Supplement to: Integrative Genome-wide Analysis of the Determinants of RNA Splicing in Kidney Renal Clear Cell Carcinoma

Kjong van Lehmann,[1,*,‡], Andre Kahles,[1], ‡ Cyriac Kandoth,[1] William Lee,[1]
Nikolaus Schultz,[1] and Oliver Stegle,[2] and Gunnar Rätsch[1]

1 *Computational Biology Center, Memorial Sloan-Kettering Cancer Center, New York, NY 10044, U.S.A*
2 *European Bioinformatics Institute, Hinxton, Cambridge, CB10 1SD, United Kingdom*

1. Supplemental Methods

While most thresholds can be considered conservative and standard, we have included them here for completeness and reproducibility sake.

1.1. *Preprocessing*

1.1.1. *Read Alignment*

We identified 282 patients with Kidney Renal Clear Cell Carcinoma (KIRC) in the The Cancer Genome Atlas (TCGA) project that had a) RNA-seq data of the tumor, b) whole-exome sequencing data of blood (germline), c) tumor as well as d) copy number variation (CNV) available. We obtained the RNA-Seq samples for provided by the CGHub data portal (`https://cghub.ucsc.edu/`). As only BAM files were available at the time the data was obtained, a custom script was used to extract read information from the given alignment files and generate valid FASTQ files for re-alignment (two FASTQ files with the reads in the same order such that the aligner can align them in pairs). All reads were aligned to the hg19 human reference genome using the alignment software STAR (version Alpha 2.2.0)[1] for computation time reasons. The annotation was enriched with splice junction information extracted from the GENCODE annotation (version 14).[2] We used the following options: `--outFilterMultimapScoreRange 2 --outFilterMultimapNmax 100 --outFilterMismatchNmax 10 --alignIntronMax 500000 --alignMatesGapMax 1000000 --sjdbScore 1 --sjdbOverhang 5`.

We further obtained whole exome sequencing samples from normal (blood) and tumor tissue taken from the same individuals. We followed the same strategy to use a custom script to convert the downloaded alignment files from BAM into FASTQ format. Alignments of all reads against the human hg19 reference genome were again performed using STAR. For this alignment, STAR was configured not to identify introns. We used the following additional options: `--sjdbScore -10 --scoreGapNoncan 0 --scoreGap -8 --alignIntronMin 100 --scoreGapGCAG 0 --scoreGapATAC 0 --scoreStitchSJshift 0 --alignIntronMax 0 --alignSJoverhangMin 200 --alignSJDBoverhangMin 200`.

In addition to the TCGA data files we also downloaded a subset (140) of the RNA-Seq samples that were generated as part of GEUVADIS[3] as well as 460 RNA-Seq samples generated in context of ENCODE.[4] We aligned these samples using the same STAR settings as were used for the TCGA RNA-Seq samples.

1.2. *Expression Estimation and Alternative Event Extraction*

Based on the RNA-Seq alignments, we used a custom python script to count expression information.All genome coordinates used for expression counting were based on the GENCODE annotation (version 14). We used SplAdder[5] to augment this annotation with additional information from the RNA-Seq samples. The code for SplAdder is publicly available (`https://github.com/ratschlab/spladder`). Using the augmented annotation, we counted an alignment towards the expression of a gene, if it shared at least one position with any exon of the gene. The expression counts were library size normalized using the expression counts of all genes that had a read count larger than 10 in at most 95% of the samples as a size factor.

Using an adapted version of SplAdder, we extracted alternative splicing events from the augmented annotation. To this end, the annotation was represented as a splicing graph and events were sampled as subsets of this graph. We only focused on alternative 3'- and 5'-splice site events as well as on exon skip events. To identify a high-confidence subset, we only retained events that had at least 10 reads spanning any of the event's introns in at least 80% of the samples. We further required that an event was alternative, that is both isoforms were expressed, in at least 1% of the samples. As a representation of relative isoform expression of an event that could be used as phenotype for the association study, we used the splicing index (PSI),[6] which was computed as follows:

- For exon skips as $PSI = \frac{a+b}{a+b+2c}$, where a and b are the number of alignments spanning the introns directly before and after the cassette exon, respectively, and c is the number of alignments confirming the skipping of the cassette exon.
- For alternative 3'/5'-events as $PSI = \frac{a}{a+b}$, where a is the number of alignments spanning the intron in the shorter event isoform and b is the number of alignments spanning the intron in the longer event isoform.

A robustness analysis of SplAdder can be found in[5] Supplemental Table 20, demonstrating reproducibility of splicing events under different thresholds.

1.3. *Tumor-specific splicing analysis*

Further, to identify tumor-specific splicing, we ranked all expressed genes by the ratio of the average number of samples that expressed a certain intron in KIRC tumor over the average number of samples expressing the intron in KIRC normals, GEUVADIS and ENCODE combined. To prevent division by zero, we added a pseudocount to the mean.

1.4. *Enrichment analysis*

The ranked list obtained from the tumor specific splicing analysis has been used for a functional enrichment analysis of gene ontology (GO) terms using the GOrilla webserver.[7]

1.4.1. *Variant Calling*

We have used the HaplotypeCaller in GATK (version 3.1.1)[8,9] to create gVCF files for each new BAM file. We followed the good-practice guidelines for variant calling with GATK and used these options: `-stand_call_conf 50.0 -stand_emit_conf 10.0 -ERC GVCF --variant_index_type LINEAR --variant_index_parameter 128000 -mbq 15 -dcov 200 --minPruning 5 -pairHMM VECTOR_LOGLESS_CACHING -S STRICT`. We then performed joint variant calling across normal and tumor samples separately using the downsampling option (`-dcov 300`).

In addition to the joint calling, we also generated calls of somatic variants using the MuTect package (version 1.1.5)[10] on 282 matching tumor normal pairs of exome sequencing samples. We retained all variants labeled as `PASS` by MuTect.

1.5. ***Quantitative Trait Analysis***

1.5.1. *Preprocessing*

The splicing index is being used as a quantitative phenotype. In order to address some unwanted properties of this phenotype we have performed an inverse normal transform on all splicing indices estimated by SplAdder. Assuming there are no ties, this transformation will address concerns about deviations from the normal distribution. However, ties in the rankings can create multi-modalities in the data and introduce an increase of false positives. To resolve such ties, we have added a small amount of random pseudo-noise in the range of 10^{-5} to each estimate before transformation. Splicing events which exclusively exhibited ties (e.g.: no variation), have been removed from analysis. We also excluded phenotypes in which less than 10% of the samples had any valid estimates. These steps are very conservative and lead to a decrease in sensitive, however they ensure a low false positive rate, which is desired in this analysis.

Tumor variant calls have been filtered in the following way: Variants that have less than 100 samples with valid calls, quality of less than 100, are multi-allelelic or indels have been removed from analysis. We further required more than 5 alternate alleles for each polymorphic position. All variants have been encoded into an additive scheme with 0 representing the homozygous reference state, 1 the heterozygous state and 2 the homozygous alternate allele. In this study we ignore the existence of variants that appear subclonally. The encoding has been standardized and centered. Finally, we have intentionally avoided any imputation of missing variant calls since somatic variants do not follow the typical linkage patterns. Instead we have done a trivial imputation by adding random variant calls according to the observed freqencies in the remaining samples.

MuTect calls have been filtered to `PASS` calls only and non-recurring (e.g., uniquely called on only one sample) variants have been removed. We used a binary encoding to represent the detection status of a somatic variant. All variant positions were centered and standardized.

1.5.2. *Mixed Model Analysis*

We have used a linear mixed model to perform our QTL analysis for splicing phenotypes. For the analysis of the GATK tumor calls our model takes into account hidden confounder using PANAMA[11] and population structure as random effects. Copy number variation (where available) from Ciriello *et al.*[12] and gene expression have been used as fixed effects in the sQTL analysis. More formally:

$$Y = \beta_0 + \beta_s x_s + \beta_c x_c + \beta_e x_e + \kappa + \pi + \epsilon \qquad (1)$$

with $\kappa \sim N(0, \sigma_k^2 K)$, $\pi \sim N(0, \sigma_p^2 P)$, $\epsilon \sim N(0, \sigma_\epsilon^2 I)$ where K is the exome wide kinship matrix, P is a kernel matrix estimated using PANAMA to account for hidden confounders (see Section 1.5.3), x_c are copy number variations from Ciriello *et al.*[12] and x_e is the gene expression estimate as described above. Genetic effects are then tested using a Likelihood ratio test where $H_0 : \beta_s = 0$. We have used LIMIX[13] in order to efficiently perform this analysis across all samples and single nucleotide variants (SNVs). Benjamini-Hochberg step-up procedure[14] has been used for FDR estimation for all SNVs per event.

For the analysis of the somatic variant call we employed the same model, however replacing the population structure random effect with a separate factor accounting for cancer heterogeneity. Thus the new model is

$$Y = \beta_0 + \beta_s x_s + \beta_c x_c + \beta_e x_e + \omega + \pi + \epsilon \tag{2}$$

with $\omega \sim N(0, \sigma_s^2 S)$ where S is a somatic genetic relatedness matrix based on MuTect calls.

1.5.3. *Kinship Calculation and PANAMA*

In order to account for confounding effects due to systematic genetic similarities and population structure we have estimated the kinship matrix K as

$$K = X^T X/n \tag{3}$$

where X is the exome-wide additive genotype matrix and n the number of loci. P has been estimated directly via PANAMA[11] based on the library size normalized expression counts.

1.5.4. *Integrative Analysis*

We have used annotations from the UCSC genome browser in order to annotate all sQTLs. We integrated ClinVar[15] data, COSMIC[16] data as well as other known GWAS loci.

2. Expression Confounding to Cancer-specific Introns

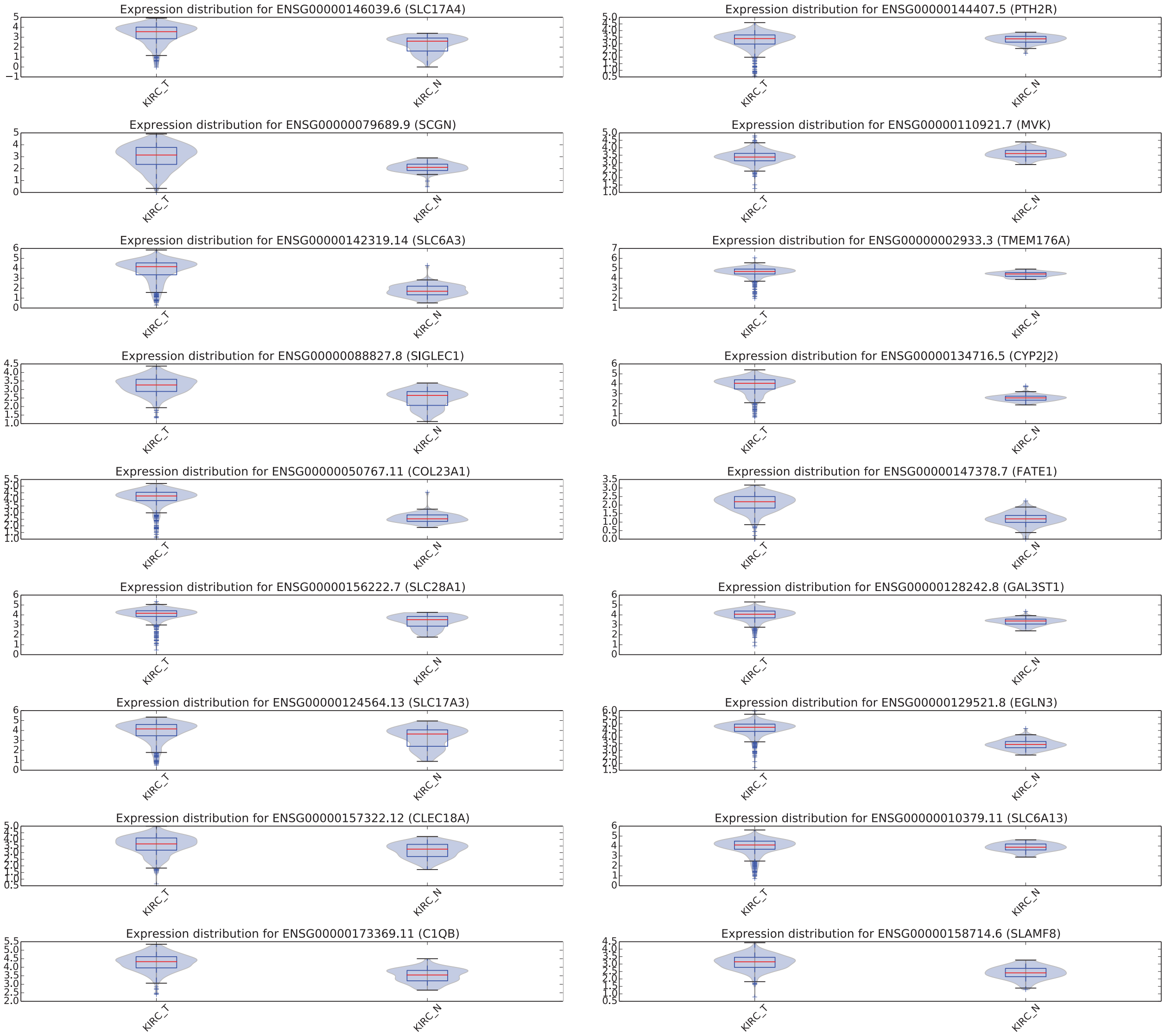

Fig. 1. Gene expression distribution in tumor and normal KIRC samples for the top ranked genes with cancer-specific introns. Expression between sampls was normalized by mean of the total counts of all genes. Expression counts are shown as $\log_{10}$ counts (y-axis) for KIRC normal/tumor (x-axis).

References

1. A. Dobin, C. a. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson and T. R. Gingeras, *Bioinformatics* **29**, 15 (January 2013).
2. J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, I. Barnes, A. Bignell, V. Boychenko, T. Hunt, M. Kay, G. Mukherjee, J. Rajan, G. Despacio-Reyes, G. Saunders, C. Steward, R. Harte, M. Lin, C. Howald, A. Tanzer, T. Derrien, J. Chrast, N. Walters, S. Balasubramanian, B. Pei, M. Tress,

J. M. Rodriguez, I. Ezkurdia, J. van Baren, M. Brent, D. Haussler, M. Kellis, A. Valencia, A. Reymond, M. Gerstein, R. Guigo and T. J. Hubbard, *Genome Research* **22**, 1760 (2012).

3. T. Lappalainen, M. Sammeth, M. R. Friedlander, P. A. t Hoen, J. Monlong, M. A. Rivas, M. Gonzalez-Porta, N. Kurbatova, T. Griebel, P. G. Ferreira, M. Barann, T. Wieland, L. Greger, M. van Iterson, J. Almlof, P. Ribeca, I. Pulyakhina, D. Esser, T. Giger, A. Tikhonov, M. Sultan, G. Bertier, D. G. MacArthur, M. Lek, E. Lizano, H. P. Buermans, I. Padioleau, T. Schwarzmayr, O. Karlberg, H. Ongen, H. Kilpinen, S. Beltran, M. Gut, K. Kahlem, V. Amstislavskiy, O. Stegle, M. Pirinen, S. B. Montgomery, P. Donnelly, M. I. McCarthy, P. Flicek, T. M. Strom, C. Geuvadis, H. Lehrach, S. Schreiber, R. Sudbrak, A. Carracedo, S. E. Antonarakis, R. Hasler, A. C. Syvanen, G. J. van Ommen, A. Brazma, T. Meitinger, P. Rosenstiel, R. Guigo, I. G. Gut, X. Estivill and E. T. Dermitzakis, *Nature* **501**, 506 (2013).
4. ENCODE Project Consortium, *Science* **306**, 636 (October 2004).
5. X. Gan, O. Stegle, J. Behr, J. G. Steffen, P. Drewe, K. L. Hildebrand, R. Lyngsoe, S. J. Schultheiss, E. J. Osborne, V. T. Sreedharan, A. Kahles, R. Bohnert, G. Jean, P. Derwent, P. Kersey, E. J. Belfield, N. P. Harberd, E. Kemen, C. Toomajian, P. X. Kover, R. M. Clark, G. Rätsch and R. Mott, *Nature* **108**, 10249 (August 2011).
6. E. T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth and C. B. Burge, *Nature* **456**, 470 (November 2008).
7. E. Eden, R. Navon, I. Steinfeld, D. Lipson and Z. Yakhini, *BMC Bioinformatics* **10**, p. 48 (January 2009).
8. M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler and M. J. Daly, *Nature Genetics* **43**, 491 (2011).
9. G. A. Van der Auwera, M. O. Carneiro, C. Hartl, R. Poplin, G. del Angel, A. Levy-Moonshine, T. Jordan, K. Shakir, D. Roazen, J. Thibault, E. Banks, K. V. Garimella, D. Altshuler, S. Gabriel and M. A. DePristo, *Current Protocols in Bioinformatics* **Chapter 11**, p. UNIT 11.10 (2013).
10. K. Cibulskis, M. S. Lawrence, S. L. Carter, A. Sivachenko, D. Jaffe, C. Sougnez, S. Gabriel, M. Meyerson, E. S. Lander and G. Getz, *Nature Biotechnology* **31**, 213 (2013).
11. N. Fusi, O. Stegle and N. D. Lawrence, *PLoS Computational Biology* **8**, p. e1002330 (January 2012).
12. G. Ciriello, M. L. Miller, B. A. Aksoy, Y. Senbabaoglu, N. Schultz and C. Sander, *Nature Genetics* **45**, 1127 (September 2013).
13. C. Lippert, F. Casale, B. Rakitsch and O. Stegle, *bioRxiv* , 0 (2014).
14. Y. Benjamini and Y. Hochberg, *Journal of the Royal Statistical Society. Series B (Methodological)* **51**, 289 (1995).
15. M. J. Landrum, J. M. Lee, G. R. Riley, W. Jang, W. S. Rubinstein, D. M. Church and D. R. Maglott, *Nucleic Acids Research* **42**, D980 (January 2014).
16. S. A. Forbes, N. Bindal, S. Bamford, C. Cole, C. Y. Kok, D. Beare, M. Jia, R. Shepherd, K. Leung, A. Menzies, J. W. Teague, P. J. Campbell, M. R. Stratton and P. A. Futreal, *Nucleic Acids Research* **39**, D945 (January 2011).