Cell Reports, Volume *23*

Supplemental Information

Increased RNA Editing May Provide a Source for Autoantigens in Systemic Lupus Erythematosus

Shalom Hillel Roth, Miri Danan-Gotthold, Meirav Ben-Izhak, Gideon Rechavi, Cyrille J. Cohen, Yoram Louzoun, and Erez Y. Levanon

Supplemental Information

Supplemental Figures

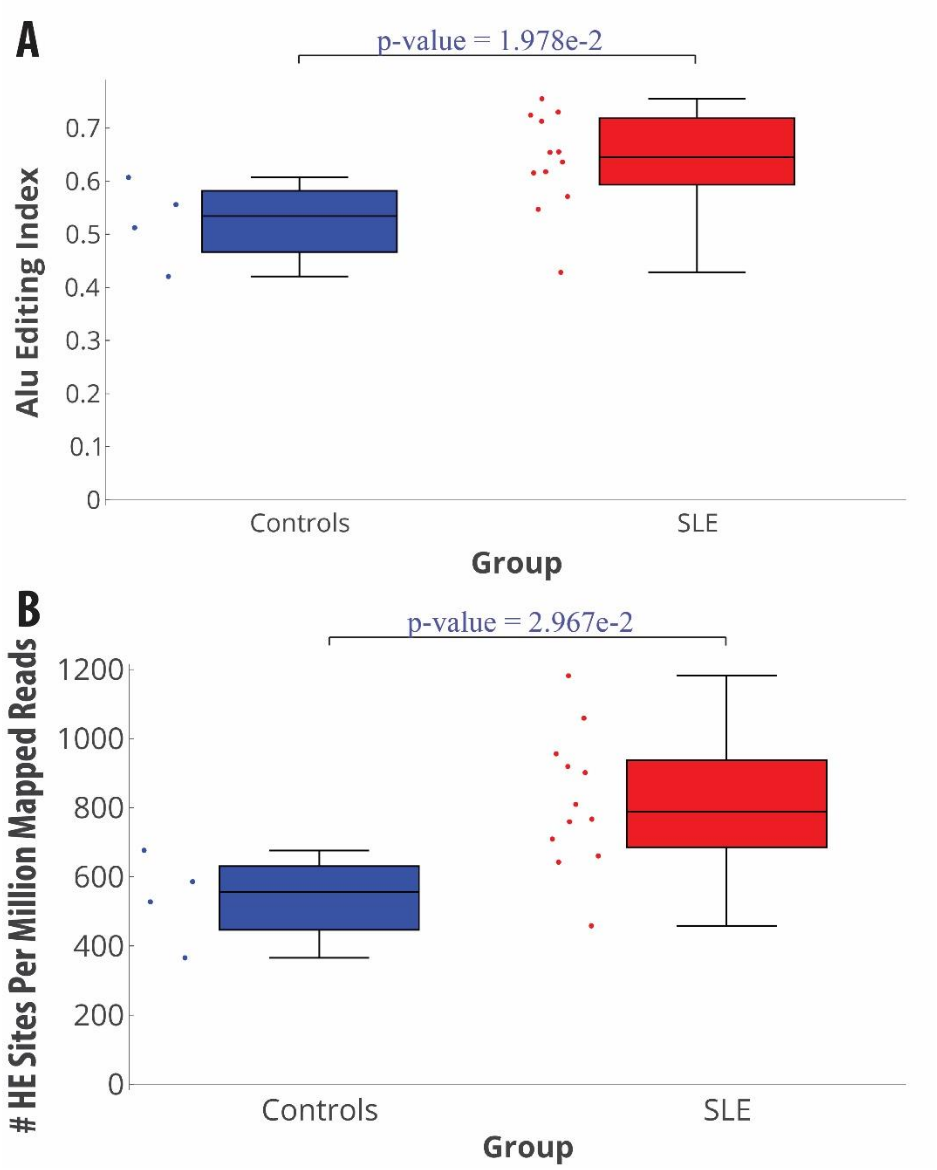

Figure S1. Editing is significantly increased in SLE patients (dataset taken from (Rai et al., 2016)), Related to Figure 1. The compared cohorts are healthy controls (blue) and SLE patients (red). Paired comparisons were evaluated using the Wilcoxon test. The global levels of editing were assessed by determining (A) the editing in *Alu* elements (*Alu* Editing Index - AEI) and (B) the number of hyper edited (HE) sites (normalized by the number of mapped reads) per sample.

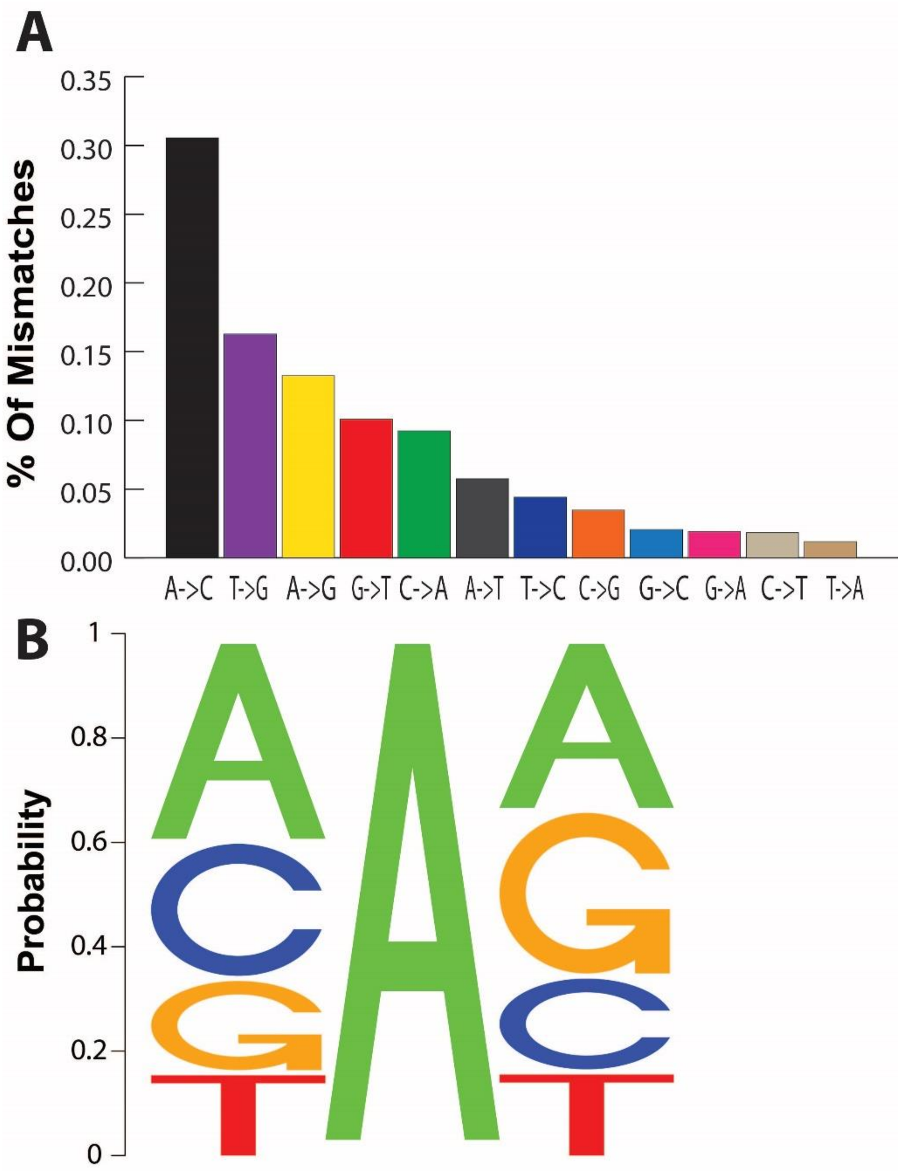

Figure S2. The signal cleanness of recoding sites at hyper edited regions, Related to Figure 3. (A) Percentage of all mismatches and (B) the nucleotide frequencies of the neighboring nucleotides around the A-to-G sites

Supplemental Tables

Table S2. Elevated APOBEC proteins levels in SLE, Related to Figure 2.

Gene name	**Fold Change between ISM-high patients and controls**	**Wilcoxon p-value**
APOBEC3A	2.58	5.25e-08
APOBEC3B	2.48	0.00237
APOBEC3F	1.80	1.96e-08
APOBEC3G	1.48	4.28e-08
APOBEC3H	1.84	0.000885

Supplemental Experimental Procedures

Reads alignment

The reads were aligned to the human genome (hg19) using STAR 2.4.2 (Dobin et al., 2013), using alignment parameters that limit the number of mismatches to 0.1 of the mapped length and accept only uniquely aligned reads (default parameters except alignIntronMax - 1000000, alignMatesGapMax - 1000000, outFilterMismatchNmax - 999, outFilterMismatchNoverReadLmax - 0.1, outFilterMultimapNmax - 1, outReadsUnmapped - Fastx, quantMode – GeneCounts). This resulted in three output files for each sample (an alignment file, a fastq file containing all the unmapped reads, and a file containing the number of reads per gene) that were used in the next analyses.

Hyper Editing

RNA editing by ADAR1 usually happens in clusters. Thus, many heavily edited reads may differ from the DNA to the extent that standard schemes fail to align them correctly (Carmi et al., 2011). A recently described pipeline enables the measurement of such heavily edited reads (Porath et al., 2014). It was observed that HE is closely correlated with ADAR1 activity (Paz-Yaacov et al., 2015; Porath et al., 2014). In this approach, the reference genome and the reads are transformed, changing all As to Gs in order to map the reads.

We extracted HE levels for every sample (this reflects the general level of editing in the sample (Porath et al., 2014)), as well as a list of the sites at which HE was detected. As input to the pipeline we used only reads that were not mapped to the genome (typically 10%-15% of reads per sample). We used the suggested parameters and Hg19 genome, except for the fraction of A-to-G out of all mismatches, which was set to 0.8. The numbers of hyper-edited sites were normalized by the number of the mapped reads in each sample. Two samples in which we identified a very high rate of clustered A-to-C substitutions (more than 70% of all clustered substitutions found in the sample) were excluded from this analysis and from all further calculations using sites detected by it, since A-to-C clustered substitutions were previously shown to result from an Illumina machine artifact (Dohm et al., 2008).

***Alu* editing index**

As most of the editing sites in the human genome (over 90%) are concentrated in *Alu* genomic regions, AEI gives a good approximation of the global editing level and ADAR1 editing efficiency in the sample (Bazak et al., 2014a).

Reads that were aligned to *Alu* elements were collected and mismatches between them and the reference genome (hg19) were identified. To get a reliable collection of mismatches, mismatches with a Phred score lower than 30 or those reported as SNPs in dbSNP (SNP build 135) were filtered out (Eisenberg et al., 2005). To eliminate putative sequence errors the results were further filtered using a probabilistic model assuming an a priori sequencing error rate of 0.001 (corresponding to Phred score of 30). This was achieved by applying Benjamini-Hochberg multiple testing correction for all *Alu* adenosines, with a false discovery rate (FDR) of 0.05. Over 90% of all *Alu* evaluated had A-to-G mismatch dominance (i.e. that the relative part of A-to-G mismatches is higher than all others combined), which further supports the cleanness of the signal. However, for the calculations of the editing levels we used all *Alu* elements (to avoid biases caused by choosing dominant A-to-G *Alu*s only) as described below .

A straight forward comparison of editing levels, based on the editing rates at all sites, is quite complicated and requires ultra-high coverage due to the fact that most adenosines in *Alu* elements are edited to at least some extent (though usually to a low degree, less than 1% of adenosines) (Bazak et al., 2014a, 2014b). Hence, any set of detected sites in a given sample may represent only a small random fraction of the editing sites. Thus, the number of reads supporting a specific mismatch is not a good measure of the effect in the whole sample.

To apply all the above to our data set, we used the AEI (*Alu* Editing Index). This index gives the average editing level across all adenosines in *Alu* elements weighted by their expression, which is the ratio of A-to-G mismatches (presumably due to deamination to inosines), over the total number of nucleotides aligned to an *Alu* repeat (both adenosines and presumed inosines). This index averages over millions of adenosines, which as previously described, (Bazak et al., 2014b) makes it rather robust to statistical noise.

Recoding Editing Sites Analyses

The examination of editing rates at non-synonymous, pre-known editing sites was limited to 60 sites, which were expressed by at least 10 samples in each group, so that n≥10 for each group of the Wilcoxon test.

Protein prediction was assessed using a list of edited sites filtered from known SNPs (dbSNP135 (Sherry et al., 2001) and highly polymorphic HLA genes, but with conserved editing sites (Pinto et al., 2014) included). The amino acid sequence (as found in the NCBI database) was queried and the "mutated" AA was planted in the right place of the sequence generating peptides of 21 AA long (where the edited AA is in the middle). The affinity of each peptide in the couple (edited and non-edited form) for the predicted HLA alleles of the samples were evaluated using NetMHCPan 4.0 (Jurtz et al., 2017) for K-mers (window) of between 8-11 AA long.

Expression analysis

The gene annotations for the expression analysis were obtained from the UCSC gene tables (Karolchik et al., 2003). The number of reads aligned to each gene was calculated by STAR.

Number of Epitopes Prediction

The number of epitopes per peptide chain was calculated by combining three algorithms: a proteasomal cleavage algorithm (Ginodi et al., 2008), a TAP binding algorithm (Peters et al., 2003), and the MLVO MHC binding algorithm (Vider-Shalit and Louzoun, 2011) as previously described in (Agranovich et al., 2013).

Since the used algorithms only apply to the 39 most common alleles in Caucasian population, the epitopes were computed for the 35 HLA alleles of the relevant samples. The results were weighted according to the allele frequency in the global human population. The efficiency of these algorithms and their quality has been systematically validated (previously) vs. epitope databases and was found to induce low false positive (FP) and false negative (FN) error rates (Maman et al., 2011a, 2011b). The number of epitopes was calculated for both the sequence and the average of three randomly scrambled versions of it, in order to assess whether the number of epitopes is affected by the sequence of the peptide or the AA composition (a biochemical effect).

Supplemental References

Agranovich, A., Maman, Y., and Louzoun, Y. (2013). Viral proteome size and CD8+ T cell epitope density are correlated: The effect of complexity on selection. Infect. Genet. Evol. *20*, 71–77.

Bazak, L., Levanon, E.Y., and Eisenberg, E. (2014a). Genome-wide analysis of Alu editability. Nucleic Acids Res. *42*, 6876–6884.

Bazak, L., Haviv, A., Barak, M., Jacob-Hirsch, J., Deng, P., Zhang, R., Isaacs, F.J., Rechavi, G., Li, J.B., Eisenberg, E., et al. (2014b). A-to-I RNA editing occurs at over a hundred million genomic sites, located in a majority of human genes. Genome Res. *24*, 365–376.

Boehm, U., Klamp, T., Groot, M., and Howard, J.C. (1997). CELLULAR RESPONSES TO INTERFERON-γ. Annu. Rev. Immunol. *15*, 749–795.

Carmi, S., Borukhov, I., and Levanon, E.Y. (2011). Identification of widespread ultra-edited human RNAs. PLoS Genet. *7*, e1002317.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: Ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Dohm, J.C., Lottaz, C., Borodina, T., and Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res. *36*, e105.

Eisenberg, E., Adamsky, K., Cohen, L., Amariglio, N., Hirshberg, A., Rechavi, G., and Levanon, E.Y. (2005). Identification of RNA editing sites in the SNP database. Nucleic Acids Res. *33*, 4612–4617.

Ferrington, D.A., and Gregerson, D.S. (2012). Immunoproteasomes: structure, function, and antigen presentation. Prog. Mol. Biol. Transl. Sci. *109*, 75–112.

Ginodi, I., Vider-Shalit, T., Tsaban, L., and Louzoun, Y. (2008). Precise score for the prediction of peptides cleaved by the proteasome. Bioinformatics *24*, 477–483.

Gommans, W.M., Tatalias, N.E., Sie, C.P., Dupuis, D., Vendetti, N., Smith, L., Kaushal, R., and Maas, S. (2008). Screening of human SNP database identifies recoding sites of A-to-I RNA editing. RNA *14*, 2074–2085.

Groettrup, M., Kirk, C.J., and Basler, M. (2010). Proteasomes in immune cells: more than peptide producers? Nat. Rev. Immunol. *10*, 73–78.

Jurtz, V., Paul, S., Andreatta, M., Marcatili, P., Peters, B., and Nielsen, M. (2017). NetMHCpan-4.0: Improved Peptide–MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. J. Immunol. *199*, 3360–3368.

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J., et al. (2003). The UCSC Genome Browser Database. Nucleic Acids Res. *31*, 51–54.

Lipsky, P.E. (2001). Systemic lupus erythematosus: an autoimmune disease of B cell hyperactivity. Nat. Immunol. *2*, 764–766.

Maman, Y., Blancher, A., Benichou, J., Yablonka, A., Efroni, S., and Louzoun, Y. (2011a). Immune-Induced Evolutionary Selection Focused on a Single Reading Frame in Overlapping Hepatitis B Virus Proteins. J. Virol. *85*, 4558–4566.

Maman, Y., Nir-Paz, R., and Louzoun, Y. (2011b). Bacteria modulate the CD8+ T cell epitope repertoire of host cytosol-exposed proteins to manipulate the host immune response. PLoS Comput. Biol. *7*, e1002220.

Paz-Yaacov, N., Bazak, L., Buchumenski, I., Porath, H.T., Danan-Gotthold, M., Knisbacher, B.A., Eisenberg, E., and Levanon, E.Y. (2015). Elevated RNA Editing Activity Is a Major Contributor to Transcriptomic Diversity in Tumors. Cell Rep. *13*, 267–276.

Peters, B., Bulik, S., Tampe, R., van Endert, P.M., and Holzhutter, H.-G. (2003). Identifying MHC Class I Epitopes by Predicting the TAP Transport Efficiency of Epitope Precursors. J. Immunol. *171*, 1741–1749.

Pinto, Y., Cohen, H.Y., and Levanon, E.Y. (2014). Mammalian conserved ADAR targets comprise only a small fragment of the human editosome. Genome Biol. *15*, R5.

Porath, H.T., Carmi, S., and Levanon, E.Y. (2014). A genome-wide map of hyper-edited RNA reveals numerous new sites. Nat. Commun. *5*, 4726.

Rai, R., Chauhan, S.K., Singh, V.V., Rai, M., and Rai, G. (2016). RNA-seq Analysis Reveals Unique Transcriptome Signatures in Systemic Lupus Erythematosus Patients with Distinct Autoantibody Specificities. PLoS One *11*, e0166312.

Reichenberger, E.J., Levine, M.A., Olsen, B.R., Papadaki, M.E., Lietman, S.A., Ohnishi, T., Matsuguchi, T., Shimokawa, H., Ohya, K., Amagasa, T., et al. (2012). The role of SH3BP2 in the pathophysiology of cherubism. Orphanet J. Rare Dis. *7*, S5.

Rusinova, I., Forster, S., Yu, S., Kannan, A., Masse, M., Cumming, H., Chapman, R., and Hertzog, P.J. (2013). INTERFEROME v2.0: an updated database of annotated interferon-regulated genes. Nucleic Acids Res. *41*, D1040–D1046.

Schroder, K., Hertzog, P.J., Ravasi, T., and Hume, D.A. (2004). Interferon-γ : an overview of signals, mechnisms and functions. J. Leukozyte Biol. *75*.

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. *29*, 308–311.

Vider-Shalit, T., and Louzoun, Y. (2011). MHC-I prediction using a combination of T cell epitopes and MHC-I binding peptides. J. Immunol. Methods *374*, 43–46.