RNA Editing Project

In Silico Analysis Pipeline of RNA Motifs

Anna Sim
PhD Student Eric Kofman
Advisor Dr. Eran Mukamel

Outline

- Background
- Research Question
- Data Description
- Pipeline
- Results
- Next Steps

Background

Many physiological and functional aspects of RNA editing are not well understood

Distinguishing true RNA editing sites is challenging due to:
- Genomic Variation [4]
- Low frequency[2]
- Various types of modifications (site-selective/ hyper-editing) [2]
- Sequencing Artifacts
- PCR Errors

By examining the motifs and providing evidence that discovered RNA edit sites are bona fide ADAR-edited sites, the project aims to quality control the filtering approach of MARINE software

Dataset Description

- 123 samples of wild-type and knockouts linked to Autism Spectrum Disorder
- Databases: REDIportal, dbSNP, Craig Venter HuRef SNPs

Example of MARINE Output:

site_id	barcode	contig	position	ref	alt	strand	count	coverage	conversion	feature_nam	feature_strar	feature_type	feature_conv
AACCGCGC/	AACCGCGC/	chr14	95544529	A	G	+	4	4	A>G	GLRX5	+	protein_codi	A>G
AACCGCGC/	AACCGCGC/	chr2	27067657	A	G	+	1	1	A>G	AGBL5	+	protein_codi	A>G
AACCGCGC/	AACCGCGC/	chr17	1783116	T	C	+	1	1	T>C	SMYD4	-	protein_codi	A>G
AACCGCGC/	AACCGCGC/	chr7	12233526	A	G	+	1	1	A>G	TMEM106B	+	protein_codi	A>G
AACCGCGC/	AACCGCGC/	chr1	37493849	T	C	+	1	1	T>C	MEAF6	-	protein_codi	A>G
AACCGCGC/	AACCGCGC/	chr4	82893670	T	C	+	1	1	T>C	THAP9-AS1	-	lncRNA	A>G
AACCGCGC/	AACCGCGC/	chr4	82893670	T	C	+	1	1	T>C	SEC31A	-	protein_codi	A>G
AACCGCGC/	AACCGCGC/	chr1	230869142	T	C	+	1	1	T>C	C1orf198	-	protein_codi	A>G
AACCGCGC/	AACCGCGC/	chr1	8012028	T	C	+	1	2	T>C	ERRFI1	-	protein_codi	A>G
AACCGCGC/	AACCGCGC/	chr2	112756248	T	C	+	1	1	T>C	CKAP2L	-	protein_codi	A>G
AACCGCGC/	AACCGCGC/	chr12	10847109	T	C	+	1	1	T>C	PRR4	-	protein_codi	A>G
AACCGCGC/	AACCGCGC/	chr12	10847109	T	C	+	1	1	T>C	PRH1	-	protein_codi	A>G
AACCGCGC/	AACCGCGC/	chr1	92837557	A	G	+	6	11	A>G	RPL5	+	protein_codi	A>G

Pipeline

get_stream.py

class FileLoader(sample.tsv, fasta.fa)

class SequenceMatcher(contig, position, feature_strand, num_neighbor)

class NucleotideCounter(expanded_stream):

generate_logo.py

folder_path
- file_1.csv
- file_2.csv
- file_3.csv

class FileLoader(folder_path, output_path)

class FigureGenerator(normalized_df, output_path, title)

file.csv

A	C	G	T
287045	260170	238542	181899
967656	0	0	0
247839	175978	391569	152270

example.png

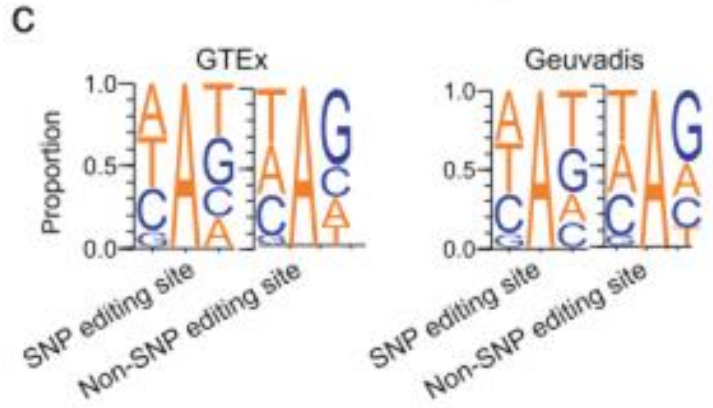
References

RESEARCH Open Access

Human A-to-I RNA editing SNP loci are enriched in GWAS signals for autoimmune diseases and under balancing selection

Hui Zhang[1,2†], Qiang Fu[1†], Xinrui Shi[1†], Ziqing Pan[1], Wenbing Yang[1], Zichao Huang[1], Tian Tang[3], Xionglei He[1] and Rui Zhang[1,4*]

The nucleotides neighboring both the non-SNP and SNP editing sites show a pattern consistent with known ADAR preference. The motif is characterized by the underrepresentation of G upstream to the editing site.



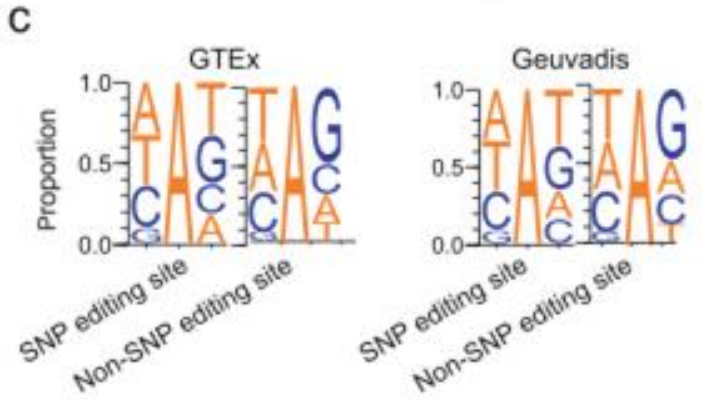
REVIEW Open Access

Rewriting the transcriptome: adenosine-to-inosine RNA editing by ADARs

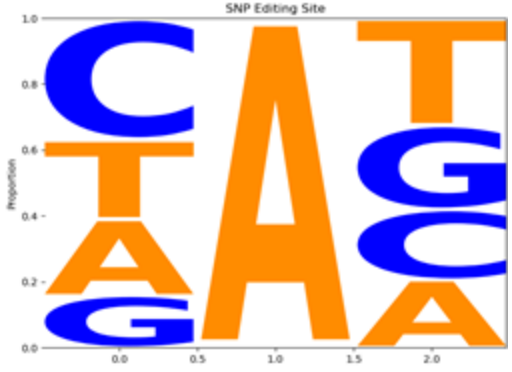
Carl R. Walkley[1,2*] and Jin Billy Li[3*]

ADAR has a preferred sequence motif neighboring the targeted adenosine, in particular the 5′ and 3′ nearest neighboring positions to the editing site, with the depletion and enrichment of G upstream and downstream of the editing site, respectively [50, 112, 113].

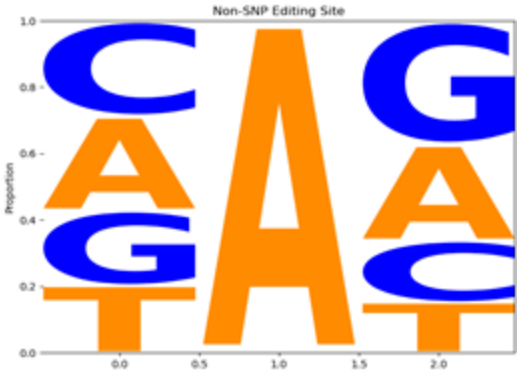
Results



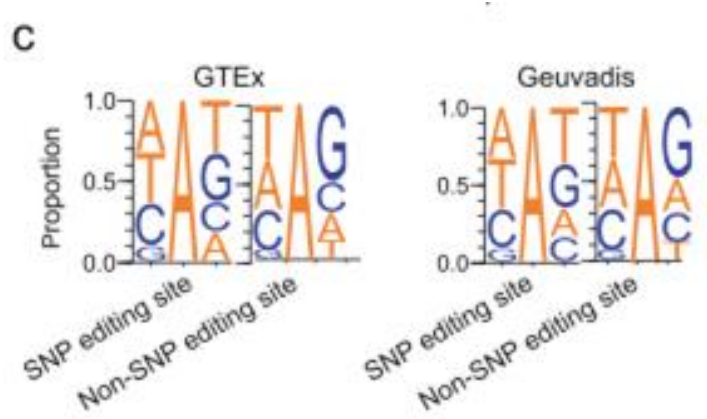
SNP Editing Site



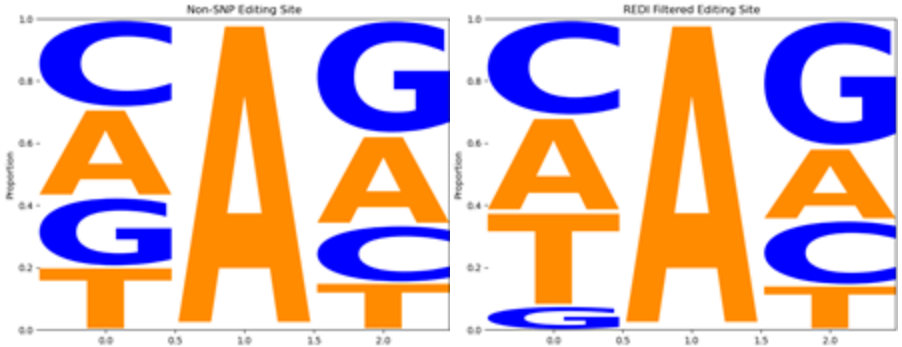
Non-SNP Editing Site

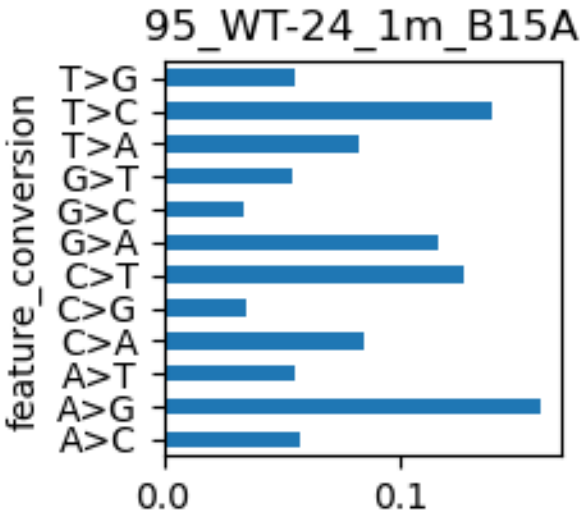


Results

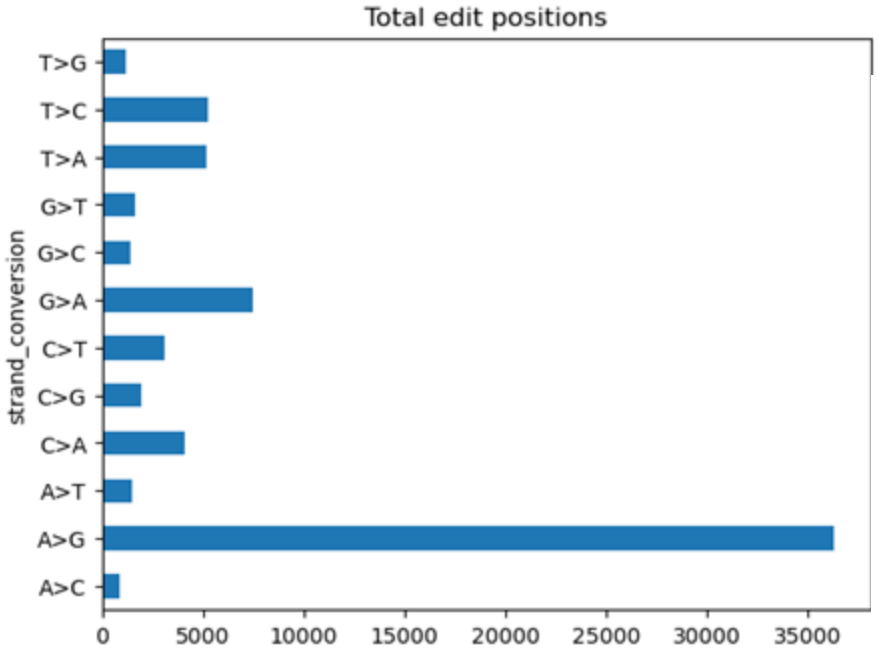


Non-SNP Editing Site REDI Filtered Editing Site





Results

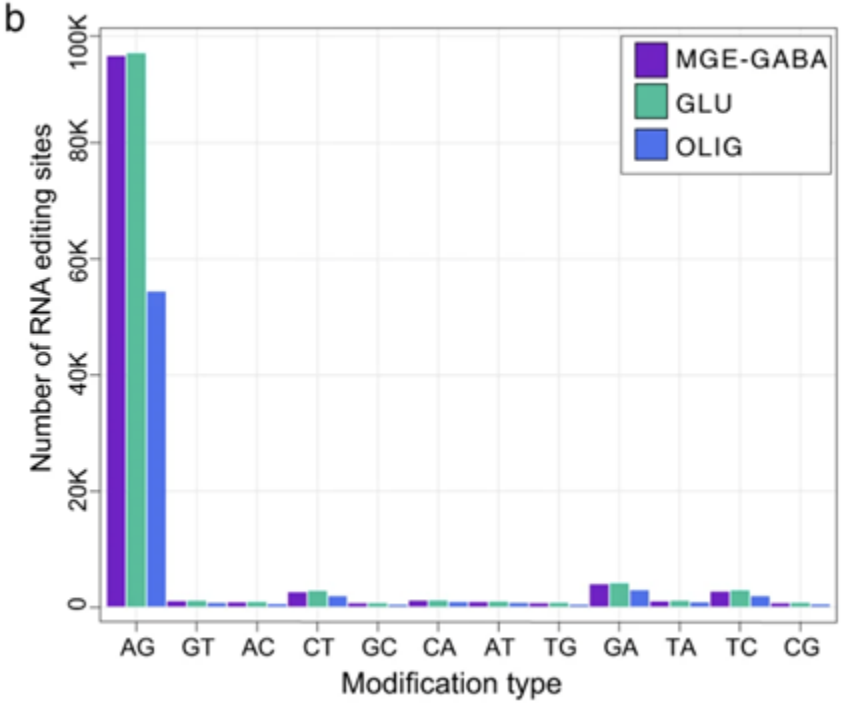


Article | Open access | Published: 30 May 2022

Cellular and genetic drivers of RNA editing variation in the human brain

Winston H. Cuddleston, Junhao Li, Xuanjia Fan, Alexey Kozenkov, Matthew Lalli, Shahrukh Khalique, Stella Dracheva, Eran A. Mukamel & Michael S. Breen

Nature Communications **13**, Article number: 2997 (2022) | Cite this article



Next Steps

- Regenerate the figures with a new filtering algorithm
- Confirm that the novel sites have similar motifs and are in line with the context of REDIportal data
- Continue to optimize the script for ease of use and efficiency