WHOLE-EXOME SEQUENCING APPROACH IN A COHORT OF MULTIPLEX ITALIAN FAMILIES WITH MULTIPLE SCLEROSIS

A. Zauli[1], C. Guaschino[1,2], E. Mascia[1], F. Esposito[1,2], M. Sorosina[1], A. Osiceanu[1], S. Peroni[1], S. Santoro[1], D. Biancolini[3], D. Lazarevic[3], S. Bonfiglio[3], D. Cittaro[3], V. Martinelli[1], G. Meola[4], S. D'Alfonso[5], G. Comi[1,2], F. Martinelli Boneschi[1,2]

[1] Laboratory of Human Genetics of Neurological diseases, CNS Inflammatory Unit , Institute of Experimental Neurology, Division of Neuroscience, San Raffaele Scientific Institute, Milan; [2] Department of Neurology, San Raffaele Scientific Institute; [3] Center for Translational Genomics and Bioinformatics, San Raffaele Scientific Institute, Milan; [4] Department of Biomedical Sciences for Health, IRCCS Policlinico San Donato, University of Milan, San Donato Milanese, Italy; [5] Interdisciplinary Research Center of Autoimmune Diseases IRCAD, University of Eastern Piedmont, Novara & Department of Health Sciences, University of Eastern Piedmont, Novara, Italy.

INTRODUCTION

Multiple Sclerosis (MS) is a chronic inflammatory and degenerative disease of the central nervous system (CNS) caused by the interplay of several genetic and environmental factors. More than 100 common variants have been associated with MS by large collaborative genome-wide association studies (GWAS)[1,2], each with individual mild effect on the disease, but they explain only a small fraction of the observed heritability of the disease (28%). Therefore, additional genetic variants or biological mechanisms should contribute to the genetic architecture of the disease, such as rare variants with effect sizes larger than the one played by common variants. Such rare/private variant can be enriched in families with a higher prevalence of the disease. **Aim**: the aim of the study is to identify rare variants involved in the predisposition of MS by performing whole–exome sequencing in affected and unaffected cases of multiplex families.

PATIENTS AND METHODS

Patients: 12 Italian families with a minimum of 3 affected relatives (1 with 5, 4 with 4, 6 with 3, and 1 with 2 affected) have been recruited as part of a large multicentric Italian study. Among them, three are Sardinian families (**Fig. 1**). Segregation patterns with the disease in recruited families suggest different mechanisms, including also epigenetic.

Whole exome sequencing (WES): The WES approach was performed in 37 relatives (27 affected and 10 unaffected, see Figure 1). The median coverage was 83.3x (95%CI: 49.1-107.4), and almost all samples have >80% of percentage target bases with at least 20x (those that are below have been resequenced) (**Fig. 2**). The enrichment protocol was the Agilent QXT v5 kit (in few cases the TruSeq Exome of Illumina), and fragments were sequenced using an Illumina (HiSeq 2500) platform in paired end mode (2x101). Reads were trimmed using trimmomatic (Version:0.32), either for low quality bases (leading a trailing quality >= 28) and for adapters. A further trimming step using cutadapt (Version: 1.8.3) was performed to clean the remaining adapters. Reads were then aligned against hg19 reference genome, using BWA ("mem"; Version: 0.7.10-r789). Following the GATK[3,4] best practices a step of InDel Realignment was performed (GATK: IndelRealigner; Version: 3.3-0-g37228af), followed by a Base Quality Recalibration process (GATK: BaseRecalibrator). Variants were called using UnifiedGenotyper (GATK: UnifiedGenotyper; stand_emit_conf=28.0, stand_call_conf=50.0). A variant quality recalibration step (GATK: VariantRecalibrator/ApplyRecalibration; Sensitivity Threshold: 0.99) separately for SNPs and short InDels was performed. A final list (Tier0) of 3.650.438 raw variant was obtained. Tier0 from WES pipeline was filtered for variants that didn't pass the sensitivity threshold of 0.99. With this filter (Tier1) the number of variants was reduced to 3.259.436. This set of variants was annotated using SnpEff[5,6] suite (Version:4.1g) and proprietary tools, for Variants' Effects, for Variant Type, and with different external databases: dbSNP[7] (Version:142), dbNSFP[8] variants (Version:2.9), dbNSFP gene (Version:2.9). These variants were furtherly annotated by their global allele frequencies (AF) using external databases: ExAC[9] (Release:0.3), 1000 Genomes[9] (Release:20100804) and dbSNP (Release: 142). We apply a further filter in order to retain only those variants that could affect coding sequence (variants with an Effect HIGH or MODERATE according to the SnpEff annotations), obtaining a final list (Tier2) of 41821 functional variants **(Figure 3)**.

RESULTS

The pedigrees of recruited families are reported in **Figure 1.** Through the bioinformatic pipeline (see Methods section), 41.821 functional variants were found in affected relatives. Filtering for a global allele frequency (AF) ≤ 5% in public databases, the number of variants was reduced to 15878 **(Figure 3)**. Selecting variants present in all affected relatives (in homozygous or heterozygous status) of ≥ 50% of families, a final number of 140 rare variants was obtained **(Figure 3).** Among the 140 variants, 17 variants in 19 genes were detected in all relatives of all families (**Table 1**). Six of these 140 variants fall in genetic loci known to be associated with MS, but they have a high frequency in unaffected relatives and Italian healthy controls, suggesting that they are private genetic variants frequent in the Italian population and not implicated with the disease **(Table 2)**. Analyses are ongoing to better understand this peculiar pattern and the segregation of these variants in families.

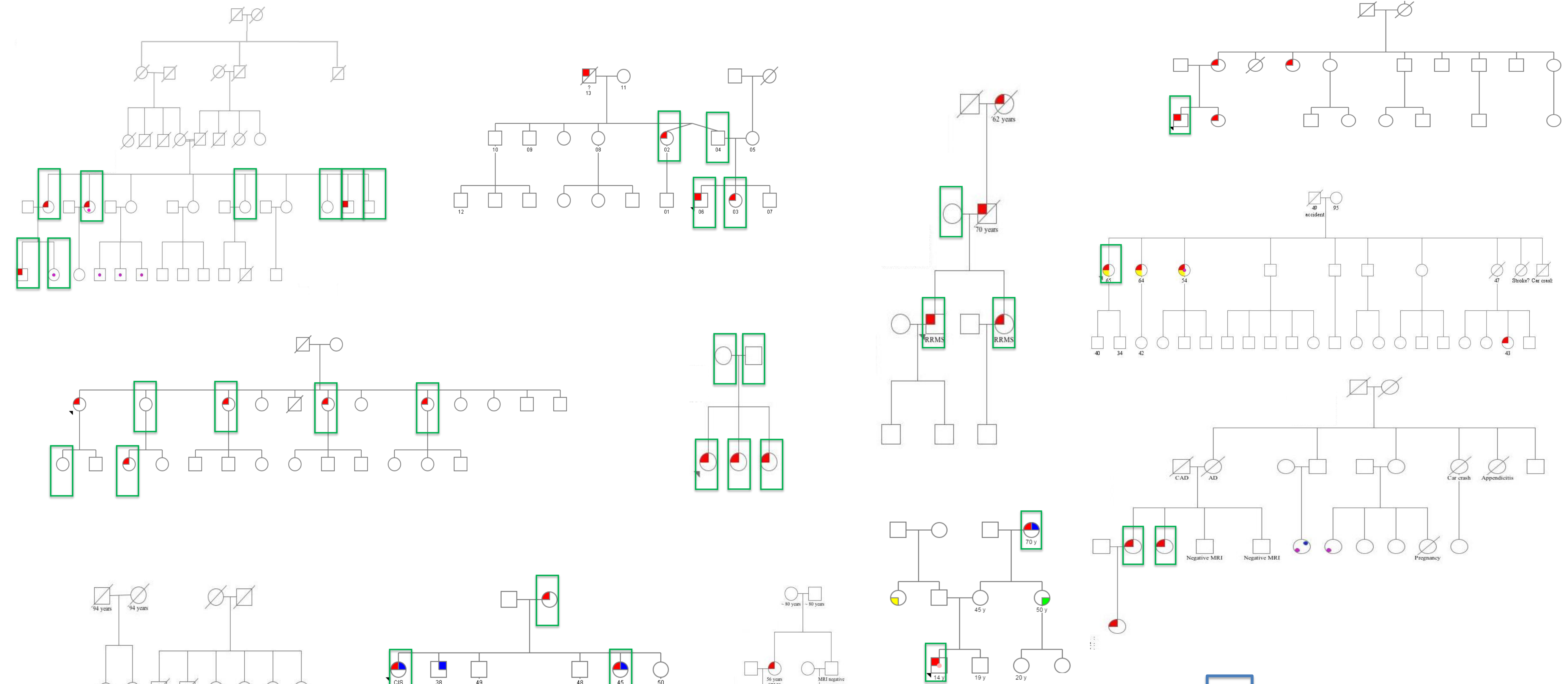


Figure 1: list of recruited families. Sequenced subjects are marked in green boxes. Subjects with red colour inside are affected with MS, blue colour is allergy, yellow thyroiditis, green aspecific autoimmune disease.

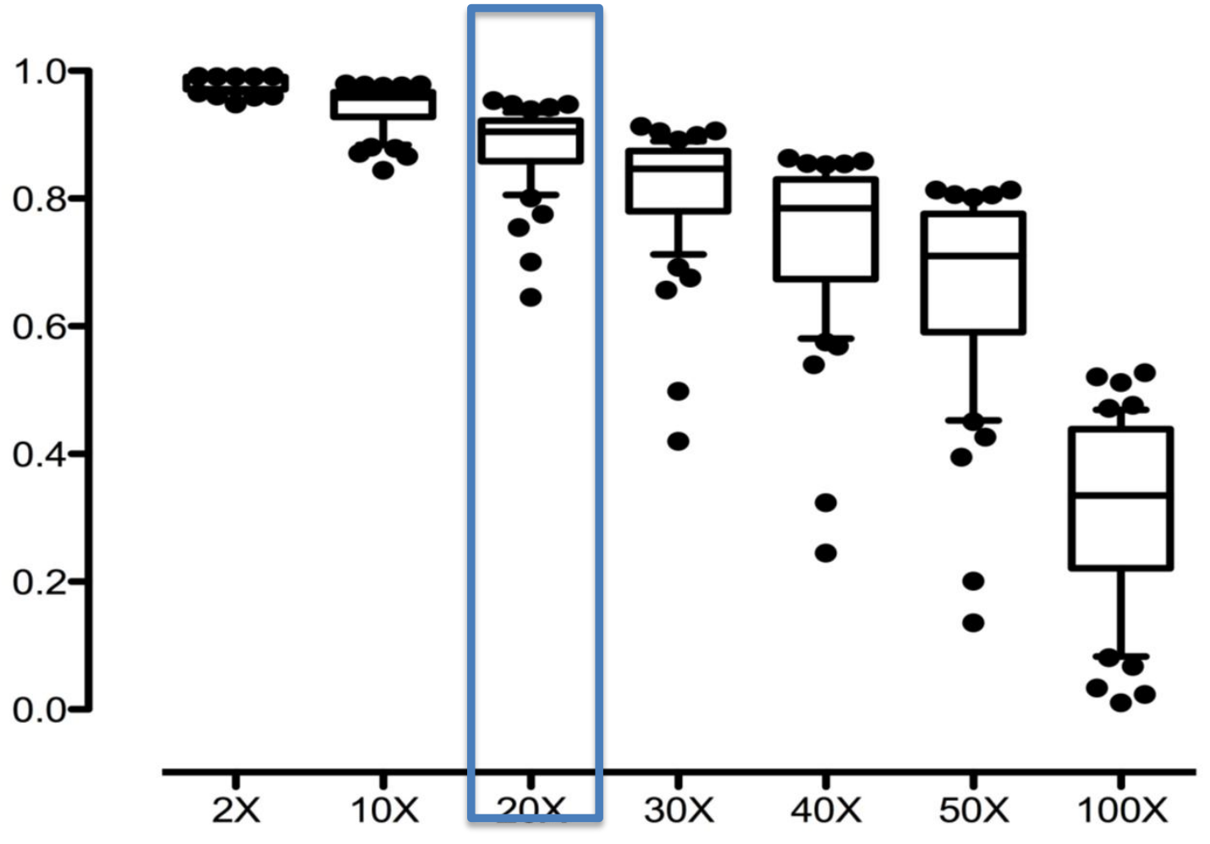
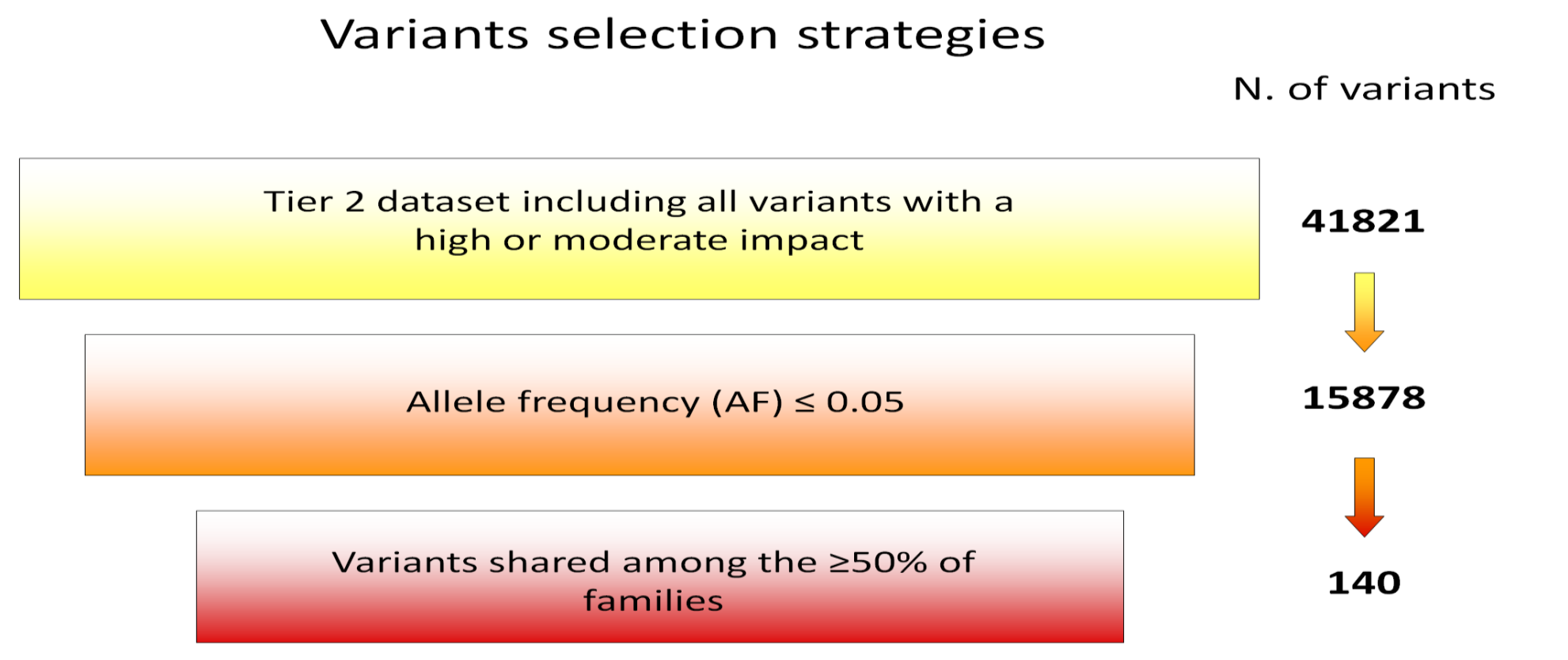


Figure 2: Percentage target bases in sequenced samples



Variants selection strategies

N. of variants

Tier 2 dataset including all variants with a high or moderate impact — 41821

Allele frequency (AF) ≤ 0.05 — 15878

Variants shared among the ≥50% of families — 140

Figure 3: filtering criteria to select variants.

AF	% sharing	N. of variants	N. of genes
≤ 0.05	≥50%	140	128
≤ 0.05	≥75%	77	73
≤ 0.05	100%	17	19

Table 1: Number of variants (and related genes) stratified according to MAF in public databases and proportion of presence in affected cases of families.

CHR:POS	Gene	Reference Allele	Alternative Allele	MAF Public Databases	MAF in Affected relatives	MAF in Unaffected Relatives	MAF in 62 Italian Healthy Controls
chr12:6777069	ZNF384	TTGCTGC	TTGC	0.0000259 (Kaviar)	0.761	0.8333	0.653
chr14:103576803	EXOC3L4	CGG	CG	0.012 (Exac)	0.389	0.45	0.508
chr15:78913067	CHRNA3	ACAG	A	0.0388999 (Kaviar)	0.542	0.8	0.829
chr16:85815220	EMC8	C	CT	0.0054557 (Kaviar)	0.85	1	0.75
chr5:36166827	SKP2	CT	CTT	0.0366528 (Kaviar)	0.519	0.65	0.51
chr8:79598666	ZC2HC1A	GTTTATTTA	GTTTA	0.0081758 (Kaviar)	0.636	0.5	0.746

Table 2: MAF of 6 rare variants in genetic MS loci

CONCLUSIONS

This is the first Italian study that uses a WES approach in MS multiplex continental Italian families to identify rare susceptibility variants, and preliminary analyses allowed to identify a first set of 140 rare functional variants present in MS affected relatives. Analyses are ongoing to explore the frequency of this set of 140 variants in the Italian healthy population to distinguish risk variants from variants specific of Italian population and their segregation with the disease within families. The WES of additional 35 affected and unaffected relatives from MS multiplex families are ongoing, as well as the clinical recruitment of additional families in collaboration with Prof. S D'Alfonso (University of Novara) and IMSCG members.

REFERENCES

1 IMSGC et al, Nature Genetics, 2013; 2 IMSGC et al, Nature 2011; 3. McKenna A et al., Genome Res. 2010; 4. Van der Auwera GA et al., Curr Protoc Bioinformatics. 2013; 5. Cingolani P et al., Fly 2012.; 6. Cingolani P et al., Front Genet. 2012. 7. Jan Sherry ST et al., Nucleic Acids Res. 2001.;8. Liu et al., Hum Mutat. 2011.; 9. Exome Aggregation Consortium et al., Biorxiv 2015; 10. The 1000 Genomes Project Consortium, Nature 2015.