FAMILY-BASED LINKAGE ANALYSIS AND FULL EXOME SEQUENCING FOR THE IDENTIFICATION OF POTENTIAL RISK VARIANTS IN IgA NEPHROPATHY

[1]Sharon Natasha Cox, [2]Francesco Pesce, [2]Julia Sarah El-Sayed Moustafa, [3]Fabio Sallustio, [1]Grazia Serino, [4]Annalisa Giampetruzzi, [5]Nicola Ancona, [2]Mario Falchi and [1,6]Francesco Paolo Schena,

[1]D.E.T.O. University of Bari, Bari, Italy, [2]Dept of Genomics of Common Disease, Imperial College, London ,UK, 1,[3]DiSTeBA, Università del Salento, S.P.6, 73100 LECCE. [4]IPSP, CNR, Bari, Italy. [5]ISSIA, CNR, Bari, Italy, [6]C.A.R.S.O. Consortium, Molecular Biology, Bari, Italy;

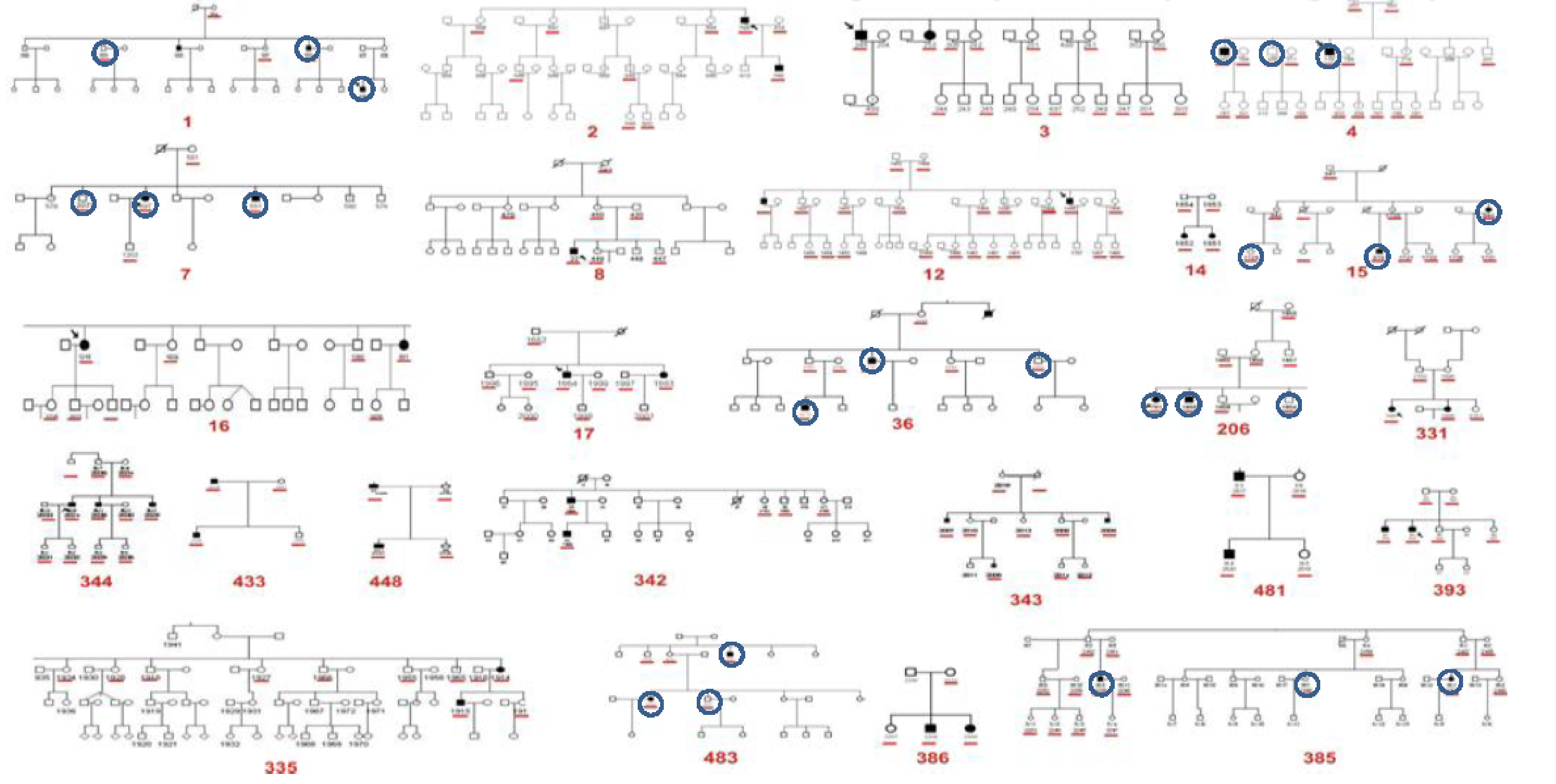
INTRODUCTION AND OBJECTIVES

IgA nephropathy (IgAN) is the most common form of primary glomerulonephritis world-wide. The pathogenetic mechanisms are still unknown, however IMMUNOLOGICAL and GENETIC FACTORS seem to play a key role in disease susceptibility. The strong genetic component is supported by familial clustering, ethnic variation in prevalence and by reports of large pedigrees containing multiple affected individuals. IgAN is a genetically complex disease depending on the complex interaction of multiple genes and environmental factors. Genome-wide linkage studies (GWLS) and genome-wide association studies (GWAS) have been performed to identify specific genetic markers involved in IgAN. Three GWLS of familial IgAN have reported linkages at the following chromosomal regions 2q36, 4q26–31, 6q22–23 and 17q12–22, however, no disease genes were identified within these areas.

Aim of our study was to find rare, high penetrant risk variants, combining family-based linkage analysis with full exome sequencing.

MATERIALS & METHODS

Multiplex families included in the linkage study and sequencing study



25 multiplex families were included in the initial linkage study, the red bars represent 217 genotyped subjects. Subjects included in our exome sequencing study are circled in blue

MATERIALS & METHODS

Linkage Analysis

Samples Preparation

DNA was extracted using the QIAGEN QIAmp Midi kit from EDTA anticoagulated peripheral blood collected from 25 Italian families of south italian ancestry.

Genotyping

Genotyping was performed using Illumina HumanCytoSNP-12 BeadChip. Data was exported from Genome Studio Software.

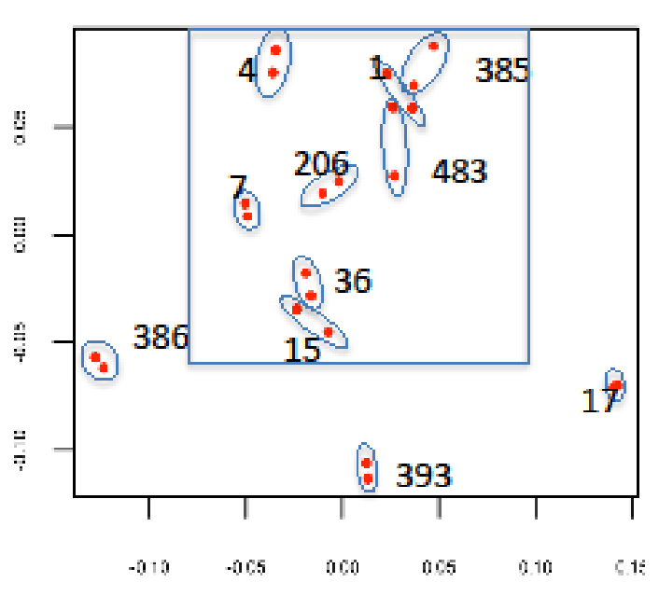
Quality Control

Individuals with more than 5% missing genotypes and more than 5% mendelian errors were excluded. Markers that failed the Hardy-Weinberg test ($P \leq 1 \times 10^{-6}$) and those with MAF≤0.05 were excluded. Genotyping errors were also detected and removed using Merlin error detection analysis (--error option).

Linkage Analysis

After quality control and pedigree corrections, 267 individuals for a total of 23 families and 227,114 SNPs were available for analysis. Non-parametric linkage (modeling linkage-disequilibrium with a r^2>0.10) analysis was performed on 16 most informative families. Linkage analysis was carried out under models of locus homogeneity or heterogeneity (specifying IgAN as an autosomal dominant trait with disease allele frequency of 0.001 and estimated penetrance of 75%).

Sequencing strategy definition



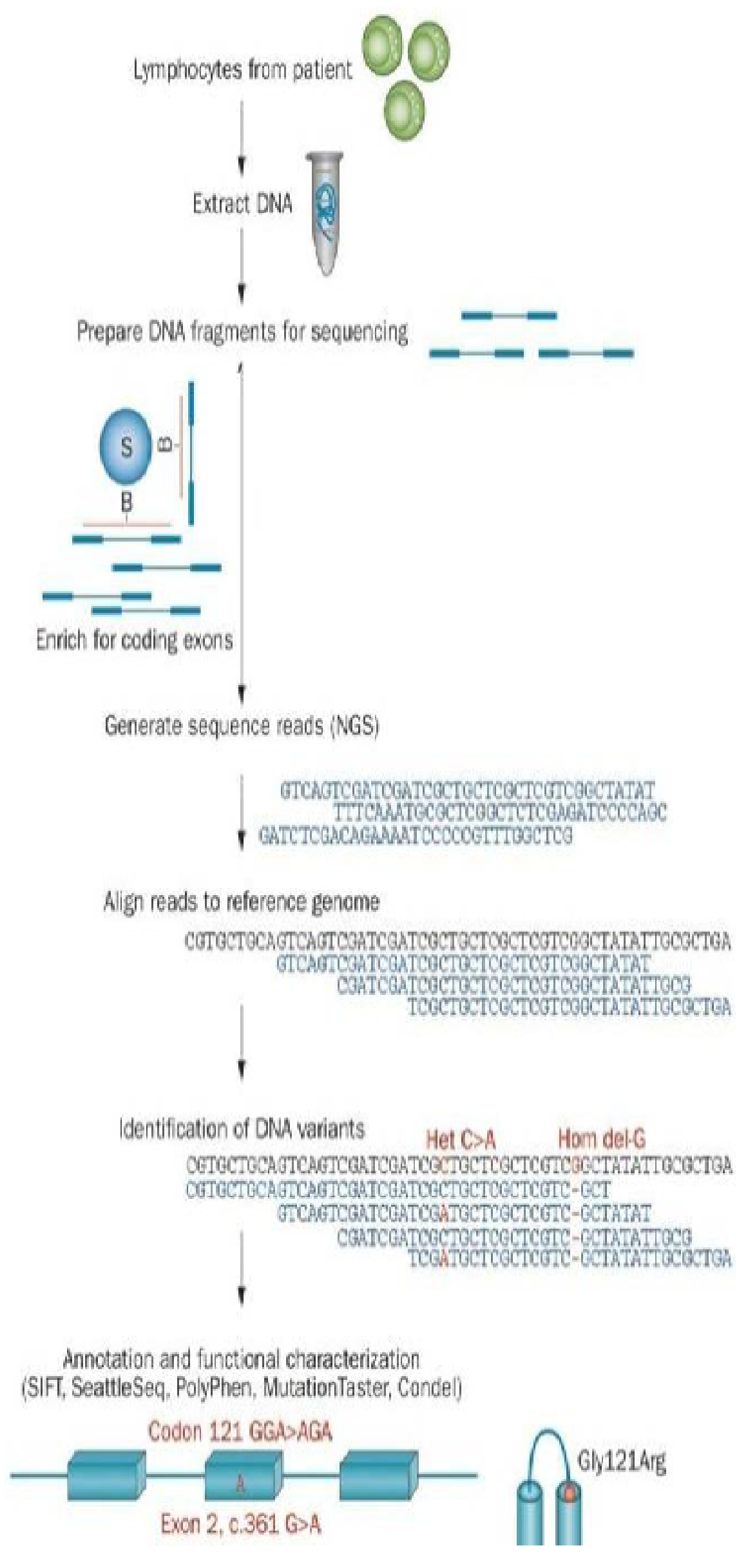
After performing linkage analysis we evaluated the genetic distance in terms of IBS between all cases in the 11 families linked to chromosome 4q26. 16 cases from the closest 8 families (blue rectangle) were prioritized for DNA sequencing(see plot).

For the selection of the internal (intra-family) negative controls we performed an IBD analysis on each of these 8 families and identified the closest relative (for each affected) with less IBD-sharing (genetically discordant) in the region of interest .

We performed full exome sequencing on 16 most informative IgAN patients belonging to 8 non consanguineous families. We also sequenced 8 familial controls.

MATERIALS & METHODS

Exome Sequencing Study


Lymphocytes from patient
Extract DNA
Prepare DNA fragments for sequencing
S
B
B
Enrich for coding exons
Generate sequence reads (NGS)
GTCAGTCGATCGATCGCTGCTCGCTCGTCGGCTATAT
TTTCAAATGCGCTCGGCTCTCGAGATCCCCAGC
GATCTCGACAGAAAATCCCCGTTTGGCTCG
Align reads to reference genome
Identification of DNA variants
Het C>A
Hom del-G
Annotation and functional characterization
(SIFT, SeattleSeq, PolyPhen, MutationTaster, Condel)
Codon 121 GGA>AGA
Exon 2, c.361 G>A
Gly121Arg

We used the illumina HiScanSQ system for exome sequencing. Target regions were captured with the TruSeq Exome Enrichment and randomly fragmented and purified using a QIAquick PCR Purification kit (Qiagen). Adapters were ligated to each end of the fragments, and the resulting DNA library purified (using QIAquick PCR Purification kit).The magnitude of enrichment of captured ligation-mediated PCR products was determined using the Agilent 2100 Bioanalyzer. Next, each captured library was loaded onto the HiScanSQ platform, and paired-end sequencing was performed with read lengths of 101 bp. Sequence reads were mapped to the reference human genome (UCSC Genome Browser hg19) using Burrows-Wheeler. Single nucleotide variants (SNVs) and small insertion/deletions (Indels) were detected following the Best Practices Workflow of Genome Analysis Toolkit (GATK) using two algorithms :Unified Genotyper and Haplotype Caller.

Variant filtration ,Sanger Sequencing and TaqMan Assays

Variants were then annotated with snpEFF (snpEff_v2_0_5) and categorized into four classes (high, moderate, low and modifier) by their functional impact. SeattleSeq SNP annotation web interface was also used for annotating vcf files. Sequence data were filtered against multiple databases, namely, dbSNP137, 1000 Genomes Project using annovar and Minimum Allele Frequency(MAF) of 0.01 was chosen as a cutoff. Variants were then visualized with Integrative Genomics Viewer (IGV). Candidate variants were validated using Sanger Sequencing. Forward and reverse PCR primers were designed for each candidate variant. Purified products were sequenced in both forward and reverse directions on an ABI3730xl DNA analyser (Applied Biosystems). Analysis of sequence data was carried out using the Chromas 2.01 software. Human reference sequences were retrieved from the UCSC Genome Browser. Segregation patterns of validated variants were studied with classical TaqMan® Assays on a StepOnePlus Real-Time PCR System.

RESULTS

Linkage Analysis

We confirmed and refined our previously published linked regions on chromosome 4q26, 6q22-23 and17q21. Linkage signals were also detected on other chromosomes. Eleven families linked to chromosome 4q24-28, that showed the highest evidence for linkage among the identified regions, were studied.

Family_ID	Genotyped (N)	Chr1	Chr2	Chr3p	Chr3q	Chr4	Chr5	Chr6	Chr8	Chr9	Chr17	Chr20	Chr22
1*	4	0.23	0.23	0.22	0.23	0.23	0.23	0.23	0.19	0.23	-0.53	0.23	0.23
3	16	-0.37	0.20	-0.32	-0.37	-0.37	0.20	0.00	0.16	0.20	-0.37	0.20	0.20
4*	18	0.00	0.30	0.28	0.00	0.30	0.09	0.30	0.25	0.00	0.30	0.00	0.00
7*	5	0.00	0.30	0.00	0.30	0.30	0.00	0.30	0.25	0.30	0.00	0.30	0.30
14	4	0.19	0.00	0.00	0.00	0.00	0.30	0.30	0.25	0.30	0.00	0.30	0.30
15*	12	0.23	0.23	-0.45	0.23	0.23	0.23	0.23	0.19	0.23	0.23	0.23	0.23
16	8	0.00	0.00	0.28	0.00	0.00	0.00	0.30	0.25	0.00	0.00	0.30	0.30
17	10	0.00	0.00	0.28	0.00	0.30	0.00	0.00	0.00	0.30	0.30	0.30	0.00
36*	7	0.35	0.35	0.33	-0.23	0.35	-0.23	-0.23	0.29	-0.23	0.35	-0.23	0.35
206*	8	0.30	0.30	0.28	0.00	0.30	0.00	0.00	0.25	0.00	0.00	0.00	0.30
343	10	0.38	-0.06	0.36	0.38	-0.06	0.28	-0.06	0.26	0.28	-0.06	-0.44	-0.06
344	12	0.20	-0.37	0.18	0.20	-0.37	0.20	-0.37	-0.26	0.20	0.20	0.20	-0.37
385*	16	0.35	-0.23	0.33	0.35	0.35	-0.23	-0.23	-0.17	-0.23	0.35	0.35	-0.23
386	5	0.00	0.30	0.00	0.00	0.30	0.30	0.30	0.25	0.00	0.30	0.00	0.00
393	6	0.00	0.00	0.28	0.30	0.30	0.00	0.30	-0.67	0.30	0.30	0.00	0.00
483*	5	0.23	0.23	-0.45	0.23	0.23	0.23	0.23	0.19	0.23	0.23	0.23	0.23
SNP		rs6577472	rs12613771	rs6550478	rs12629557	rs17006113	rs2731665	rs2064687	rs11166903	rs11792985	rs2256020	rs915039	rs2252528
LOD		2.09	1.79	1.59	1.62	2.39	1.61	1.61	1.70	2.12	1.59	1.98	1.79

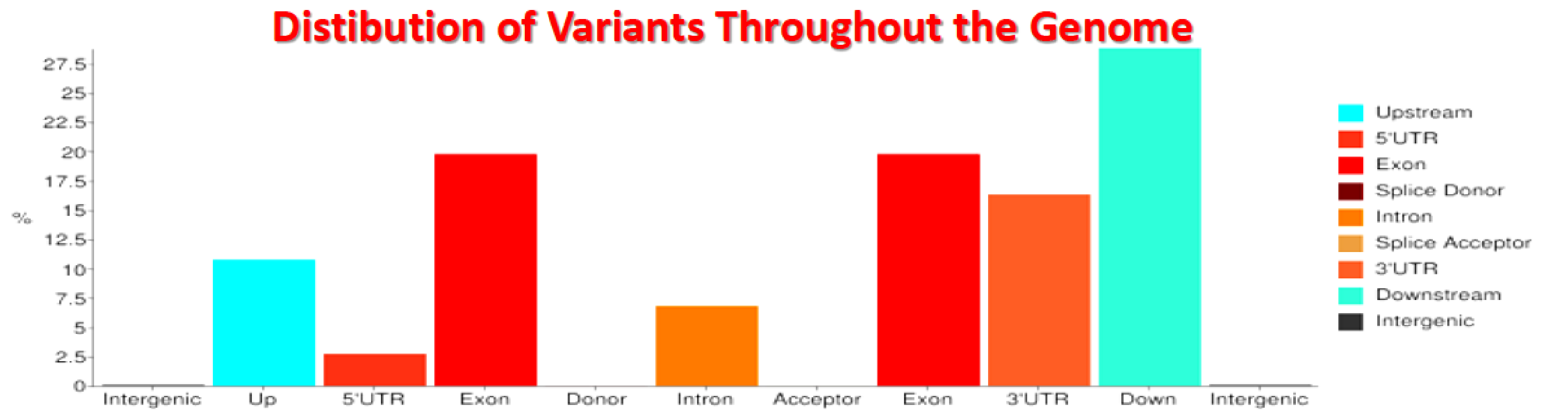
The table shows the families linked to each chromosomal region; positive and negative linkage signals are highlighted in red and blue respectively. The partial LOD contribution for each family is reported in each box. For each chromosome the top LOD score is also reported (* sequenced families).

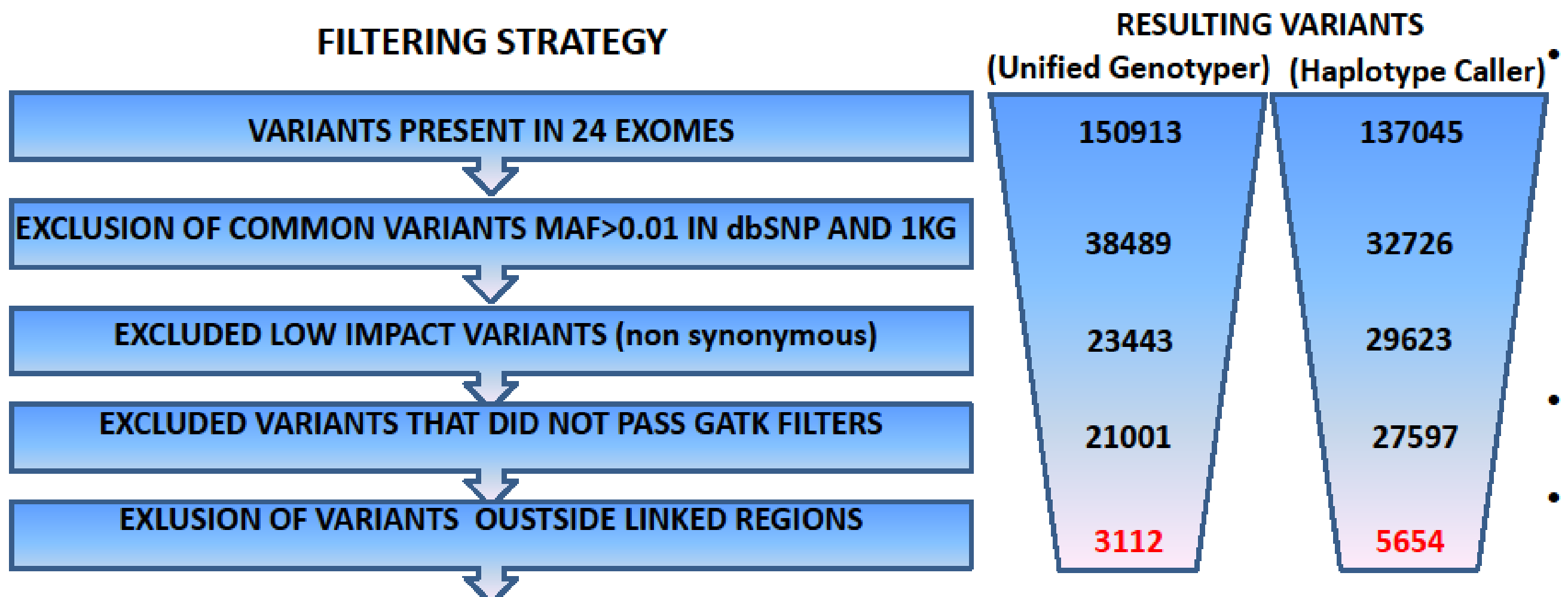
RESULTS

Exome Sequencing Study

FAMILY ID	1			4			7			15			36			206			385			483		
SAMPLE ID	552	156	553	779	781	205	551	597	2073	1725	1728	1724	1741	2373	2400	761	1859	1854	2172	2220	2160	2535	2541	2557
STATUS	IgAN	IgAN	CONTROL	IgAN	IgAN	CONTROL	IgAN	IgAN	CONTROL	IgAN	IgAN	CONTROL	IgAN	IgAN	CONTROL	IgAN	IgAN	CONTROL	IgAN	IgAN	CONTROL	IgAN	IgAN	CONTROL
SEQUENCING AND MAPPING DATA																								
raw data Yield (Mbases)	4363	3658	4170	4084	745	2792	3829	4063	3135	2262	2838	5140	4201	3749	4210	2437	3699	4318	922	3871	4491	5940	3269	4159
n°Reads (M)	43.2	36.2	41.3	40.4	7.4	27.6	37.9	40.2	31.0	22.4	28.1	50.9	41.6	37.1	41.7	24.1	36.6	42.8	9.1	38.3	44.5	58.8	32.4	41.2
% mapped reads to Genome	98.59	98.29	98.59	98.58	98.55	98.48	98.29	96.52	96.13	97.28	97.15	98.2	98.4	98.4	98.58	98.42	98.54	98.46	96.56	97.3	98.27	98.17	97.33	97.38
EXOME CAPTURE																								
% mapped reads to target region	44.25	42.27	42.95	42.64	44.56	42.93	42.27	43.61	43.09	42.75	42.88	43.24	42.49	42.58	43.88	44.75	44.35	42.15	45.85	43.76	42.3	42.95	44.66	43.39
mean coverage target region	26.54	22.36	24.56	23.87	4.54	16.44	22.36	24.11	18.44	13.22	16.76	24.32	23.79	21.35	24.63	14.5	21.74	24.29	5.72	23.15	24.98	24.24	19.93	24.59
mean mapping quality	50.16	50	50.17	50.14	50	49.99	50.13	49.65	49.18	49.75	49.16	49.75	50.00	50.10	50.04	49.97	50.00	50.00	49.08	49.00	50.72	50.22	49.93	50.08
VARIANT CALLING																								
Unified Genotyper	45794	43065	45913	44714	24256	40897	45442	44435	41481	36509	40356	46305	50311	48924	50099	44443	48825	50437	23629	41296	43789	46940	40125	43406
Haplotype Caller	33647	30571	33568	32554	10681	27454	32936	32990	29383	23492	27782	35594	41168	39286	40793	33787	39417	41248	10774	29154	32012	36323	27361	31396

Distibution of Variants Throughout the Genome





FILTERING STRATEGY

VARIANTS PRESENT IN 24 EXOMES

EXCLUSION OF COMMON VARIANTS MAF>0.01 IN dbSNP AND 1KG

EXCLUDED LOW IMPACT VARIANTS (non synonymous)

EXCLUDED VARIANTS THAT DID NOT PASS GATK FILTERS

EXLUSION OF VARIANTS OUSTSIDE LINKED REGIONS

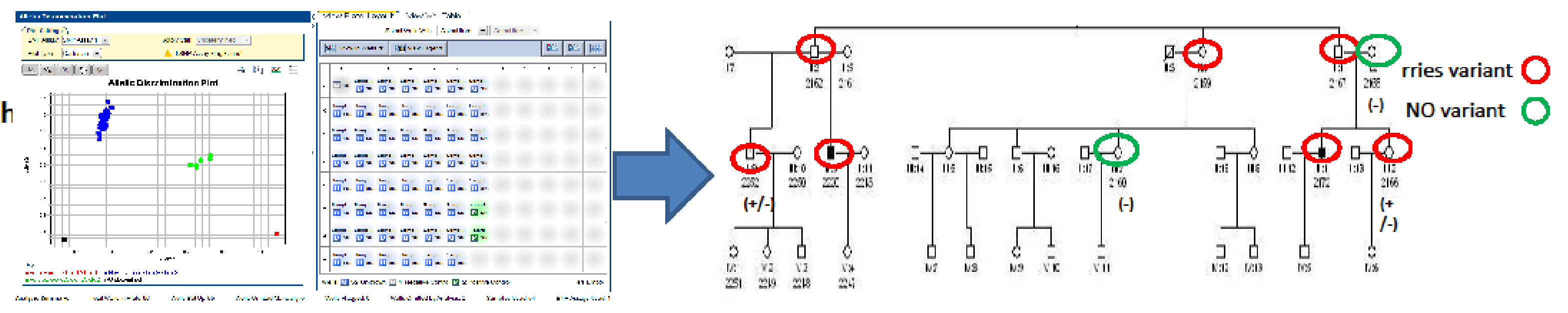
RESULTING VARIANTS (Unified Genotyper)	(Haplotype Caller)
150913	137045
38489	32726
23443	29623
21001	27597
3112	5654

- At the end of the filtering strategy we selected for each family the most important segregating variants based on their potential functional role and on their profile on the IGV.
- 34 variants where prioritized for Sanger sequencing.
- 26 variants were confirmed with Sanger sequencing.

FOR EACH FAMILY WE SELECTED VARIANTS THAT CO-SEGREGATE IN AFFECTED PATIENTS AND ARE ABSENT IN FAMILIAL CONTROLS

FAMILY ID	1		4		7		15		36		206		385		483	
SAMPLE ID	552	156	779	781	551	597	1725	1728	1741	2373	761	1859	2172	2220	2535	2541
VARIANTS	10		9		6		8		18		16		4		12	
FOLLOWED UP VARIANTS	SETD5 CYP11B2 THADA BCLAF1		CHD5 CAMDK2		TG COX10-AS1 INFA21 USP6 RPUSD3		CAAP1 LDLRAP1 CAMDK2 B4GALT5		THRA USP22 NUP210 UBE2G1 CDC27		EDEM1 TOM1L2 ERAL1		CDK12 CHD5 BCLAF1 IL22RA2 MIRLET7B		DFFA UBE4B SCL6A6 JADE SQSTM1 CAMDK2	
SANGER VALIDATED VARIANTS	SETD5 CYP11B2 THADA		CHD5 CAMDK2		TG INFA21 RPUSD3		CAAP1 LDLRAP1 B4GALT5		THRA UBE2G1 CDC27		EDEM1 ERAL1		CHD5 BCLAF1 IL22RA2 MIRLET7B CDK12		DFFA UBE4B SCL6A6 JADE SQSTM1	
SEGREGATION ANALYSIS	SETD5		CAMDK2		TG		CAAP1		THRA		ERAL1		MIRLET7B		DFFA JADE	

Segregation Analysis with TaqMan® Assays



Segregation with the affection status was found in each family. The MAF of candidate variants was also evaluated and checked in 200 population controls. We found that none of the studied variants were present in-house controls. Furthermore the variants were checked in the ExAC BROWSER and we verified that 29 of the followed-up variants were private variants (the remaining showed a MAF < 0.0001). Variants passing the segregation analysis were tested on 200 other familial / sporadic IgAN patients and were not found

CONCLUSIONS

- We confirmed and refined our previously published linked regions, linkage signals were also detected on other chromosomes.
- Within the linked regions , we identified and validated 26 high penetrant risk variants.
- The validated variants were very rare and segregated within affected individuals and were absent in all controls.
- Our exome sequencing data supports the hypothesis that IgAN is a disease characterized by extensive genetic heterogeneity with multiple genes affecting disease onset. This is supported by the finding that a single variant common to all our IgAN families wasn't detected.
- Different variants belonging to the same gene (CHD5 CAMDK2) were also detected.

Segregation analysis on the remaining Sanger validated variants will need to be performed. The functional role of these variants will need to be uncovered.

D) Genetic diseases and molecular genetics.

Sharon N. Cox

DOI: 10.3252/pso.eu.52era.2015

ePosters supported by F. Hoffmann- La Roche Ltd.