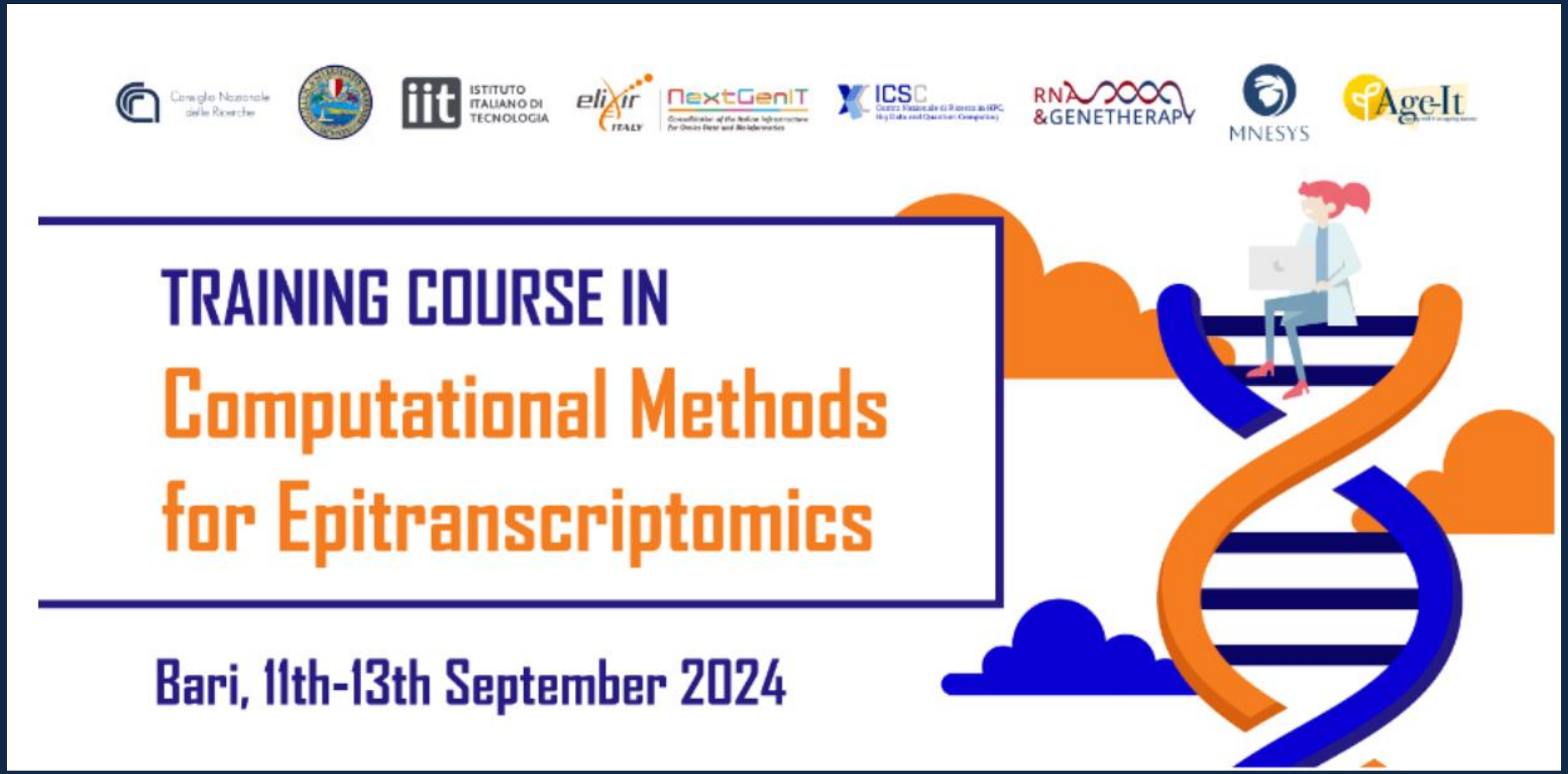
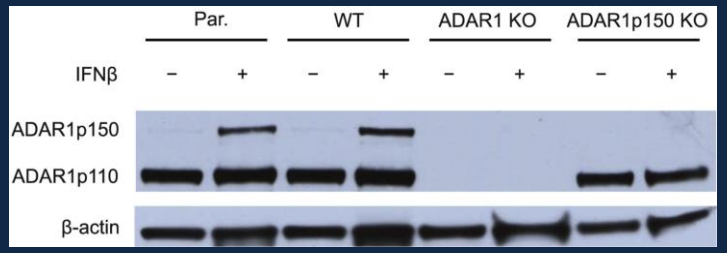
Computational approaches for detection and quantification of A-to-I RNA-editing



Dr. Domenico Alessandro Silvestris PhD RTD-A
Dipartimento di Bioscienze, Biotecnologie e Ambiente (DBBA), UNIBA email: domenico.silvestris@uniba.it

De novo detection of RNA editing sites by RNA-seq data

- preparing the necessary folders and input files



mkdir SRR5564268.ADAR1KO
mkdir SRR5564272.ADAR1KO
mkdir SRR5564273.ADAR1KO
mkdir SRR5564274.wildtype
mkdir SRR5564275.wildtype
mkdir SRR5564276.wildtype

DATASET

Series GSE99249 Query DataSets for GSE99249

Status	Public on Feb 09, 2018
Title	Determine the ADAR1 editome during IFN response.
Organism	Homo sapiens
Experiment type	Expression profiling by high throughput sequencing
Summary	ADAR1 is an interferon (IFN) inducible RNA editing enzyme that converts adenosine (A) to inosine (I). Here we identified ADAR1 and ADAR1p150 specific editing events by conducting RNA-seq on WT, ADAR1 KO, and ADAR1p150 KO cells.
Overall design	Mock or IFNb (1nM) treat WT, ADAR1 KO, and ADAR1p150 KO cell lines. For each genotype, three independent cell clones were used.
Web link	https://www.ncbi.nlm.nih.gov/pubmed/29395325
Contributor(s)	Rice CM, Chung H, Calis JJ, Rosenberg BR
Citation(s)	Chung H, Calis JJA, Wu X, Sun T et al. Human ADAR1 Prevents Endogenous RNA from Triggering Translational Shutdown. *Cell* 2018 Feb 8;172(4):811-824.e14. PMID: 29395325
BioProject	PRJNA386593 Homo sapiens Transcriptome or Gene expression

CRISPR/CAS9 genome engineering derived HEK293T cell culture clones (human embryonic kidney cultures)

Total RNA, Illumina Strand- oriented, 150 bp, paired end reads [*RPL7L1* gene selected]

cp /home/instructor_2/data/SRR5564268.ADAR1KO/*.bam* /home/student_X/SRR5564268.ADAR1KO
cp /home/instructor_2/data/SRR5564272.ADAR1KO/*.bam* /home/student_X/SRR5564272.ADAR1KO
cp /home/instructor_2/data/SRR5564273.ADAR1KO/*.bam* /home/student_X/SRR5564273.ADAR1KO
cp /home/instructor_2/data/SRR5564274.wildtype/*.bam* /home/student_X/SRR5564274.wildtype
cp /home/instructor_2/data/SRR5564275.wildtype/*.bam* /home/student_X/SRR5564275.wildtype
cp /home/instructor_2/data/SRR5564276.wildtype/*.bam* /home/student_X/SRR5564276.wildtype

https://drive.google.com/drive/folders/14kUimiDtd_CRWRB6ZRqMe--7M2_GtYd4?usp=sharing

FOR EACH SAMPLE EXECUTE THE FOLLOWING COMMAND LINES

RNA_BAM = /home/student_X/SAMPLE/output.bam # **change this parameter each time according to the sample name**
DNA_BAM = /home/instructor_2/WGS/output.bam
REF_FASTA = /home/instructor_2/accessory_files/PAR_masked-GRCh38.primary_assembly.genome.fa
Splice_sites_list = /home/instructor_2/accessory_files/gencode.v45.primary.splicesites.txt
REDI_OUT = denovo_pipeline

Go to each SAMPLE DIRECTORY (SRR5564268.ADAR1_KO, …, …,…) with cd

activate conda environment (python 2.7 + pysam 0.20.0)

source /home/instructor_2/miniconda3/bin/activate reditools1

RNA editing call with reditools1 RUN 1 collect all sites with all types of substitution (with strand correction) with pre-set default filters for minimum coverage (10 reads), minimum editing level (10%), minimum mapping quality (255 for RNA and 60 for DNA), minimum number of bases to support the variation (3) and excluding all the positions not supported by DNA [**N.B.** this type of setting is more recommended to detect sites in non-repeated regions; to detect editing in *Alu* regions, characterized by very low editing levels and low coverage, a less stringent setting of the filters is suggested.]

python /home/instructor_2/exe/REDItools/main/REDItoolDnaRna.py -i RNA_BAM -j DNA_BAM -o REDI_OUT -f REF_FASTA -d -D -s 2 -g 2 -S -m 60,255 -l -L -p -P -u -U -e -E -N 0.00 -z -t 40 -w Splice_sites_list -a 10-10 -A 10-10 -V

filtering of reditools output tables to obtain only positions with AG variation AND invariants at the DNA level

cd REDI_OUT/DnaRna…. # go to REDItools RUN 1 output directory

cat outTable | awk '{if ($11 == "AG" && $19 == "-") print $0;}' > filtered_pos

$11 == "AG" selects only positions with AG variation
$19 == "-" selects only positions with no variation in the DNA

N.B. in the event that it is not possible to carry out strand correction (for example with an unstranded RNA-Seq) we will filter all AG and TC variations (awk '{if ($11 == "AG" && $19 == "-" || $11 == "TC" && $19 == "-") print $0;}').

convert editing candidates sites in GFF format for further filtering

python /home/instructor_2/exe/REDItools/accessory/TableToGFF.py -i filtered_pos -s -t -o filtered_pos.gtf

```
USAGE: python TableToGFF.py [options]
Options:
-i              Table file from REDItools
-s              Sort output GFF
-t              Tabix output GFF (requires Pysam module)
-b              Buffer size (as number of lines) [32000] (requires -s)
-T              Temporary directory (requires -s)
-o              Outfile [outTable_518310875.gff]
-h              Print this help
```

RNA editing call with reditools1 RUN 2 only on the list of previously filtered positions (provided as gtf) extracting the reads supporting each AG position

Go to each SAMPLE DIRECTORY (SRR5564268.ADAR1_KO, …, …,…) with cd

REDI_OUT = denovo_pipeline2

python /home/instructor_2/exe/REDItools/main/REDItoolDnaRna.py -i RNA_BAM -j DNA_BAM -o REDI_OUT -f REF_FASTA -d -D -s 2 -g 2 -S -m 60,255 -l -L -p -P -u -U -e -E -N 0.00 -z -R -t 40 --reads --addP -w Splice_sites_list -a 10-10 -A 10-10 -T filtered_pos.sorted.gff.gz

remove sites within high similarity regions: reads supporting AG variations are realigned against the reference genome with pblat in order to identify reads mapping to paralogous regions and mark them as "badreads"

cd REDI_OUT # path to REDItools RUN 2 directory

/home/instructor_2/exe/pblat/pblat -t=dna -q=rna -stepSize=5 -repMatch=2253 -minScore=20 -minIdentity=0 REF_FASTA file reads.psl # file = path to outReads_ file

/home/instructor_2/exe/REDItools/accessory/readPsl.py reads.psl badreads.txt

RNA editing call with reditools1 RUN 3 recalls the previously detected edited positions, this time not considering the "badreads"

Go to each SAMPLE DIRECTORY (SRR5564268.ADAR1_KO, …, …,…) with cd

REDI_OUT = denovo_pipeline3

python /home/instructor_2/exe/REDItools/main/REDItoolDnaRna.py -i RNA_BAM -j DNA_BAM -o REDI_OUT -f REF_FASTA -d -D -s 2 -g 2 -S -m 60,255 -l -L -p -P -u -U -e -E -N 0.00 -z -R -t 40 -W AG -w Splice_sites_list -a 10-10 -A 10-10 -T ${ full path to dir1}/filtered_pos.sorted.gff.gz -b ${full path to dir2}/badreads.txt

OPTIONAL STEPS:

SNP annotation and filtering (even if we have DNA from the same sample or cell line as support, as a further cleanup of possible false positives in regions poorly covered by WES/WGS, we can use the list of known SNPs downloaded from UCSC and eliminate all positions that coincide with a SNP)

Go to denovo_pipeline3/DnaRna… directory

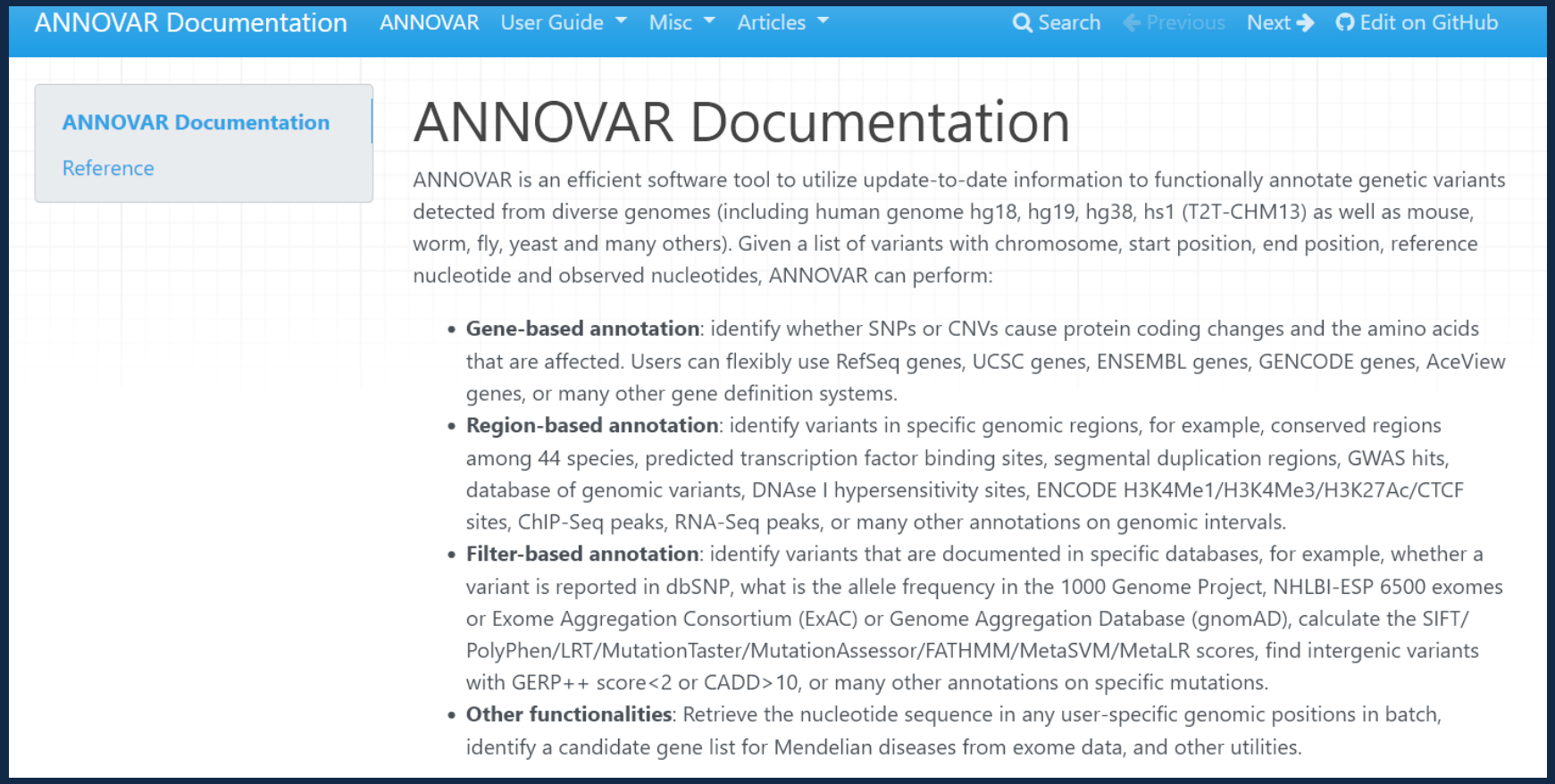
python /home/instructor_2/exe/REDItools/accessory/AnnotateTable.py -a /home/instructor_2/accessory_files/cDNA_snp151.sorted.gtf.gz -n snp151 -i outTable -o outTable.snp –u

cat outTable.snp | awk '{if ($21 == "-") print $0;}' > outTable.snp.filt

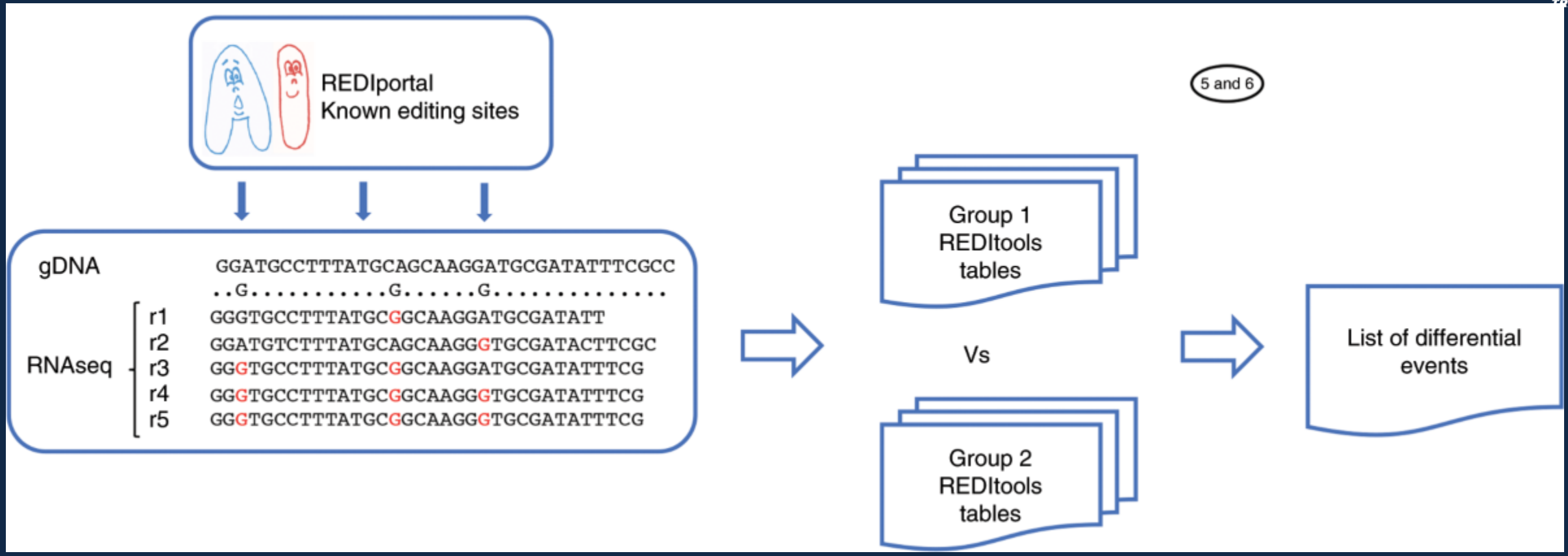
Annotation of sites in REDIportal using the list of sites already stored in the database downloadable from http://srv00.recas.ba.infn.it/webshare/ATLAS/donwload/TABLE1_hg38.txt.gz

python /home/instructor_2/exe/REDItools/accessory/AnnotateTable.py -a /home/instructor_2/accessory_files/atlas_hg38_sorted.gtf.gz -n ed -k R -c 1 -i outTable.snp.filt -o outTable.snp.filt.ed –u

A-to-G candidates can be annotated in more detail with dedicated tools such as ANNOVAR and annotation at the gene level (e.g. Gencode) or region level (e.g. RepeatMask).



Differential RNA editing



RNA editing has been shown to be implicated in a myriad of patho/physiological conditions from a functional point of view, it is therefore important to be able to statistically compare single site editing levels between conditions (e.g. healthy vs diseased or treated vs untreated).

The identification of differential RNA editing is still an open question. Nonetheless, dysregulated RNA editing at recoding events can be calculated employing the Mann-Whitney U-test described in Silvestris et al. (2019) or the statistical pipeline proposed by Tran et al. (2019) Both pipelines are embedded with the get_DE_events.py script.

Prepare a comma separated sample informations file (e.g wt_vs_ko.sif) required as input by the get_DE_events.py script. Run sample_status_file_creator.py providing:

- A csv sample file containing the main informations about each sample to be used in the experiment.
- A name for Samples group1 (e.g. wt)
- A name for Samples group2 (e.g ko)

es.
SRR5564268.ADAR1KO,GROUPB,ko

Run the get_DE_events.py script (Mann-Whitney U-test) on multiple REDItools tables following the sample/Group subdivions reported in the sample informations file (.sif). The option -sig yes in combination with -cpval 2 (BH correction), returns only significantly edited positions. MtsA and mtsB, represents the minimum threshold of samples per group on which the statistical tests are applied.

N.B. I strongly suggest using the BH test for p-value correction rather than Bonferroni which could be too stringent!

```
rm /home/student_X/SRR*/denovo_pipeline3/DnaRna_*/parameters.txt  # get_DE_events script works only on outTables
python /home/instructor_2/data/get_DE_events.py -input_file wt_vs_ko.sif > DE_res
```

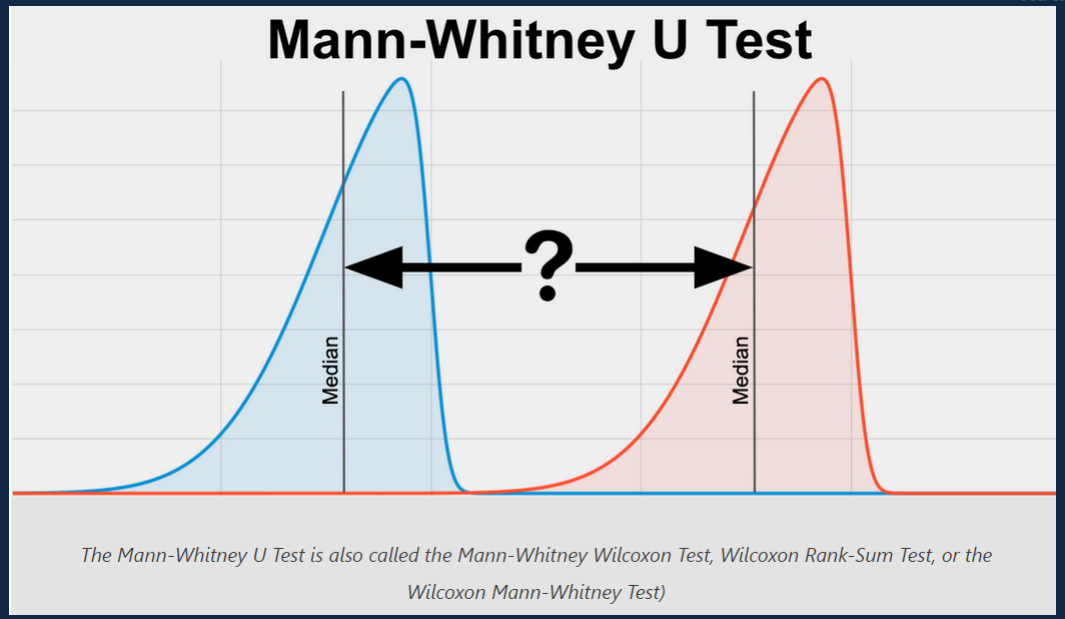
Alternatively, run the get_DE_events.py script on the same samples applying the the statistical pipeline proposed by Tran et al. (2019)

```
python /home/instructor_2/data/get_DE_events.py -linear -input_file wt_vs_ko.sif
```

```
usage: get_DE_events.py [-h] [-c MIN_COVERAGE] [-cpval PVALUE_CORRECTION]
                        [-input_file SAMPLES_INFORMATIONS_FILE]
                        [-gene_pos_file GENE_POS_FILE] [-f MIN_EDIT_FREQUENCY]
                        [-mtsA GROUPA_MIN_SAMPLE_TESTING]
                        [-mtsB GROUPB_MIN_SAMPLE_TESTING]
                        [-sig ONLY_SIGNIFICANT]
                        [-siglevel STATISTICAL_SIGNIFICANCE] [-linear]
                        [-graph] [-chr_col CHR_COLUMN] [-rsite]

optional arguments:
  -h, --help            show this help message and exit
  -c MIN_COVERAGE       Coverage-q30
  -cpval PVALUE_CORRECTION
                        1 --> Bonferroni correction / 2 --> Benjamini hochberg
  -input_file SAMPLES_INFORMATIONS_FILE
                        Comma separated file e.g: Sample,Group,Type
                        SRR1093527,GROUPA,BrainCerebellum...
                        SRR1088437,GROUPB,ArteryTibial... etc
  -gene_pos_file GENE_POS_FILE
                        nonsynonymous_table_NONREP derived from Rediportal
                        NOTE: A gene_pos_file is required by -graph or -rsite
                        options. An example file can be found at "https://gith
                        ub.com/BioinfoUNIBA/QEdit/blob/master/Example_files/no
                        nsynonymous_table_NONREP_2BS.txt"
  -f MIN_EDIT_FREQUENCY
                        Editing Frequency
  -mtsA GROUPA_MIN_SAMPLE_TESTING
                        min percentage of groupA samples
  -mtsB GROUPB_MIN_SAMPLE_TESTING
                        min percentage of groupB samples
  -sig ONLY_SIGNIFICANT
                        Return only significant editing events
  -siglevel STATISTICAL_SIGNIFICANCE
                        cutoff level to reject H0 hypothesis default 0.05
  -linear               Enable linear statistical model
  -graph                R graph compatible table containing the following
                        columns: Edited_Site | Delta_mean | log_padjstd |
                        color NOTE: THIS OPTION CAN BE USED ONLY IN
                        COMBINATION with -Gene_pos_file
  -chr_col CHR_COLUMN   If set to "yes" a chromosome_position column will be
                        aded to R graph table. NOTE: THIS OPTION IS SPECIFIC
                        FOR -graph & -Gene_pos_file COMBINATION
  -rsite                If set to "yes" all recoding sites will be shown in
                        the output table. NOTE: THIS OPTION ONLY WORKS IN
                        DEFAULT MODE.
```

N.B. get_DE_events.py needs SciPy and pandas



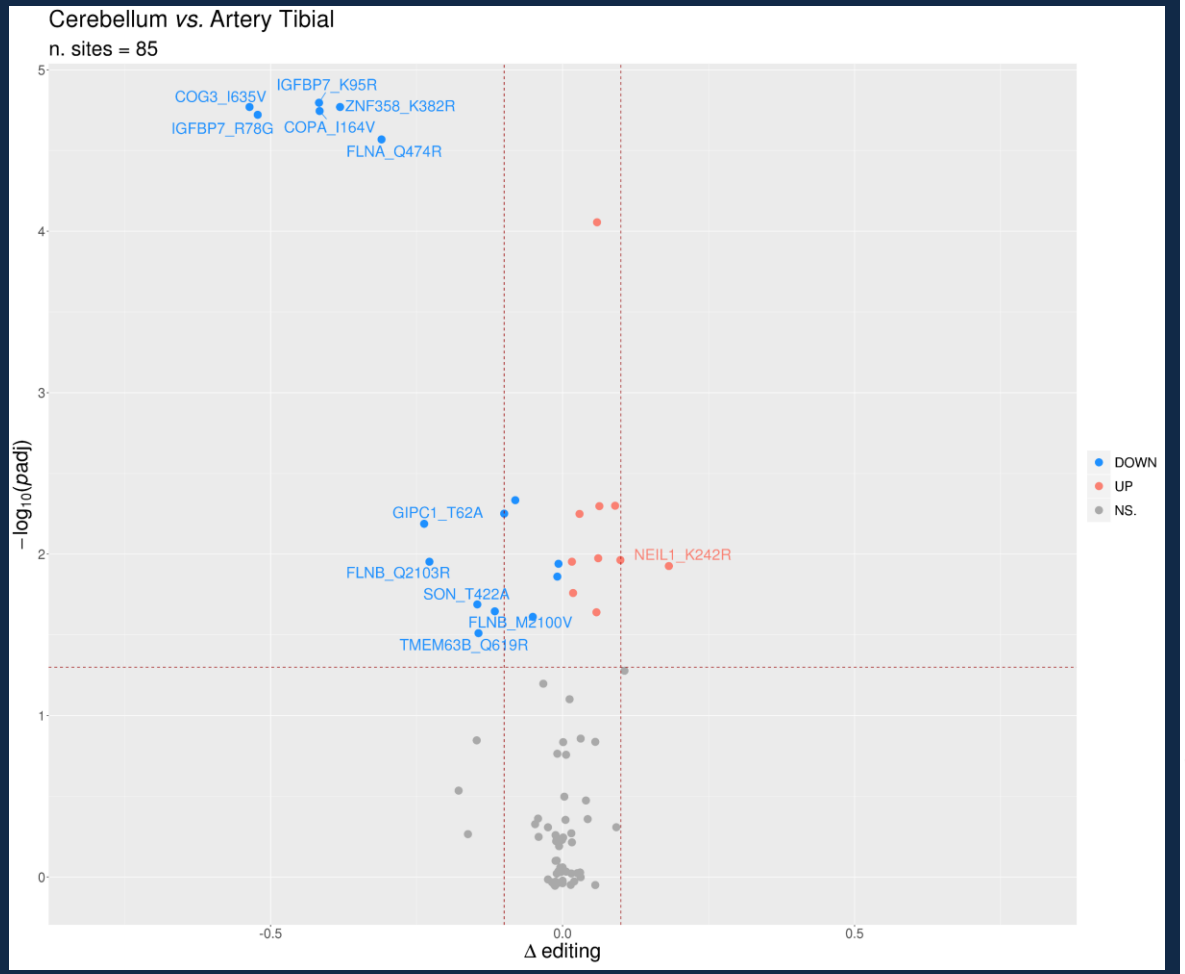
The **Mann-Whitney U Test** is a non-parametric statistical test used to determine if 2 groups are significantly different from each other on your variable of interest. Your variable of interest should be continuous and your 2 groups should have similar values on your variable of interest. Your 2 groups should be independent (not related to each other) and you should have enough data (more than 5 values in each group, though it also depends on how big the difference is between groups).

chromosome	position	editing_type	SRR5564268.ADAR1KO_ko	SRR5564272.ADAR1KO_ko	SRR5564273.ADAR1KO_ko	SRR5564274.wildtype_wt	SRR5564275.wildtype_wt	SRR5564276.wildtype_wt	[groupA_samples/groupB_samples]	delta_diff	pvalue (Mannwhitney)
chr6	42886104	AG	1.00^AG^29	1.00^AG^17	1.00^AG^14	1.00^AG^18	1.00^AG^17	1.00^AG^20	[3, 3]	0.0	1.0
chr6	42888224	AG	0.20^AG^48	0.32^AG^58	0.17^AG^14	0.58^AG^62	0.59^AG^139	0.60^AG^105	[3, 3]	-0.36	0.080856
chr6	42888313	AG	0.25^AG^97	0.35^AG^98	0.23^AG^32	0.27^AG^36	0.32^AG^70	0.26^AG^51	[3, 3]	-0.007	0.662521
chr6	42888372	AG	0.27^AG^135	0.40^AG^148	0.45^AG^95	0.96^AG^150	0.93^AG^252	0.88^AG^183	[3, 3]	-0.55	0.080856
chr6	42888378	AG	0.12^AG^61	0.19^AG^73	0.24^AG^47	0.69^AG^108	0.71^AG^203	0.62^AG^132	[3, 3]	-0.49	0.080856
chr6	42889245	AG	0.21^AG^207	0.24^AG^221	0.21^AG^105	0.26^AG^115	0.24^AG^171	0.23^AG^118	[3, 3]	-0.023	0.261155
chr6	42882951	AG	-	-	0.47^AG^7	0.33^AG^8	0.45^AG^5	[0, 3]	-	-	
chr6	42888234	AG	-	-	0.15^AG^16	0.18^AG^44	0.16^AG^32	[0, 3]	-	-	
chr6	42888241	AG	-	0.12^AG^20	-	0.31^AG^31	0.29^AG^71	0.27^AG^50	[1, 3]	-	-
chr6	42888291	AG	-	-	0.47^AG^66	0.49^AG^146	0.53^AG^117	[0, 3]	-	-	
chr6	42888309	AG	-	-	0.18^AG^20	0.29^AG^54	0.26^AG^44	[0, 3]	-	-	
chr6	42888317	AG	-	-	0.18^AG^24	0.20^AG^48	0.16^AG^31	[0, 3]	-	-	
chr6	42888362	AG	-	-	0.24^AG^40	0.31^AG^85	0.19^AG^38	[0, 3]	-	-	
chr6	42888367	AG	-	-	0.11^AG^18	0.13^AG^34	-	[0, 2]	-	-	
chr6	42888380	AG	-	-	0.12^AG^16	0.19^AG^42	0.11^AG^20	[0, 3]	-	-	
chr6	42888438	AG	-	-	0.28^AG^60	0.29^AG^118	0.28^AG^83	[0, 3]	-	-	
chr6	42889179	AG	-	-	0.63^AG^279	0.66^AG^512	0.68^AG^381	[0, 3]	-	-	
chr6	42889293	AG	-	-	0.32^AG^116	0.30^AG^150	0.33^AG^139	[0, 3]	-	-	
chr6	42889373	AG	-	-	0.20^AG^78	0.20^AG^129	0.20^AG^104	[0, 3]	-	-	
chr6	42881344	AG	-	-	-	-	0.25^AG^4	[0, 1]	-	-	
chr6	42883020	AG	-	-	-	-	0.27^AG^3	[0, 1]	-	-	

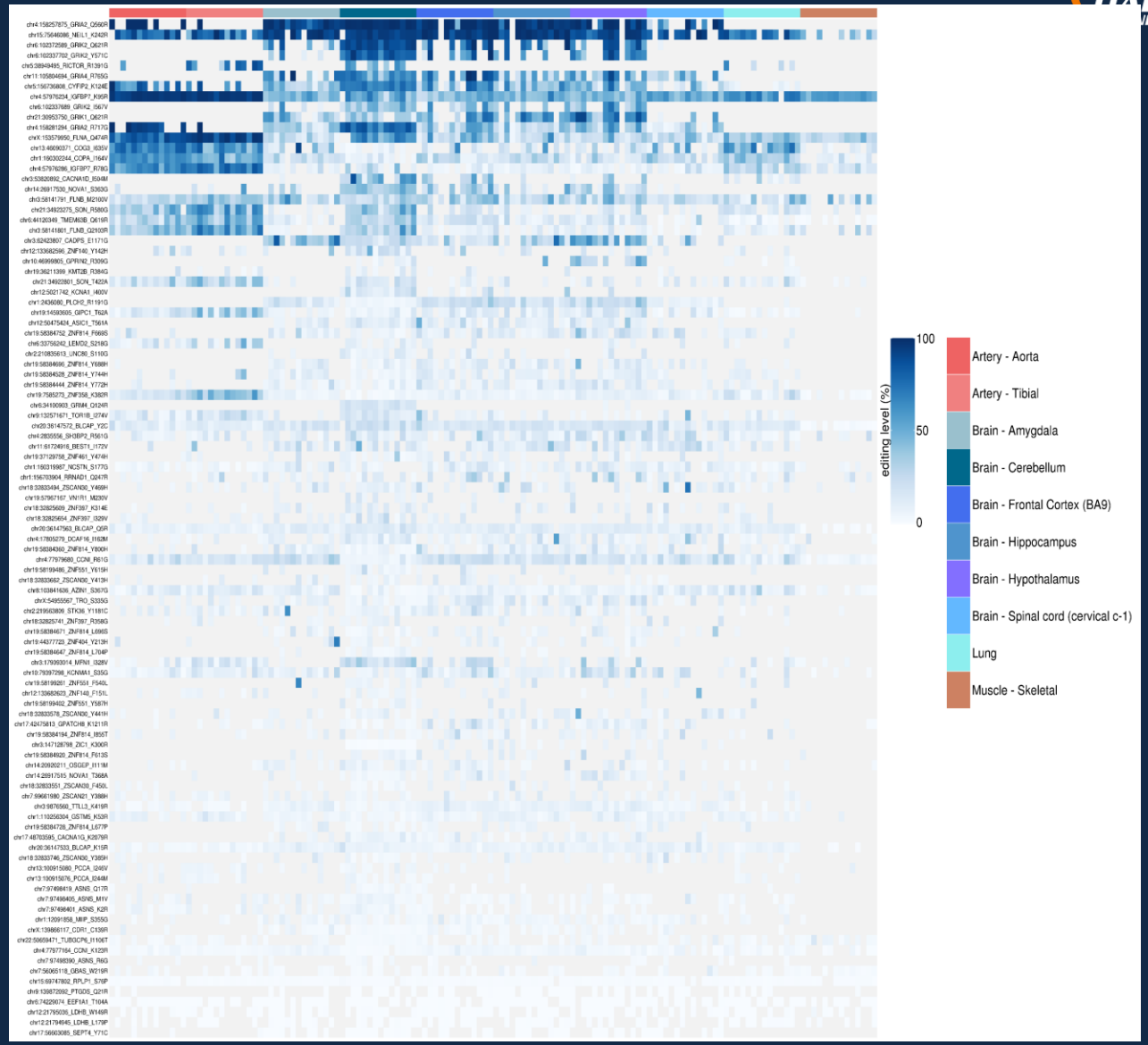
(reditools1) [instructor_2@wn-gpu-8-3-2 data]$

Editing levels visualization examples (https://github.com/BioinfoUNIBA/QEdit):

volcano plot



heatmap



REDItools KNOWN

REDItoolKnown.py has been developed to explore the RNA editing potential of RNA-Seq data sets using known editing events. Such events can be downloaded from REDIportal database (es. grep nonsynonymous to select recoding editing events)) or generated from supplementary materials of a variety of publications. Known RNA editing events have to be stored in TAB files.

TAB files are simple textual files with at least three tabulated columns including:

- genomic region (generally the chromosome name according to the reference genome)
- coordinate of the position (1-based)
- strand (+ or -). You can also indicate strand by 0 (strand -), 1 (strand +) or 2 (+ and - or unknown)

genomic region	coordinate	strand
chr21	10205589	-
chr21	10205629	-
chr21	15411496	+
chr21	15412990	+
chr21	15414553	+
chr21	15415901	+
chr21	15417667	+
chr21	15423330	+

TAB files must be coordinate sorted. In unix/linux environment they can be sorted by the sort command:

```
sort -k1,1 -k2,2n mytable.txt > mytable.sorted.txt
```

Example:

REDItoolKnown.py -i rnaseq.bam -f reference.fa -l knownEditingSites.tab

```
USAGE: python REDItoolKnown.py [options]
Options:
-i              BAM file
-I              Sort input BAM file
-f              Reference in fasta file
-l              List of known RNA editing events
-C              Base interval to explore [100000]
-k              List of chromosomes to skip separated by comma or file
-t              Number of threads [1]
-o              Output folder [rediFolder_268123878]
-F              Internal folder name [null]
-c              Min. read coverage [10]
-q              Min. quality score [30]
-m              Min. mapping quality score [30]*
-O              Min. homoplymeric length [5]
-s              Infer strand (for strand oriented reads) [1]
-g              Strand inference type 1:maxValue 2:useConfidence [1]
-x              Strand confidence [0.70]
-S              Strand correction
-G              Infer strand by gff annotation (must be sorted, otherwise use -X)
-X              Sort annotation files
-K              File with positions to exclude
-e              Exclude multi hits
-d              Exclude duplicates
-p              Use paired concardant reads only
-u              Consider mapping quality
-T              Trim x bases up and y bases down per read [0-0]
-B              Blat file for correction
-U              Remove substitutions in homopolymeric regions
-v              Min. num. of reads supporting the variation [3]
-n              Min. editing frequency [0.1]
-E              Exclude positions with multiple changes
-P              File containing splice sites annotations
-r              Num. of bases near splice sites to explore [4]
-H              No Table Header
-h              Print this help

*This value may change according to the aligner:
        - For Bowtie use 255
        - For Bowtie2 use 40
        - For BWA use 30
        - For RNA-STAR use 255
        - For HISAT2 use 60
        - For Tophat1 use 255
        - For Tophat2 use 50
        - For GSNAP use 30
```


THANK YOU
FOR YOUR ATTENTION!