Deciphering the role of non-coding variants in the etiology of neurodegenerative diseases by massively parallel reporter assay

Lucia Corrado1, Fjorilda Caushi1, Beatrice Piola*1, Endri Visha1, Erica Melone1, Diego Cotella1, Laura Follia1, Martina Tosi1, Fabiola De Marchi2, Luca Magistrelli3, Letizia Mazzini2, Alfredo Brusco4, Sandra D'alfonso1

1. University of Eastern Piedmont, Dept. of Health Sciences, Novara, Italy, 2. University of Eastern Piedmont, Maggiore Della Carità Hospital, Novara, Italy, 3. University of Eastern Piedmont, Maggiore Della Carità Hospital, Department of Neurology and ALS Centre, Novara, Italy, 4. University of Torino, Department of Medical Sciences, Torino, Italy

Background

NGS technology succeed to identify a large number causative and susceptibility genes for Neurodegenerative diseases (NDDs) , such as Alzheimer's disease (AD), Parkinson's disease (PD), Amyotrophic Lateral Sclerosis (ALS), Frontotemporal Dementia (FTD) and Spinocerebellar Ataxia (SCA), however a missing heritability was reported. The majority of reported studies have been focused mainly on the coding regions. Whole Genome Sequencing is a powerful tool, although it is still challenging the analysis of the huge amount of the non-coding variants such as in promoters, enhancers, 5' and 3' UTR. For this reason, they are poorly investigated, thus, even though NGS has increased the rate of genetic detection, variants in non-coding regions may explain part of the missing heritability.

Aim of the study

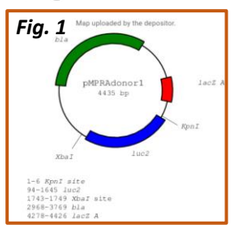
The general aim of this study is to investigate a possible role of rare non-coding variants in NDDs from WGS data on gene expression regulation by an high throughput technology enables to analyze transcriptional activities of thousands of regulatory elements at the same time and the prediction of their possible pathogenic impact on gene expression.

Materials and Methods

Variants selection: From WGS data of 140 patients affected by the 4 above mentioned neurodegenerative disease , we selected rare (MAF<0.01) non-coding variants located in the 5' flanking regions of a panel of 696 genes previously associated to NDDs. Variant prioritization was performed according to the overlap with potentially functional genomic signatures from UCSC genome browser annotations:

1. Presence of CpG islands
2. Sequences relevant to the regulation of transcription from the ENCODE Project
3. Presence of candidate cis-Regulatory Elements (cCREs) from the ENCODE Registry
4. Sequence annotated in GeneHancer as containing regulatory elements, gene transcription start sites, interactions (associations) between regulatory elements and genes
5. Sequence annotated by Open Regulatory Annotation (ORegAnno) which displays literature-curated regulatory regions, transcription factor binding sites, and regulatory polymorphisms.

Massively Parallel Reporter Assay (MPRA): The MPRA is a method useful to test functionally thousands of possible regulatory elements, screening them in parallel.
This is achieved by adding genetic barcodes specific for the putative regulatory sequence to a reporter gene. Instead of measuring fluorescence as a readout, RNA-seq can be performed to determine which regulatory elements were active and how many transcripts of their respective reporter genes were made as a result (with each one identifiable by its unique barcode).
MPRA assay has been applied to analyze DNA variants located in different genomic positions, such as in promoter regions (using pMPRAdonor1 vector) and in enhancer/silencer sequences (using pMPRAdonor2 vector) (Addgene) containing an ORF (figure 1 and 2).



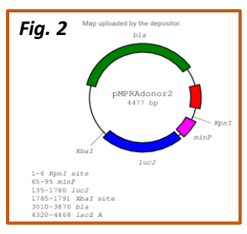


Figure 1: structure of pMPRAdonor1 containing the LUC2 ORF .

Figure 2: structure of pMPRAdonor2 containing the LUC2 ORF and the minimal promoter (minP).

A: Oligonucleotide library design: An oligonucleotide library synthesis (OLS) was created for MPRA application and subjected to two cloning procedures before transfection into eukaryotic cells.
Oligonucleotide sequence was formed by 145 bp (with 72 bp upstream and 72 bp downstream each variant in the two different allelic forms). Two flanking ePCR primers which are common for all the designed oligos and two restriction sites for XbaI and KpnI restriction enzymes (RE) between the tag and the variant were added. The final probes were 200 bp long (figure 3).

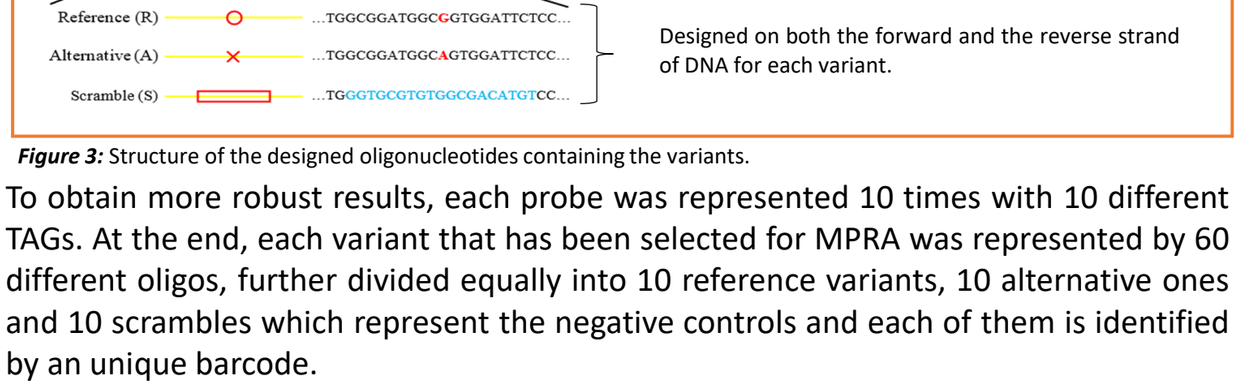


Figure 3: Structure of the designed oligonucleotides containing the variants.

To obtain more robust results, each probe was represented 10 times with 10 different TAGs. At the end, each variant that has been selected for MPRA was represented by 60 different oligos, further divided equally into 10 reference variants, 10 alternative ones and 10 scrambles which represent the negative controls and each of them is identified by an unique barcode.
A total of 2460 probes were generated by using a programmable microarray-based oligonucleotide synthesizer and the final library was synthesized by Agilent Technologies (Santa Clara, CA).

B: Emulsion PCR (ePCR): The ePCR step was aimed to amplify the library and to add SfiI restriction sites at the extremities of each probe of the library (required for the following cloning step) (figure 4).

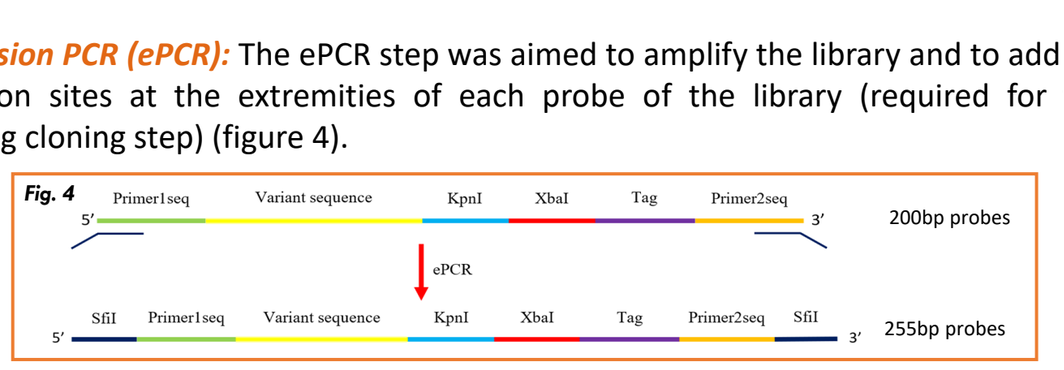


Figure 4: Structure of the designed oligonucleotides containing the variants after the e PCR.

C: First cloning step in DH5α E.Coli cells: The resulting library is transformed into DH5α E.Coli cells, after being ligated with pMPRA1 backbone obtained by digestion through SfiI enzyme of pMPRA1 plasmid (Addgene) → pMPRA1+OLS called intermediate plasmid library was obtained.

D: Second cloning step in DH5α E.Coli cells: The resulting intermediate plasmid library is transformed into DH5α E.Coli cells, after being ligated with LUC2 ORF obtained by pMPRAdonor1 digestion with KpnI and XbaI enzymes or minP+LUC2 ORF obtained by pMPRAdonor2 digestion with KpnI and XbaI enzymes → pMPRA1+OLS+ LUC2 and pMPRA1+OLS+minP+LUC2 were obtained.

E: Eukaryotic cells transfection and data analysis: Each library was transfected at least four times (two transfections for each promoter type) into HEK293T cells and SH-SY5Y cells (ongoing) aiming >100 times higher number of transfected cells than the library complexity. Total RNA was isolated (Qiagen miRNeasy Tissue/Cells Advanced Mini Kit). cDNA was then synthesized using Superscript II and Oligo (dT) Primer, from which only short sequences encompassing 10 bp unique barcodes were amplified using Herculase II Fusion DNA Polymerase (Agilent Technologies) and primers introducing Illumina adapter sequences. Sequence libraries were also prepared using 10ug of input DNA in the same way. All obtained PCR products was sequenced on an Illumina MiSeq instrument as 151-nt single-end reads. By calculating the log ratio of mRNA counts to DNA counts, the measurement of the transcriptional activity of each barcode's corresponding DNA sequence was obtained. For each transfection Tag counts, Per Million sequencing reads (TPM) values will be calculated by dividing each tag count by the total number of sequence-matching tag counts divided by a million. TPM ratio will be then taken as RNA TPM over input DNA TPM and log2 converted.
Bioinformatics analysis of NGS data was performed using R version 4.2.0, mpra package and mpralm function.

Results and conclusions

Selection of variants with a possible impact on gene expression from WGS data

From 140 patients' WGS data : variants localized in 696 genes known to be causative or susceptibility factors of NDDs → 870 rare non-coding variants in upstream or UTR sequences; Prioritization of rare variants with a possible role in gene expression regulation → 41 variants annotated as mapping in possible regulatory regions: obtained using UCSC-GRCh38/hg38 ENCODE-regulation tracks → 19 upstream variants, 22 5'UTR variants: depending on their position identified in UCSC Genome Browser

MPRA library synthesis

A total of 2460 probes comprising selected variants and characterized by different tags were synthesized. The library was subjected to a first amplification for the inclusion of cloning sites.

Final product after second cloning step in DH5α E.Coli cells

After the second cloning step the resulting constructs are composed by the pMPRA1 plasmid backbone with the OLS and the LUC2 ORF, when the variants are tested as promoters (figure 5 and 7), and the pMPRA1 plasmid backbone with the OLS and the minimal promoter followed by the LUC2 ORF, when the variants are tested as enhancer/silencer (figure 6).

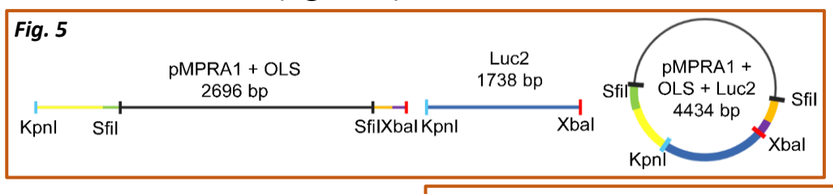
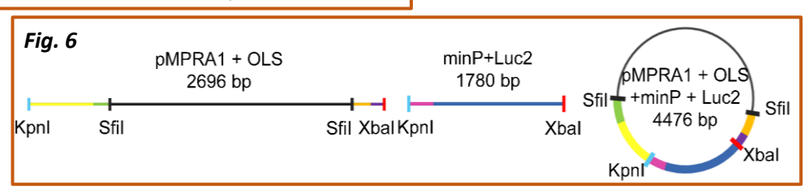


Figure 5: Graphical representation of the ligation of pMPRA+OLS and LUC2 ORF, used to test variants in putative promoter regions.

Figure 6: Graphical representation of the ligation of pMPRA+OLS and LUC2 ORF, used to test variants in putative enhancer/silencer regions .



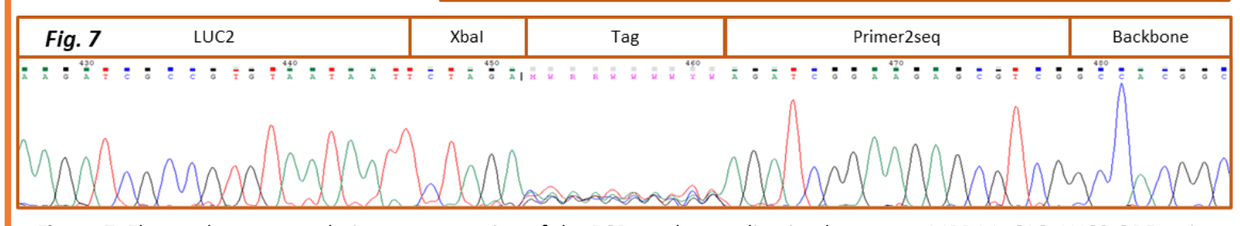


Figure 7: Electropherogram relative to sequencing of the PCR product on ligation between pMPRA1+OLS+LUC2 ORF; using forward primers that anneal on LUC2 and reverse primers on pMPRA1 backbone. The ligation sites between LUC2 ORF and Tag through XbaI site and between primer2seq of the OLS insert and pMPRA1backbone through SfiI site are visible.

Final NGS after eukaryotic cells transfection and data analysis

After the second cloning step, transfection of HEK293T and of SH-SY5Y (ongoing) eukaryotic cells was performed. An amplicon based NGS analysis "Nextera DNA CD indexes" was then performed for both the extracted cDNA and input DNA" following the "Illumina DNA Prep" protocol. Up to now, we have performed the assay on the HEK293T cell line, and the analysis showed that, among the 41 variants located in putative regulatory regions, 4 of them seem to deregulate gene expression, with a statistically significant P value, lower than 1x10-4 (table 1).
Of these variants, two belong to ALS patients (the ones located in *CST3* and *UBQLN2* genes), while the others belong to patients affected by ataxia.

Tab. 1	Gene	Disease	Observed effect
chr20:23637984	CST3	ALS	DOWNREGULATION
chrX:56563656	UBQLN2	ALS	UPREGULATION
chr10:100267697	CWF19L1	Ataxia	DOWNREGULATION
chr12:24562578	SOX5	Ataxia	NEUTRAL

Table 1: Table reporting the 4 selected variants, with a statistically significant adjusted p value.

These are rare variants as their frequencies are not reported in any database (GnomAd ExomeAll, GnomAd GenomeAll, 1000G).
It has already been observed that 2 of these variants mapped in genes *UBQLN2 and CWF18LK1*, are associated with patients' disease; the other 2 variants are localized in *CST3* and *SOX5* genes, which are not directly associated with the investigated disorders, even though they are known to be related to other neurological disorders.
UBQLN2 encodes ubiquilin-2, a member of the ubiquilin family of proteins that regulate the degradation of ubiquitinated proteins by the proteasome. Mutation in the *UBQLN2* are known to be causative of ALS-15 with or without FTD.
CWF19L1 encodes a member of the CWF19 protein family. Mutations in this gene have been associated with autosomal recessive spinocerebellar ataxia-17 and mild cognitive disability.
CST3 belongs to the type II cystatin gene family and is located in the cystatin locus. It encodes the most abundant extracellular inhibitor of cysteine proteases.
SOX5 gene (SRY-Box Transcription Factor 5) encodes a member of the SOX (SRY-related HMG-box) family of transcription factors involved in the regulation of embryonic development and in the determination of the cell fate. The encoded protein may act as a transcriptional regulator after forming a protein complex with other proteins.
In conclusion, we succeed to fine tune the MPRA assay on HEK293T cell line, demonstrating that this powerful technique can be applied simultaneously to a wide number of non-coding variants whose pathogenetic role needs to be evaluated.
The same assay will be performed on SH-SY5Y cell line (neuroblast from neural tissue), a more specific cell line usable to study NDDs. The resulting selected variants, will be further analyzed with other functional assays, to assess their possible role in the disease.

UPO AGING PROJECT UNIVERSITÀ DEL PIEMONTE ORIENTALE

Authors declare no conflict of interest.