Supplementary online materials

Multiple Rare Genetic Variants Co-Segregating with Familial IgA Nephropathy all Convey into a Single Immune-Related Network

Sharon Natasha Cox, $1,288$, Francesco Pesce, $1,38$, Julia Sarah El-Sayed Moustafa, $3,4$, Fabio Sallustio¹, Grazia Serino,^{1,5} Charalambos Kkoufou,³, Annalisa Giampetruzzi,⁶, Nicola Ancona,⁷, Mario Falchi,^{3,4*}, and Francesco Paolo Schena,^{1,2*}, on behalf of the European IgAN Consortium.

¹ Dept of Emergency and Organ Transplantation. University of Bari Aldo Moro, Bari, Italy.

 2 C.A.R.S.O. Consortium, University of Bari, Italy;

³Dept of Genomics of Common Disease, Imperial College London, London, UK.

⁴Dept of Twin Research and Genetic Epidemiology, King's College London, London, UK.

⁵IRCCS "de Bellis", Laboratory of Experimental Immunopathology, 70013 Castellana Grotte, Bari, Italy.

⁶ Dept of Soil, Plant and Food Sciences, University of Bari Aldo Moro, Bari, Italy.

⁷ISSIA, CNR, Bari, Italy,

\$ These authors contributed equally *Francesco Paolo Schena and Mario Falchi shared senior authorship &Corresponding Author

Running Title: Rare Genetic Variants in IgA Nephropathy

Key Words: KIDNEY DISEASE**,** FAMILY MEDICINE**,** GENE POLYMORPHISM**,** GENETICS**,** GLOMERULONEPHRITIS.

Correspondence: Cox Sharon Natasha, C.A.R.S.O. Consortium, University of Bari, 70100 Valenzano (Ba), Italy. Phone: 0039/080-4602121; FAX: 0039/080-4602121; e-mail: sharonnatasha.cox@uniba.it

DESCRIPTION OF SUPPLEMENTAL DATA:

Supplementary Text: Previous linkage study by Bisceglia *et al* as compared to this one.

Analysis of variant segregation in extended families

Supplementary Methods: Sample donors, Microarray Genotyping, whole-exome sequencing, variant selection and Sanger Sequencing .

REFERENCES

Supplementary Figures S1-S7 Supplementary Tables S1-S6.

Supplementary text:

Previous linkage study by Bisceglia et al as compared to this one

The previous linkage studies by Gharavi AG et al [1] and Bisceglia et al [2]) were performed using 400 microsatellites, here we adopted a SNP based strategy using a finer genetic map of 300,000 SNPs. Between studies we have a partial overlap. Families 1, 3, 4, were included in the first linkage study perfomed by Gharavi AG et al (Nat Genet 2000; 26: 354-7). Families 3, 4, 7, 15, 16, 206, 343, 344, 385, 386, 393 were also present in our previous linkage study perfomed by Bisceglia L et al (Am J Hum Genet 2006; 79: 1130-4). The families 14,17,36,483 were added exclusively for this study. Compared to the previous papers, some subjects from each family have been added for genotyping.

Analysis of variant segregation in extended families

We investigated whether variants identified and validated in the previous steps were also co-segregating with urinary abnormalities, i.e. persistent microscopic hematuria and/or proteinuria, in unaffected family members of the extended families used for the WES.

In family 1, all validated variants (*CHD5* 1:g6163696g>a; *FAM179A* 2:g29249757ac>a; *IL22RA2* 6:g137465358c>t; *CDK12* 17:g37689446c>t; *MIRLET7BHG* 22:g46453973t>c) segregated in a family member who had microscopic hematuria (558) and was absent in those displaying negative urinalysis (553).

In Family 4 we found that out of the three validated variants (*IFNA21* 9:g21165905c>t; *ATRAID* 2:g27439820a>g; *RPUSD3* 3:g9880772t>c). *ATRAID* and *RPUSD3* segregated perfectly family members with IgAN and with persistent microscopic hematuria in family members 215, 202, 198. The *IFNA21* variation, on the other hand, was present in only one of the family members with microscopic hematuria (202).

In family 7, we evaluated the segregation of the variant 1:g25894878c>g within the *LDLRAP1* gene. This variant was absent in family members that had normal urinalysis (579, 2073, 1202) and was carried by a family member who presented some episodes of microscopic hematuria (581).

In family 15, five validated variants (*DFFA* 1:g10527277g>c; *JADE1* 4:g129783008t>a; *SLC6A6* 3:g14528787a>g; *SQSTM1* 5:g179264117a>g; *UBE4B* 1:g10190827c>t) segregated in subjects 1731 and 1720, both characterized by persistent microscopic hematuria and/or proteinuria. Only the non-synonymous coding variant S365R *JADE1* segregated in all three subjects characterized (1723, 1731 and 1720) by urinary abnormalities.

In family 36, all tested variants (*CYP11B2* 8:g143993975c>t, *SETD5* 3:g9515095c>a, *PTPRG* 3:g62063912g>a, *THADA* 2:g43455302g>a) segregated in two obligate carriers (2396 and 2397) and in a family member characterized by episodes of microscopic hematuria.

For family 206, the variants *THRA* 17:g38233146c>t, *UBE2G1* 17:g4173166g>a and *CDC27* 17:g45197967a>g that segregated with the affection status also segregated with an individual characterized by persistent microscopic hematuria (1858), while only *UBE2G1* gene variant segregated with a second individual characterized by persistent microscopic hematuria.

In family 385, the *CAMKD2* 4:g114374628t>a and the *CHD5* 1:g6162250g>gac variant was evaluated for segregation in the extended family. This variant was present in family members with persistent microscopic hematuria and/or proteinuria and in the subject 2162, an obligate carrier with a normal phenotype.

For family 483, we evaluated the co segregating variant 3:g5259973a>g within the *EDEM1* gene. This variant was absent in family members who had normal urinalysis apart from the obligate carrier 2542 with normal urinalysis.

Supplementary methods:

Sample donors

The genome-wide linkage analysis involved 34 biopsy-proven familial IgAN patients and 112 relatives from 16 Italian kindreds of South Italian ancestry (Table 1, Figure S1). Recruitment strategies and criteria for diagnosis have been reported elsewhere and are available on the European IgAN Consortium Web site (www.igan.net)[3]. Briefly , familial IgAN was diagnosed when at least 2 family members had biopsy-proven IgAN, the remaining family members were all checked for urinalysis. Some family members were affected by persistent microscopic hematuria and/or proteinuria and were depicted with an unknown status. Unaffected family members had at least three documented negative urinalysis. An independent cohort of 240 biopsy-proven IgAN patients and 113 HBD were included in the study for custom TaqMan SNP genotyping assays. Written informed consent was obtained from all study participants. The study was carried out according to the principles of the Declaration of Helsinki and was approved by the local Institutional Ethics Review Board.

Microarray Genotyping

DNA was isolated from whole blood of IgAN patients and HBD by Qiagen QIAamp DNA Blood Midi Kit (Qiagen Srl, Milan, Italy). DNA was quality-checked on agarose gels and quantified using a microvolume spectrophotometer (NanoDrop 2000; Thermo Fisher Scientific Inc.). Genotyping was performed using Illumina HumanCytoSNP-12 BeadChip containing 300,000 selected tag SNPs with a median marker spacing of 6.2 kbases. Data was exported from Genome Studio Software. Microarray data of the IgAN patients and relatives genotyped on Illumina HumanCytoSNP-12 have been previously published by our group under accession number GSE44974 at the GEO (http://www.ncbi.nlm.nih.gov/geo/)[4].

Whole-exome sequencing

We performed whole exome sequencing on 16 most informative IgAN patients belonging to 8 non-consanguineous families and 8 intra-familial controls. For the selection of the internal (intra-family) negative controls, we performed an Identical By Descent (IBD) analysis on each of these 8 families and identified the closest relative (for each affected) with the least IBD-sharing (genetically discordant) in the region of interest.[5, 6] Three micrograms (µg) of genomic DNA was used for generating each exome library. Genomic DNA was randomly sheared using a Nebulizer (Life Technologies), adapters were ligated to each end of the fragments and purified using a QIAquick PCR Purification kit (Qiagen). Target regions were captured with the TruSeq Exome Enrichment and the illumina HiScanSQ system was used for sequencing. 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 101bp. Image analysis was performed with default parameters of Illumina RTA v1.13 pipeline and demultiplexing was performed with CASAVA 1.8.2 (Illumina). Sequence reads were mapped to the reference human genome (UCSC Genome Browser hg19) using the Burrows-Wheeler aligner (BWA; version 0.5.9 r16)[7] with default parameters. Alignments were converted from sequence alignment map (SAM) format to sorted, indexed binary alignment map (BAM) files (SAMtools version 0.1.18; http://sourceforge.net). The Picard tool was used to remove duplicate reads.

Efficiency of alignment and qualimap was used for evaluating alignment data (qualimap v2.1.2, http://qualimap.bioinfo.cipf.es/). The Best Practices Workflow of Genome Analysis Toolkit (GATK version 2.7-4, http://www.broadinstitute.org) was used for improving the alignments and for genotype calling with recommended parameters[8]. BAM files were re-aligned with the GATK IndelRealigner, and base quality scores were recalibrated by the GATK base quality recalibration tool. Genotypes were called at first with the GATK UG (version 2.7-4) and the GATK VariantRecalibrator tool was used to score variant calls by a machine-learning algorithm and to identify a set of high-quality variants using the Variant Quality Score Recalibration (VQSR) procedure. GATK was used to filter high-quality variants with hard filtering criteria (variant confidence score ≥30, mapping quality ≥40, read depth ≥5, and strand bias FS filter <60). At a later date, our exome data was also re-processed using the newer GATK algorithm HC (version v3.3). To be conservative, we decided to retain and evaluate both variant lists generated by HC and UG. These two tools are based on different algorithms[9]. HC in variable regions of the genome discards the existing mapping information and reassembles the reads (*de novo* assembly of haplotypes). UG uses a Bayesian genotype likelihood model and estimates the most likely genotype calls and simply looks for a coincident haplotype event in the reads. Both methods evaluate haplotypes using an affine gap penalty Pair Hidden Markov Model[10]. Variants were then annotated with the software snpEFF[11] (snpEff $v2$ 0 5, http://snpeff.sourceforge.net/download.html) and categorized into four classes (high, moderate, low and modifier) and the functional impact of coding variants was also predicted. Low impact variants were predicted by snpEFF and filtered out as they were synonymous coding and "assumed to be mostly harmless or unlikely to change protein behaviour" as decribed in manual (http://snpeff.sourceforge.net/SnpEff_manual.html). Sequence data were filtered against multiple databases, using annovar (http://annovar.openbioinformatics.org, version 2013Aug23) and Minor Allele Frequencies (MAF) of the called variants were compared against dbSNP 137 (ftp://ftp.ncbi.nih.gov/snp/) and 1000 Genomes Project, where we filtered against the European cohort (April 2002 release, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/). Then, variants were visualized with Integrative Genomics Viewer (IGV, version 2.3.36)[12]. The concordance between the genotypes of the variants identified through exome sequencing and bead SNP genotyping was evaluated using the PLINK[13] program. Concordance was of 96% and 94% for UG and HC, respectively.

Whole-exome sequencing data for all families have been deposited in the SRA database of NCBI (http://www.ncbi.nlm.nih.gov/sra) and are available under the following study accession No. SRP061415

Variant selection and Sanger Sequencing

Segregating variants in affected individuals were selected based on their scaled C-scores from the Combined Annotation Dependent Depletion (CADD) webserver (http://cadd.gs.washington.edu)[14]. The phred-like scores ("scaled C-scores") ranging from 1 to 99, are based on the rank of each variant relative to all possible 8.6 billion substitutions in the human reference genome (http://cadd.gs.washington.edu/info) and candidate variants were validated using Sanger Sequencing. The variants were validated at two different sites in the UK and in Italy. For the sequencing carried out in the UK, PCR primers were designed using the Primer 3 software. PCR amplification was carried out using the Clontech Advantage 2 PCR kit (Takara), following the manufacturer's recommended conditions. PCR products were purified prior to sequencing using exoSAP-IT (Affymetrix Inc), following the manufacturer's recommended conditions. For the sequencing carried out in Italy, forward and reverse PCR primers were designed for each candidate variant using the Primer Designer™ Tool (Life Technologies). PCR amplification was performed using AmpliTaq Gold® 360 DNA Polymerase (Life Technologies, Italia) and products were checked on agarose gels 2% and purified using the QIAquick PCR Purification kit (Qiagen). Purified products were sequenced in both forward and reverse directions on an ABI 3730xl 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.

REFERENCES1 Gharavi AG, Yan Y, Scolari F*, et al.* IgA nephropathy, the most common cause of glomerulonephritis, is linked to 6q22-23. *Nat Genet* 2000; **26:** 354-7.

² Bisceglia L, Cerullo G, Forabosco P*, et al.* Genetic heterogeneity in Italian families with IgA nephropathy: suggestive linkage for two novel IgA nephropathy loci. *Am J Hum Genet* 2006; **79:** 1130-4.

³ Schena FP, Cerullo G, Torres DD*, et al.* The IgA nephropathy Biobank. An important starting point for the genetic dissection of a complex trait. *BMC Nephrol* 2005; **6:** 14.

⁴ Sallustio F, Cox SN, Serino G*, et al.* Genome-wide scan identifies a copy number variable region at 3p21.1 that influences the TLR9 expression levels in IgA nephropathy patients. *Eur J Hum Genet* 2015; **23:** 940-8.

5 Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* 2002; **30:** 97-101.

6 Powell JE, Visscher PM, Goddard ME. Reconciling the analysis of IBD and IBS in complex trait studies. *Nat Rev Genet* 2010; **11:** 800-5.

7 Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009; **25:** 1754-60.

8 DePristo MA, Banks E, Poplin R*, et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 2011; **43:** 491-8.

9 Pirooznia M, Kramer M, Parla J, Goes FS, Potash JB, McCombie WR, Zandi PP. Validation and assessment of variant calling pipelines for next-generation sequencing. *Hum Genomics* 2014; **8:** 14.

10 Krogh A, Brown M, Mian IS, Sjolander K, Haussler D. Hidden Markov models in computational biology. Applications to protein modeling. *J Mol Biol* 1994; **235:** 1501-31.

11 Cingolani P, Platts A, Wang le L*, et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* 2012; **6:** 80-92.

12 Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* 2013; **14:** 178-92.

13 Purcell S, Neale B, Todd-Brown K*, et al.* PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; **81:** 559-75.

14 Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 2014; **46:** 310-5.

Supplementary Figure 1. Sixteen multiplex families included in the linkage study, the red bars represent 146 genotyped subjects. Squares and circles represent males and females, respectively; arrows indicate probands and the slash indicate deceased individuals. Filled and unfilled symbols indicate IgAN affected and unaffected individuals, respectively. The symbols with a dot indicate individuals who have not received or have discordant urinalysis. Symbols with a vertical line indicate individuals with documented urinary abnormalities (persistent microscopic hematuria and/or proteinuria). Families 1, 3, 4, were present in the first linkage study perfomed by Gharavi AG et al (Nat Genet 2000; 26: 354-7). Families 3, 4, 7, 15, 16, 206, 343, 344, 385, 386, 393 were also present in our previous linkage study perfomed by Bisceglia L et al (Am J Hum Genet 2006; 79: 1130-4). Families 14,17,36,483 were new and added exclusively for this study.

393

Supplementary Figure 2. Plot of LOD score statistics from NPL analysis for the chromosomes in which the score exceeded the 1.5 level.

Supplementary Figure 3. Pedigrees included in the exome sequencing study

Eight pedigrees were included in the exome sequencing study. Squares and circles represent males and females, respectively; the arrows indicate probands and the slashes indicate deceased individuals. Filled and unfilled symbols indicate IgAN affected and unaffected individuals, respectively. The symbols with a dot indicate individuals who had not received or had intermittent microscopic hematuria. Symbols with a vertical line indicate individuals with urinary abnormalities. Horizontal grey lines show individuals on whom whole-exome sequencing has been performed.

Supplementary Figure 4. Number of variants called by Unified Genotyper (UG) and Haplotype Caller (HC). Venn diagram depicts the number of unique and common variants when called by UG and UC (A). A total of 48598 variants have been called by HC and UG, of these 27724 (57%) were in common, 13735 (28%) were unique to HC and 7139 (15%) were unique to $\dot{\mathsf{U}}$ G. (B)

A

B

Supplementary Figure 5. Percentage distributions of gene variantions as depicted in the output of SnpEff: Variant analysis. The percentage of variations within each gene region did not differ when called by Unified Genotyper (A) and Haplotype Caller (B). A)

Supplementary Figure 6. The two variants predicted within the *BCLAF1* gene were excluded as they were actually determined by the presence of a deletion in a repeated (AAAAAC)n region (and not by a G/A substitution). Heterozygous deletion found in the repeated (AAAAAC)n region of the BCLAF1 gene. Deletion detected in the flipped reverse sequence of two affected individuals (552, 156) and absent in the intra-familal control

Supplementary Figure 7. We evaluated BCLAF1 complex genomic region in six unrelated HBD. We found that the heterozygous deletion in the repeated (AAAAAC)n region of the BCLAF1 gene was also found in three unrelated healthy blood (control1, Control2, control3) and absent in another three unrelated healthy blood donors, for this reason we decided to exclude this variation from further analysis. genomic region in six
e repeated (AAAAAC)n
d healthy blood donors

Supplementary Table 1 - Summary of the exome sequencing results

^aTarget region contains 62Mb of genomic DNA including exons, flanking 3'UTR, 5'UTR, predicted microRNA and other non coding RNA. b number of variants called for each individual by two distinct algorithms from the Genome Analysis Toolkit (GATK): Haplotype Caller and Unified Genotyper. Abbreviations: CNTR: Intra-familial control; Mbases: mega bases; M:Million.

Locus	TOP LOD SCORE	^b LEFT $LOD=0$	^c Top Hit SNP	PRIGHT $LOD=0$	LEFT bp position	Top Hit Bp position	RIGHT bp position
1p36	2.1	rs3094315	rs6577472	rs7555884	752566	8206130	35373878
2p21	1.8	rs2702068	rs12613771	rs7600065	24689512	44490660	61019699
3p22	1.6	rs17440919	rs6550478	rs12054271	1238954	37526013	65027043
3q29	1.6	rs6778567	rs12629557	rs13083786	192611340	194057168	197811684
4q26	2.4	rs7693338	rs17006113	rs7658837	94766621	120714886	151117651
5q35	1.6	rs17738444	rs2731665	rs34865693	169444363	176857270	180695849
6q22	1.6	rs723318	rs2064687	rs13215778	107280314	124376164	146900563
8q24	1.7	rs7813493	rs11166903	rs35756786	130208039	140282987	146245372
9p24	2.1	rs10814410	rs11792985	rs10813550	46587	7664428	31351194
17q21	1.6	rs11651767	rs2256020	rs16977176	4073398	46050635	70321327
20q13	2.0	rs2206633	rs915039	rs11697347	42736729	56596617	62903830
22q13	1.8	rs5754779	rs2252528	rs9615919	34369696	39219932	48940644

Supplementary Table 2. Genomic intervals considered for variant filtrationa

^aHuman reference genome hg19. b^sNPs where the LOD score dropped to zero.°SNP where the top LOD score was observed.

Supplementary Table 3. List of co-segregating gene variants identified by exome sequencing

Ref = reference allele; Alt = alternate allele; CADD SCORE: PHRED-like (-10*log10(rank/total)) scaled C-score ranking a variant
relative to all possible substitutions of the human genome (8.6x10^9)

Supplementary Table 4. Left and right PCR primers designed to amplify the genomic region containing the variant of interest used for Sanger validation

***** primers designed using the Sanger Primer Designer™ Tool (Life Technologies)

Supplementary Table 6. List of potential drugs targeting the network.

