

E-mail address: thomasw@scripps.edu.

http://dx.doi.org/10.1016/j.gdata.2016.12.006

Scripps Research Institute. Informed written consent was obtained from all study subjects.

2.2. CD4 + T cell culture and experimental conditions

CD4 + T cell cultures were generated with CD4 + T cells isolated from the peripheral blood of normal adult human female donors, as described previously [1]. Briefly, the purified cells were grown and maintained in RPMI 1640 culture medium (+ 10% FBS, 100 U/ml Penicillin, Mediatech) at 37 °C and 5% CO_2, activated with Dynabead Human T-Activator CD3/CD28 (Thermo) for 48 h, and cultured in the presence of 30 U/ml of IL2. While kept at a density of 0.5×10^6 to 2×10^6 cells/ml, the cells were expanded for 7 days and aliquoted into freezing media (90% FBS/10% DMSO) for storage at − 80 °C. Thawed aliquots were then expanded for an additional 7 days prior to experimental use.

For all sequencing experiments (Table 1), the CD4 + T cell culture was used to compare "resting" and "activated" states. After thawing and expansion, cells were activated with Dynabead Human T-Activator CD3/CD28 for 48 h to produce the "activated" state, while "resting" cells were cultured unchanged for the same 48 h window.

For all microarray experiments (Table 2), cells were transfected by Amaxa Nucleofection (Lonza) following the manufacturer's instructions (program CL-120), with specific siRNA added to a final concentration of 300 nM. Cells were then cultured for 48 h prior to activation, cultured for an additional 24 h and collected. Specific siRNA sequences and product information was previously described [1].

2.3. U2AF2 RNA binding protein immunoprecipitation (RIP)

As described previously [1], frozen cell pellets were resuspended in RIPA buffer (Sigma) supplemented with RNaseOUT (Thermo), Complete EDTA-free protease inhibitor cocktail tablets (Roche, Inc.), and phosphatase inhibitors. Lysates were sonicated, with insoluble material discarded, supplemented with enzyme inhibitors (RNAse, phosphatase, and protease), and quantified for protein. Lysates were then pre-cleared with mouse IgG-conjugated Protein G beads (Thermo) and incubated with a mouse monoclonal U2AF2 antibody (U4758, Sigma) conjugated to Protein G beads. The beads were then washed, digested with RNAse A, and RNA/Protein complexes were eluted two times.

2.4. Total RNA isolation, RNA-seq

Total RNA was purified from cell aliquots using the RNeasy purification kit (Qiagen) according to the manufacturer's protocol. DNA Digestion was performed on the columns with the RNase-Free DNase Set (Qiagen). The concentration of RNA was measured using the Qubit Fluorometer (Thermo) and RNA quality was visually inspected using the Eukaryote Total RNA Nano chip on the 2100 Bioanalyzer (Agilent). RNA-seq libraries were made using the TruSeq RNA Sample Preparation Kit (Illumina) preceded by a PolyA purification step to select only polyadenylated mRNA. Libraries were pooled (4 samples per pool) and clustered using the Illumina cBot system with TruSeq PE Cluster Kit reagents, followed by sequencing on the Illumina HiSeq 2000 system with TruSeq SBS Kit v3 reagents.

Table 1
Summary of type of RNA, replicates, library preparation method, read type, and read depth for each cell state in each sequencing experiment.

State	RNA	Replicates	Library prep	Single/paired	Read depth (mean)
Resting	Total	2	TruSeq	Paired	50×10^6
Activated	Total	2	TruSeq	Paired	50×10^6
Resting	U2AF2 RIP	6	Script-Seq v2	Single	39×10^6
Activated	U2AF2 RIP	6	Script-Seq v2	Single	33×10^6

Table 2
Summary of type of RNA, replicates, and high throughput gene expression analysis method for each siRNA knockdown experiment.

siRNA	RNA	Replicates	Gene Exp. Platform (Amp. Kit)
Control	Total	3	HTA 2.0 (WT PLUS)
U2AF1	Total	3	HTA 2.0 (WT PLUS)
SYNCRIP	Total	3	HTA 2.0 (WT PLUS)
ILF2	Total	3	HTA 2.0 (WT PLUS)
SRRM2	Total	3	HTA 2.0 (WT PLUS)
Control	U2AF2 RIP	6	HiSeq 2000 (Script-Seq v2)
U2AF1	U2AF2 RIP	3	HiSeq 2000 (Script-Seq v2)
SYNCRIP	U2AF2 RIP	3	HiSeq 2000 (Script-Seq v2)

2.5. RIP-seq

The two eluates from each of the U2AF2 RIP samples were pooled and combined with TRIZOL LS reagent to separate the RNA fraction. The RNA was washed, precipitated, and treated with the RNase Free DNase Set (Qiagen) while cleaned up on an RNeasy column. The purified RNA was used as input for Script-Seq v2 RNA-Seq Library Preparation kit (Epicentre) and the protocol was performed according to the manual with the following exceptions: 1) No PolyA purification or rRNA depletion steps were performed; 2) DNA Clean and Concentrator-5 - PCR/DNA clean columns (Zymo Research) were used to isolate the cDNA before the PCR step. Libraries were pooled (6 samples per pool) and clustered using the Illumina cBot system with TruSeq SR Cluster Kit reagents, followed by sequencing on the Illumina HiSeq 2000 system with TruSeq SBS Kit v3 reagents.

2.6. Sequence mapping and normalization

Sequencing and post-processing was performed by the TSRI Next Gen Sequencing Core Facility which delivers the raw data as FASTQ files. Initial quality assessment of FASTQ files was performed with FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). For quality control purposes, reads with a quality score below 25 ($Q < 25$) were removed using fastx_clipper from the FASTX Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html). Furthermore, reads containing sequencing adapters were trimmed using Trim Galore (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). The remaining reads were mapped to the human genome build hg19 using TopHat version 2.0.9 [2]. Read alignments with low mapping quality ($q < 20$) were removed using SAMtools (http://samtools.sourceforge.net/) and the markDuplicates function within the Picard suite (http://picard.sourceforge.net/) was used to remove optical

Table 3
Summary of batch (Run ID), replicate number, cell state, and raw and aligned reads for each sample type analyzed by sequencing.

Sequencing	Run ID	Replicate	State	Raw reads	Aligned reads
RNA	R53	1	Resting	93393572	80065567
RNA	R53	1	Activated	117843824	100589736
RNA	R53	1	Resting	114655128	97972797
RNA	R53	1	Activated	116873022	98736926
RIP	R67	1	Resting	34447112	23443371
RIP	R67	1	Activated	31498888	26314895
RIP	R80	1	Resting	32456157	24785103
RIP	R80	1	Activated	19904038	15811632
RIP	R91	1	Resting	69100755	28827670
RIP	R91	1	Activated	34153807	21342545
RIP	R91	2	Resting	43634669	11331093
RIP	R91	2	Activated	51557362	14454475
RIP	R107	1	Resting	28139551	13930714
RIP	R107	1	Activated	35096072	20586253
RIP	R107	2	Resting	28380320	22840389
RIP	R107	2	Activated	27067774	19961774

duplicates. See Table 3 for numbers of total sequenced and aligned reads for each sample.

Using the R environment (http://cran.us.r-project.org/) and open source packages available from Bioconductor (http://www.bioconductor.org/), the aligned reads within each BAM file were assigned to both the exon and gene level definitions of all known genes in hg19 as defined by the Bioconductor package TxDb.Hsapiens.UCSC.hg19.knownGene. Custom scripts were used to generate tables of counts for all samples at both the exon and gene level. For gene level counts, the values were normalized by conversion to FPKM (Fragment Per Kilobase of exon model per million mapped reads, [3]), which accounts for variation in gene size and total reads in each sample. For RIP-seq experiments, as with any type of pulldown experiment, an additional normalization step would typically entail lowering expression estimates based on subtracting read counts from a replicated specificity control (i.e. non-specific immunoglobulin for the species of the antibody). However, the mouse IgG used as a control for the U2AF2 antibody was unable to co-precipitate any RNA that could be used for generating usable sequencing data. As such, no background subtraction step was performed as part of the analysis.

2.7. RNA-seq differential gene expression and splicing

Based on previous work [4], genes with log2(FPKM) values below −2 in more than half of the samples were removed due to insufficient signal relative to the noise. The read counts for the remaining genes were input into the edgeR package [5] to measure differential gene expression (FDR < 0.05). Alternative splicing analysis was performed using AltAnalyze (http://www.altanalyze.org/) [6], which uses the ASPIRE algorithm (default threshold > 0.2) to determine differential splicing between two groups.

2.8. RIP-seq differential binding

Prior to testing for significant differences between the two groups in the RIP-seq data, a technical analysis of the data was performed to assess the extent to which a standard RNA-seq analysis pipeline could be used to interpret the RIP-seq data. As discussed in [1], the overall range of expression values is smaller in the RIP-seq data relative to RNA-seq which results in fewer detectable genes when using the empirically established RIP-seq FPKM threshold of 2. For those genes with usable data, there is consistently greater within-group (WG) variation at all levels of expression in the RIP-seq data relative to the RNA-seq data, while the between-group (BG) variation is similarly distributed (Fig. 1A, B). For the RIP-seq data, this is further demonstrated by the distribution of the log of the F-scores (BG/WG) as a function of relative expression level (Fig. 1C, blue line). For RNA-seq, noise (WG) is reduced as expression levels increase resulting in a trend of rising F-scores (Fig. 1C, red line). For the RIP-seq data, a higher noise floor coupled with the lack of noise reduction as expression increase results in a relatively lower and flat F-score distribution.

Thus, with the understanding that the RIP-seq had diminished power to detect differences between Resting and Activated CD4+ T cells relative to RNA-seq, the standard sequencing analysis pipeline was used. As such, the number of significantly different genes ("differentially bound" to U2AF2) identified by edgeR was lower in the

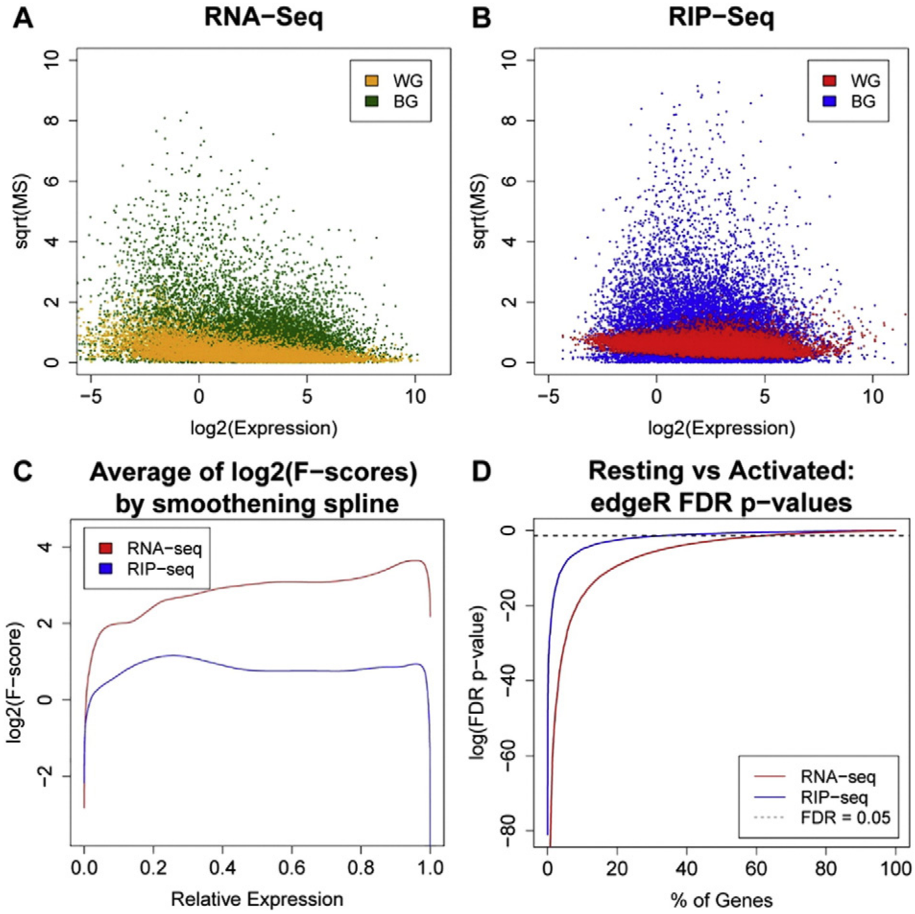

Fig. 1. A) Within group (WG, gold) and between group (BG, green) variance (square root of the mean difference [sqrt(MS)]) for each genes as a function of the log2(expression) for all genes in the RNA-seq experiment. B) Within group (gold) and between group (blue) variance as a function of the log gene expression for all genes in the RIP-seq experiment. C) Smoothened spline plot of log2(F-scores) as a function of relative expression for RNA-seq (red) and RIP-seq (blue) experiments. D) Distribution of FDR adjusted p-values for the comparison of resting and activated samples in both sequencing experiments.

RIP-seq, as demonstrated by the percentage of total genes with FDR p-values below the threshold (Fig. 1D). Furthermore, there was a lack of power to detect differences at the exon level which precluded an analysis for differential splicing. In order to make specific conclusions about alternatively spliced genes that are bound to U2AF2 during the process of T-cell activation [1], we looked at the overlap of genes that were alternatively spliced in the RNA-seq data and "differentially bound" or not "differentially bound" in the RIP-seq data.

2.9. Microarrays and RIP-Chip

To globally identify transcripts regulated by U2AF2 interacting proteins, total RNA was extracted from cells that had been treated with control or specific siRNAs. Specifically, the input into the PLUS kit (Affymetrix) was 1 μg of this DNAse treated total RNA. For RIP-Chip, a U2AF2 RNA-binding protein immunoprecipitation was performed on the U2AF1 and SYNCRIP siRNA treated cells and 125 ng of purified, DNAse treated, RNA from this procedure was used as input for the amplification protocol. Labeled, fragmented cRNA was hybridized on the Human Transcriptome Array 2.0 and scanned to generate CEL files. Three replicate experiments were performed for each group with the exception of the U2AF2 RIP from the control siRNA treated cells which was performed six times. The CEL files were input into the Affymetrix Expression Console software to generate normalized, log-transformed RMA values. Genes were filtered out of downstream analysis if the log2(RMA) expression value was <6 in 50% or more of the arrays. The remaining genes were imported into the R environment for differential expression analysis by the limma package (adjusted p-value < 0.05) [7,8]. For splicing analysis, CEL files were directly input into the AltAnalyze package and the default parameters were used to call differentially spliced genes. To plot examples of significantly differentially spliced transcripts between control siRNA and specific siRNA at the junction and exon level, Splicing Index (SI) values were calculated using the Transcript Analysis Console software (TAC, Affymetrix, Inc.).

2.10. Functional enrichment analysis

For each list of genes (i.e. differentially expressed, differentially spliced, differentially bound) identified with high throughput experiments, an analysis was performed to identify, pathways, functions, and ontological categories that were significantly enriched. The GO-Elite function (within AltAnalyze) was used to perform enrichment analysis (adjusted p-value < 0.05) on numerous ontologies and annotation databases including GeneOntology, KEGG, Biomarkers, and Wikipathways. An immune specific ontology called ImmuneMap, was used to identify significantly enriched immune related pathways and functions defined as those whose adjusted p-values were below a threshold of 0.05). The adjusted p-values were calculated using a hypergeometric distribution followed by a multiple testing correction using the Benjamini-Hochberg step-up procedure to control the false discovery rate [9].

3. Discussion

We describe the generation and analysis of multiple datasets including both Illumina deep RNA sequencing and Affymetrix GeneChip microarray experiments. Sequencing experiments measured transcript and isoform level changes in Activated human CD4+ T cells relative to Resting cells using both total RNA and RNA associated with the splicing factor U2AF2. The whole transcriptome scope of these data allowed for a complete characterization of the activation-induced changes in expression and splicing in this novel CD4+ T cell culture system. The U2AF2 RIP-seq data facilitated identification of the specific subset of genes that are post-transcriptionally regulated by U2AF2 and its interacting partners after T cell activation.

The microarray experiments measured changes in the transcriptome of Activated CD4+ T cells treated with control siRNA or siRNA for specific U2AF2 interacting proteins. The effects of knocking down U2AF1, SRRM2, ILF2, and SYNCRIP were measured with total RNA, while changes in RNA associated with U2AF2 were measured after knockdown of U2AF1 and SYNCRIP. For these data, the subset of genes affected by knockdown was uniquely characterized by the number, identity, and direction of change of the genes in the list. In addition, these characteristics in combination with the list of significantly overrepresented functional categories provided insight into the post-transcriptional regulatory role of each U2AF2 interacting protein as well as context for its interaction with U2AF2 during T cell activation.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

Dedicated to the late Daniel R Salomon. We thank N Salomonis for assistance with Alt-Analyze software; B Modena and SM Kurian for work on ImmuneMap; Eigen Peralta and the members of TSRI NGS and Microarray Core Facility for sample preparation and data collection. This work was supported by National Institute of Health —U01 GM094653 to TCW.

References

[1] T.C. Whisenant, E.R. Peralta, L.D. Aarreberg, N.J. Gao, S.R. Head, P. Ordoukhanian, et al., The activation-induced assembly of an RNA/protein interactome centered on the splicing factor U2AF2 regulates gene expression in human CD4 T cells. PLoS One 10 (12) (2015), e0144409. http://dx.doi.org/10.1371/journal.pone.0144409 (PubMed PMID: 26641092; PubMed Central PMCID: PMC4671683).

[2] C. Trapnell, L. Pachter, S.L. Salzberg, TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25 (9) (2009) 1105–1111, http://dx.doi.org/10.1093/bioinformatics/btp120 (PubMed PMID: 19289445; PubMed Central PMCID: PMC2672628).

[3] A. Mortazavi, B.A. Williams, K. McCue, L. Schaeffer, B. Wold, Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods 5 (7) (2008) 621–628, http://dx.doi.org/10.1038/nmeth.1226 (PubMed PMID: 18516045).

[4] T. Hart, H.K. Komori, S. LaMere, K. Podshivalova, D.R. Salomon, Finding the active genes in deep RNA-seq gene expression studies. BMC Genomics 14 (2013) 778, http://dx.doi.org/10.1186/1471-2164-14-778 (PubMed PMID: 24215113; PubMed Central PMCID: PMC3870982).

[5] M.D. Robinson, D.J. McCarthy, G.K. Smyth, edgeR: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26 (1) (2010) 139–140, http://dx.doi.org/10.1093/bioinformatics/btp616 (PubMed PMID: 19910308; PubMed Central PMCID: PMC2796818).

[6] D. Emig, N. Salomonis, J. Baumbach, T. Lengauer, B.R. Conklin, M. Albrecht, AltAnalyze and DomainGraph: analyzing and visualizing exon expression data. Nucleic Acids Res. 38 (2010), http://dx.doi.org/10.1093/nar/gkq405 (PubMed PMID: 20513647; PubMed Central PMCID: PMC2896198, (Web Server issue):W755-62).

[7] M.E. Ritchie, B. Phipson, D. Wu, Y. Hu, C.W. Law, W. Shi, et al., Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 43 (7) (2015), e47. http://dx.doi.org/10.1093/nar/gkv007 (PubMed PMID: 25605792; PubMed Central PMCID: PMC4402510).

[8] G.K. Smyth, Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat. Appl. Genet. Mol. Biol. 3 (2004), 3. http://dx.doi.org/10.2202/1544-6115.1027 (PubMed PMID: 16646809).

[9] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B Methodol. 57 (1) (1995) 289–300.