A novel nonsense variant in SUPT20H gene associated with Rheumatoid Arthritis identified by Whole Exome Sequencing of multiplex families

Maeva Veyssière, Javier Perea, Laétitia Michou, Anne Boland, Christophe Caloustian, Robert E Olaso, Jean-François Deleuze, Francois Cornelis, Elisabeth Petit-Teixeira, Valérie Chaudru

PLOS ONE

RESEARCH ARTICLE

A novel nonsense variant in *SUPT20H* gene associated with Rheumatoid Arthritis identified by Whole Exome Sequencing of multiplex families

Maëva Veyssiere[1]*, Javier Perea[1], Laetitia Michou[2], Anne Boland[3], Christophe Caloustian[3], Robert Olaso[3], Jean-François Deleuze[3], François Cornelis[4], Elisabeth Petit-Teixeira[1], Valérie Chaudru[1]

1 GenHotel—Univ Evry, University of Paris Saclay, Evry, France, **2** Division of Rheumatology, Department of Medicine, CHU de Québec-Université Laval, QC, Québec, Canada, **3** Centre National de Recherche en Génomique Humaine—François Jacob Institute, CEA, Evry, France, **4** GenHotel-Auvergne, Auvergne University, Genetic Department, CHU Clermont-Ferrand, Clermont-Ferrand, France

* maeva.veyssiere@univ-evry.fr

OPEN ACCESS

Citation: Veyssiere M, Perea J, Michou L, Boland A, Caloustian C, Olaso R, et al. (2019) A novel nonsense variant in *SUPT20H* gene associated with Rheumatoid Arthritis identified by Whole Exome Sequencing of multiplex families. PLoS ONE 14(3): e0213387. https://doi.org/10.1371/journal.pone.0213387

Editor: Obul Reddy Bandapalli, German Cancer Research Center (DKFZ), GERMANY

Received: October 10, 2018

Accepted: February 19, 2019

Published: March 7, 2019

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by the ARTHRITIS Fondation COURTIN (http://www.fondation-arthritis.org/) and by the Genopole (https://www.genopole.fr/). Both were granted to EPT. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

The triggering and development of Rheumatoid Arthritis (RA) is conditioned by environmental and genetic factors. Despite the identification of more than one hundred genetic variants associated with the disease, not all the cases can be explained. Here, we performed Whole Exome Sequencing in 9 multiplex families (N = 30) to identify rare variants susceptible to play a role in the disease pathogenesis. We pre-selected 77 genes which carried rare variants with a complete segregation with RA in the studied families. Follow-up linkage and association analyses with pVAAST highlighted significant RA association of 43 genes (p-value < 0.05 after 10^6 permutations) and pinpointed their most likely causal variant. We re-sequenced the 10 most significant likely causal variants (p-value $\leq 3.78*10^{-3}$ after 10^6 permutations) in the extended pedigrees and 9 additional multiplex families (N = 110). Only one SNV in *SUPT20H*: *c.73A>T (p.Lys25*)*, presented a complete segregation with RA in an extended pedigree with early-onset cases. In summary, we identified in this study a new variant associated with RA in *SUPT20H* gene. This gene belongs to several biological pathways like macro-autophagy and monocyte/macrophage differentiation, which contribute to RA pathogenesis. In addition, these results showed that analyzing rare variants using a family-based approach is a strategy that allows to identify RA risk loci, even with a small dataset.

Introduction

Rheumatoid Arthritis (RA) is one of the most frequent autoimmune disease, affecting 0.3 to 1% of the worldwide population. Since the discovery of HLA locus as a risk factor for autoimmune diseases [1] and specifically *HLA-DRB1* for RA, more than 100 RA genetic factors were

Competing interests: The authors have declared that no competing interests exist.

identified by Genome Wide Association Studies (GWASs) [2,3]. However, the effect of these genetic risk factors is too weak to explain the entire RA genetic component. Indeed, the heritability attributed to *HLA-DRB1* shared-epitope (SE) alleles was estimated between 11% [4] and 37% [5]. While GWASs loci identified outside the HLA locus only explain an additional five percent of RA heritability [6].

Several hypotheses have been proposed to explain the unknown part of this complex disease genetic component. One of these hypotheses relies on the fact that rare variants, which are poorly detected by GWASs, contribute to the risk of complex diseases. During the last decade, the development of Next Generation Sequencing has facilitated the detection of such variants. Hence, several studies used exome sequencing to identify rare to low frequency variants and evaluate their contribution to RA risk. For this purpose, two studies used a candidate gene approach based on exome sequencing [7,8]. A first study, based on a population of European ancestry, showed an aggregation of non-synonymous rare variants contributing to RA risk into *IL2RA* and *IL2RB* loci [7]. Another one, which studied Korean RA cases and healthy controls, allowed the identification of a weak association of rare missense variants in 17 different genes [8]. In this latter group, the *VSTM1* gene gave rise to an over-expression of its mRNA product in RA cases [9]. However, both studies restricted their analysis to a limited number of candidate genes. Thus, they may have missed variations contributing to RA risk present outside the loci of those candidates. Other studies [10,11] succeeded at identifying new RA risk loci by analyzing rare coding variants extracted after Whole Exome Sequencing (WES). Thus, an association between RA and *PLB1* gene [11] and several other genes involved in the production of reactive oxygen species (ROS) such as *NDUFA7* and *SCO1* [10] has been observed using respectively a familial-based strategy and a case-control analysis. Both studies, conducted in non-European populations, support the strategy of sequencing the whole exome of RA cases to identify new RA candidate variants.

In our study, we aimed at identifying new loci associated with RA in the French population by focusing our research on rare coding variants. For this purpose, we sequenced RA cases and healthy relatives from 9 multiplex families which carried *HLA-DRB1* risk alleles. However, even in the cases who carried the SE alleles, we can observe some clinical and genetic heterogeneity. Hence, we should be able to identify both genetic factors modulating the effect of *HLA-DRB1* SE and acting independently from HLA loci.

Results

Overview

In this study, we sequenced the whole exome of 19 RA cases and 11 healthy individuals belonging to 9 multiplex families (discovery set). We applied a three steps strategy (Fig 1) to prioritize the sequenced loci and identify RA risk variants. First, we selected genes carrying rare variants which were family specific and with a high damaging predicted effect. Then, we assessed the potential combined effect of rare and low-frequency variants in these genes on RA risk using pVAAST [12]. Finally, for the 10 genes with the strongest genetic association, we re-sequenced the leading effect variant in the validation set. This final step allowed us to validate their family specific co-segregation with RA.

Description of the discovery set

The male: female ratio was 6:13 in the discovery set selected for WES (described in Table 1), which is similar to the ratio observed in RA affected populations [13,14]. All the RA-affected individuals carried at least 1 shared epitope allele of *HLA-DRB1* (18 of 19 had 2 SE alleles). Fifteen affected participants (79%) were positive for Anti-Citrullinated Peptide Antibodies

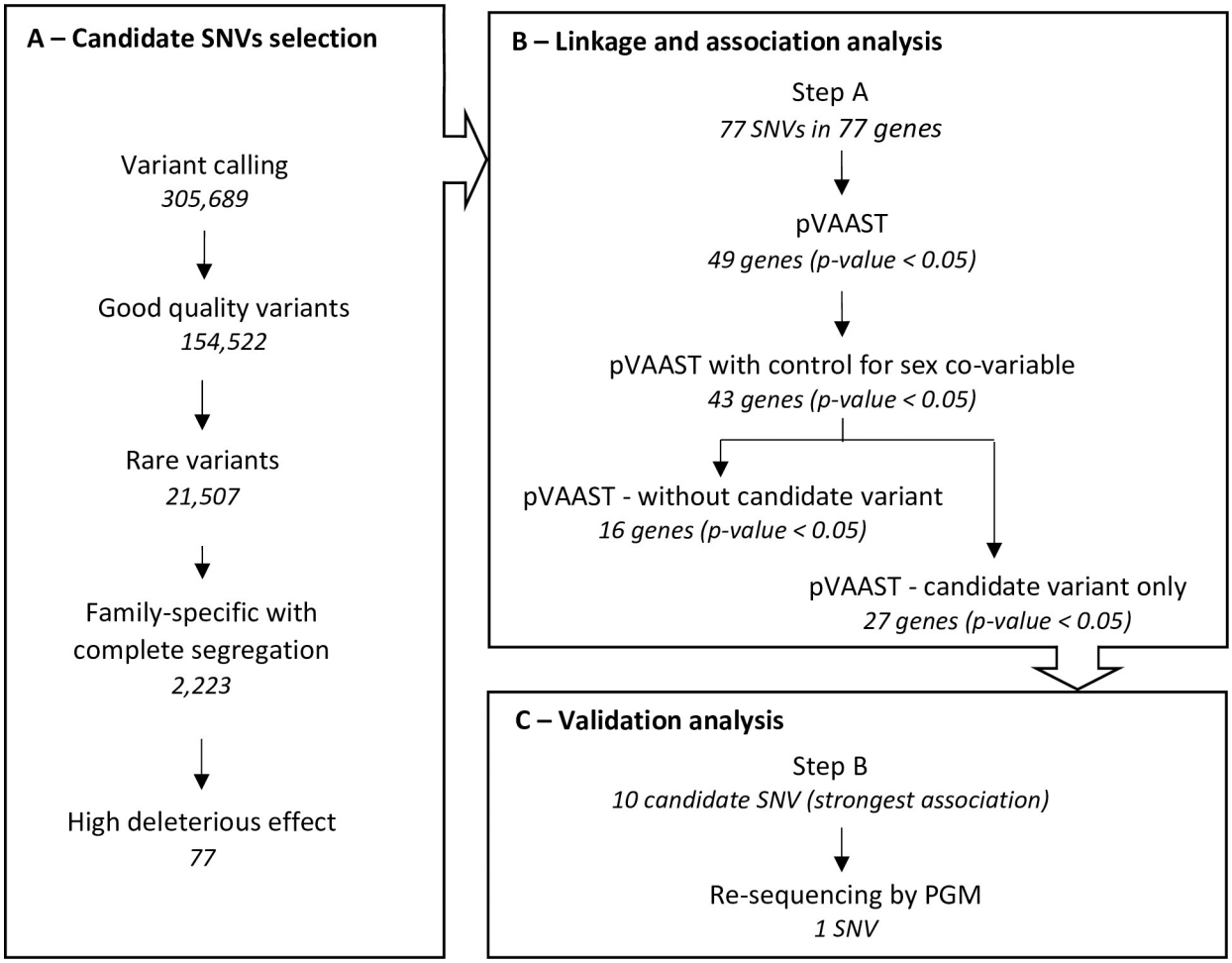

Fig 1. Description of the study design. This schema of the study presents the 3 main steps of the analysis and, the resulting number of variants selected through each sub-analysis. (A) The selection of candidate variants was performed on the discovery set and resulted in the identification of 77 candidate SNVs (B) The association analysis was applied on the discovery set and 98 controls extracted from the 1000 genomes project. The 43 genes significantly associated with RA were re-tested after removing the candidate variant identified in step A and, the candidate variant was tested alone (C) The SNVs with the strongest RA association were re-sequenced in the extended families of the discovery, plus 9 new multiplex families.

https://doi.org/10.1371/journal.pone.0213387.g001

(ACPA) and Rheumatoid Factor (RF), one (5%) was positive for the RF alone and two (11%) were negative for both. These two seronegative cases belonged to the same family (referred as family 5). The affected members of families 3, 4 and 9 were diagnosed with early RA (the mean age at diagnosis was of 31, 36 and 26 years old, respectively).

To check whether the 98 selected controls would not inflate the type I error during the burden test, we performed a PCA based on these samples and the discovery set. The genetic variability (<2.12% of the total variability) observed in the PCA (Fig 2) was due to the variability between the selected families and not between families and controls.

Whole Exome Sequencing and RA risk candidate variants detection

More than 91% of the targeted regions were sequenced with a coverage $\geq$ 30X. We extracted from these sequences 154,644 high-quality variants, including 64,747 (42%) in exons. We observed a mean of 26,679 exonic variants per sample and a SNV transition/transversion ratio equal to 3.01 which is consistent with previous studies [15].

Table 1. Description of the discovery set.

Pedigree id	Affected by RA[a]											Unaffected by RA[a]										
	All	*Male*	*Female*	*ACPA[b]*			*RF[c]*			*Age of diagnosis*		*All*	*Male*	*Female*	*ACPA[b]*			*RF[c]*			*Sampling age*	
				-	+	*NA*	-	+	*NA*	*≥40 years old*	*<40 years old*				-	+	*NA*	-	+	*NA*	*≥40 years old*	*<40 years old*
Only RA reported																						
2	2		2		2			2		1		1		1	1			1			1	
3	3	2	1		3			3		1	1	1		1	1			1				
4	2	2			2			2		1	1	1		1	1			1			1	
5	2		2	2			2			2		1		1	1			1			1	
6	2		2		2			2				1	1			1		1			1	
7	2		2		2			2				2	1	1	1	1		2			1	
RA and other AID reported																						
9	2	1	1		2			2			2											1
10	2		2	1	1			2		1	1	2	1	1	2			2			1	1
11	2	1	1		1	1		1	1			2		2	2			2			1	1

[a]: Rheumatoid Arthritis

[b]: Anti-Citrullinated Peptides Antibodies

[c]: Rheumatoid Factor

https://doi.org/10.1371/journal.pone.0213387.t001

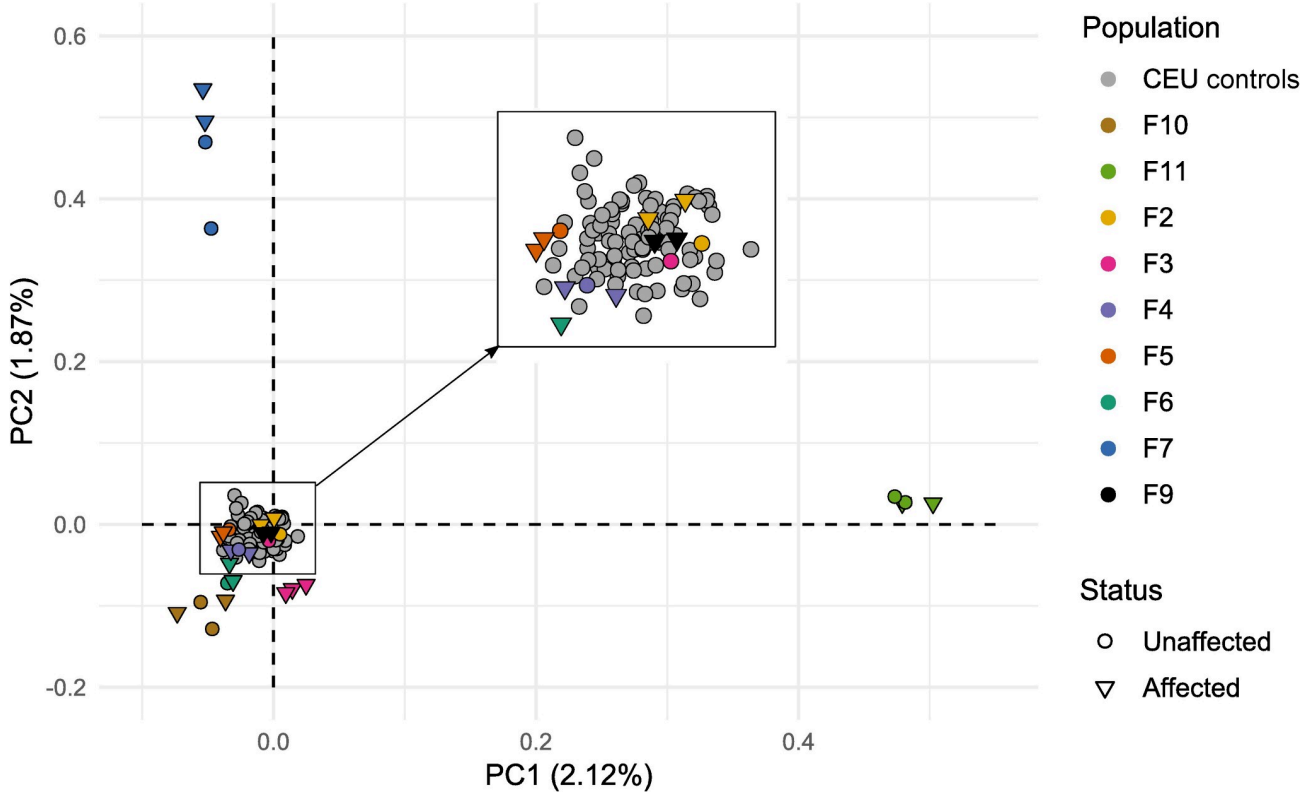

Fig 2. Principal component analysis (PCA) results of samples processed for pVAAST analysis. The PCA is based on the 30 samples from the discovery set and 98 CEU controls extracted from the 1000 genomes databases. This plot represents the 2 first principal components which respectively account for 2.12% and 1.87% of the genetic variability. The color code represents the population source: each family sequenced by WES has its own color described in the legend, the CEU controls are in grey. The shape of each dot represents the RA status of the represented individual.

https://doi.org/10.1371/journal.pone.0213387.g002

Under the hypothesis that rare genetic variants linked to RA would segregate with the affected phenotype within the multiplex families, we selected in the pool of 21,507 rare variants (MAF < 1%), 2223 family-specific variants shared by all the RA-affected relatives (but absent from unaffected members). Knowing that rare variant with high predicted biological effect may contribute to the genetic predisposition of common disease, we extracted, from the 2223 variants, 77 SNVs with a high deleterious predicted effect on protein. We based our evaluation of this impact on SNPeff ("HIGH" or "MODERATE") and CADD phred score (≥30). All these 77 rare family-specific variants were heterozygous in affected carriers. They were detected in 77 different genes, not previously associated with RA.

Evaluation of candidate variant association with RA

We reduced the set of candidate loci to 49 genes (64% of the genes tested with pVAAST) which were significantly associated with RA ($pVAAST_{p\text{-}value} < 0.041$ after 10^6 permutations). Considering the difference in RA prevalence rate between men and women, we performed pVAAST test while controlling for the variable "sex". A total of 43 genes still had a significant association with RA ($pVAAST_{p\text{-}value} < 0.05$– see Table 2).

We categorized the 43 genes into two groups by looking at the association scores of their variants (Table 3). Sixteen genes (group 1) carried several variants, including our best rare candidate SNV, which contributed to the score of the gene ($SNV_{score} > 0$). In the other twenty-seven genes (group 2), only the family-specific variant participated to the gene score. In all the 43 genes, the variant with the highest score is the rare family-specific SNV that was selected previously. Thereafter, we will refer to this variant as the candidate variant. To validate the leading effect of the candidate variants, we performed again the burden test by including first, only the candidate SNVs, and then, all variants except these candidates (results in Table 3). All the 43 candidate SNVs, except *SUSD5*: *c.526C>T* ($pVAAST_{p\text{-}value} = 0.21$) and *SMYD5*: *c.662G>A* ($pVAAST_{p\text{-}value} = 0.067$), were significantly associated with RA ($pVAAST_{p\text{-}value} \leq 0.037$). Concerning the test without the candidate SNV, three genes in the first group (20%), *SUSD5*, *SMYD5* and *MNS1*, were still significantly associated with RA ($pVAAST_{p\text{-}value} \leq 0.036$). So, several rare variants in our dataset contributed to the association of these genes with RA. However, none of the genes in the second group were associated with RA after excluding the candidate variant. This observation confirmed that the observed RA-association within these genes was in fact due to the family-specific variant.

Validation in extended pedigrees

We selected the top 10 RA candidate variants, with the strongest association with RA, for re-sequencing to confirm their family specificity and their co-segregation with RA (Table 4). We analyzed further 8 out of 10 variants for which we were able to produce sequences and, we validated the familial specificity for all of them except *PCCA*. Four variants (in *SUPT20H*, *SLC9A6*, *TIMM44* and *NEK1* genes) were shared by all affected members in the families who carried them. However, only one candidate variant showed a complete segregation with RA. This variant is a heterozygous non-sense SNV, *c.73A>T (p.Lys25*)*, located in *SUPT20H* gene which introduces a premature codon stop at the beginning of the gene (in the fourth exon) (Fig 3). This variant, not reported to date, has been deposited in ClinVar database prior to publication.

Genotyping of *SNV c.73A>T* (p.Lys25*) in family 3 and in trios

The SNV *c.73A>T (p.Lys25*)* carried by *SUPT20H* was genotyped using a customized assay and digital PCR. First, genotyping results in family 3 validated the presence of the rare A allele

Table 2. Genes showing association signal in pVAAST analysis.

Gene		*pVAAST result*[b]		*Number of variants*	
Name	***Chromosome: position***[a]	***Without co-variable control***	***With sex co-variable control***	***Contributing to association***	***In the gene***
SUSD5	3: 33,191,537–33,260,707	0.0003 (22.22)	0.00063 (22.22)	3	7
TMOD2	15: 52,043,758–52,108,565	0.0094 (7.07)	0.002 (7.07)	1	1
AURKB	17: 8,108,056–8,113,918	0.00053 (15.93)	0.0023 (15.93)	2	4
NEK1	4: 170,314,426–170,533,780	0.0021 (8.56)	0.0031 (8.56)	1	7
DHRS7C	17: 9,674,751–9,694,614	0.0094 (7.13)	0.0041 (7.13)	2	7
GOLGA3	12: 133,345,495–133,405,444	0.0057 (13.15)	0.0057 (13.15)	2	10
CDKN2B	9: 22,002,902–22,009,362	0.004 (8.37)	0.0057 (8.37)	1	1
CCDC189	16: 30,768,744–30,774,031	0.0032 (9.82)	0.006 (9.82)	1	1
INTS5	11: 62,414,320–62,420,774	0.031 (13.35)	0.0068 (13.35)	1	1
SUPT20H	13: 37,583,449–37,633,850	0.004 (15.15)	0.0068 (15.15)	1	3
EPB41L4A	5: 111,478,138–111,755,013	0.024 (18.19)	0.0068 (18.19)	2	8
KLHL1	13: 70,274,726–70,682,591	0.002 (12.31)	0.0071 (12.31)	2	5
SPATA13	13: 24,553,944–24,881,212	0.021 (13)	0.0087 (13)	2	12
PTK2B	8: 27,168,999–27,316,903	0.014 (12.72)	0.0097 (12.72)	2	12
NOLC1	10: 103,911,933–103,923,627	0.011 (12.89)	0.01 (12.89)	1	5
SMYD5	2: 73,441,350–73,454,365	0.016 (8.36)	0.011 (8.36)	2	3
GPR35	2: 241,544,848–241,570,676	0.011 (13.72)	0.012 (13.72)	1	12
PCCA	13: 100,741,269–101,182,686	0.036 (7.7)	0.013 (7.7)	1	3
PTGR1	9: 114,312,002–114,362,135	0.011 (7.89)	0.014 (7.89)	1	2
SBNO1	12: 123,773,656–123,849,390	0.0097 (11.01)	0.015 (11.01)	2	6
ABCC1	16: 16,043,434–16,236,931	0.014 (15.17)	0.015 (15.17)	2	8
EIF2AK2	2: 37,326,353–37,384,208	0.014 (9.01)	0.015 (9.01)	1	2
TRAK2	2: 202,241,930–202,316,302	0.014 (6.45)	0.016 (6.45)	1	6
C10orf53	10: 50,887,697–50,918,307	0.009 (13.76)	0.017 (13.76)	1	3
ADRA2B	2: 96,778,707–96,781,984	0.019 (7.03)	0.019 (7.03)	1	3
SLC9A6	X: 135067958–135129423	0.015 (6.57)	0.02 (6.57)	1	1
RSG1	1: 16,558,195–16,563,657	0.0074 (9.85)	0.023 (9.85)	1	2
ABHD5	3: 43,731,605–43,775,863	0.0094 (12.41)	0.024 (12.41)	1	3
PHOX2A	11: 71,950,121–71,956,708	0.017 (12.36)	0.026 (12.36)	1	1
CCDC24	1: 44,457,031–44,462,200	0.012 (6.41)	0.026 (6.41)	1	3
TIMM44	19: 7,991,603–8,008,805	0.015 (11.37)	0.027 (11.37)	2	8
STC2	5: 172,741,716–172,756,506	0.013 (8.74)	0.027 (8.74)	1	2
MNS1	15: 56,713,742–56,757,335	0.019 (12.17)	0.028 (12.17)	2	5
KCNA3	1: 111,214,310–111,217,655	0.036 (8.1)	0.028 (8.1)	1	1
FOXK2	17: 80,477,589–80,602,538	0.021 (11.71)	0.03 (11.71)	2	6
STRN3	14: 31,363,005–31,495,607	0.034 (5.39)	0.033 (5.39)	1	4
CIDEA	18: 12,254,318–12,277,594	0.031 (9.33)	0.034 (9.33)	1	3
VCL	10: 75,757,872–75,879,918	0.031 (9.33)	0.035 (9.33)	1	3
TUBGCP5	15: 22,833,395–22,873,892	0.016 (7.21)	0.036 (7.21)	2	3
SLC25A16	10: 70,237,756–70,287,231	0.024 (12.35)	0.037 (12.35)	1	1
GALNT10	5: 153,570,290–153,800,544	0.041 (3.93)	0.046 (3.93)	3	3
SECISBP2	9: 91,933,421–91,974,557	0.02 (7.32)	0.047 (7.32)	1	3
USE1	19: 17,326,155–17,330,638	0.026 (12.36)	0.049 (12.36)	1	2

The presented genes are ordered according to the association p-value of pVAAST test with sex co-variable control.

[a]: position on the human reference genome hg19

[b]: p-values computed from 10^6 permutations (score)

https://doi.org/10.1371/journal.pone.0213387.t002

Table 3. Association analysis of candidate SNVs in 43 candidate genes.

Gene		*Candidate variant (CV)*		*pVAAST* [d]	
Name	*Group*[a]	*Chromosome: position*[b]	HGVS annotation[c]	*CV only*	*Without CV*
TIMM44	1	19: 7997576	c.923C>A	0.0017 (11.37)	0.26 (2.58)
SLC9A6	2	X: 135122270	c.1763C>T	0.002 (6.57)	1 (0)
TMOD2	2	15: 52060595	c.263C>T	0.0023 (7.07)	1 (0)
INTS5	2	11: 62415889	c.1663C>T	0.0024 (13.35)	1 (0)
PTK2B	1	8: 27255130	c.29G>A	0.0028 (6.45)	0.17 (5.15)
EPB41L4A	1	5: 111504714	c.1828C>T	0.003 (13.76)	0.41 (4.03)
GPR35	2	2: 241569447	c.171C>G	0.0031 (13.72)	1 (0)
NEK1	2	4: 170506525	c.782G>A	0.0035 (8.6)	1 (0)
PCCA	2	13: 100921021	c.898G>A	0.0037 (7.7)	1 (0)
SUPT20H	2	13:37622040	c.73A>T	0.0038 (15.15)	1 (0)
MNS1	1	15: 56736015	c.724C>T	0.0038 (15.15)	0.036 (2.64)
DHRS7C	1	17: 9676206	c.608T>C	0.0041 (6.98)	0.41 (0.84)
NOLC1	2	10: 103917202	c.331C>T	0.0043 (12.89)	1 (0)
RSG1	2	1: 16558713	c.607C>T	0.0044 (10.72)	0.93 (0.01)
ABHD5	2	3: 43753284	c.590G>C	0.0046 (12.41)	1 (0)
AURKB	1	17: 8109861	c.637G>C	0.0051 (10.39)	0.14 (3.54)
CCDC189	2	16:30768844	c.949C>T	0.0051 (9.82)	1 (0)
CDKN2B	2	9: 22006147	c.256G>A	0.0052 (8.37)	1 (0)
KLHL1	1	13: 70681732	c.100G>T	0.0068 (9.53)	0.11 (2.77)
PTGR1	2	9: 114345759	c.488A>T	0.0074 (7.89)	1 (0)
GOLGA3	1	12: 133363022	c.3026G>A	0.0077 (8.59)	0.24 (3.17)
TRAK2	2	2: 202252596	c.1526G>A	0.014 (6.45)	1 (0)
ABCC1	1	16: 16208739	c.3196C>T	0.016 (9.33)	0.28 (3.86)
C10orf53	2	10: 50901917	c.195C>A	0.017 (13.76)	1 (0)
STC2	2	5: 172755066	c.131G>T	0.018 (8.74)	1 (0)
CIDEA	2	18: 12262897	c.214C>T	0.018 (9.33)	1 (0)
SBNO1	1	12: 123806235	c.2170G>T	0.019 (10.21)	0.9 (0.17)
TUBGCP5	1	15: 22864369	c.2327G>A	0.019 (7.21)	1 (0)
EIF2AK2	2	2: 37347149	c.1201G>A	0.02 (9.01)	1 (0)
SPATA13	1	13: 24868977	c.3367C>T	0.02 (6.79)	0.076 (6.46)
ADRA2B	2	2: 96780655	c.1234G>A	0.021 (7.03)	1 (0)
KCNA3	2	1: 111216792	c.640C>T	0.022 (8.1)	1 (0)
USE1	2	19: 17330089	c.490C>T	0.023 (12.36)	1 (0)
VCL	2	10: 75855428	c.1558C>T	0.023 (9.33)	1 (0)
CCDC24	2	1: 44461756	c.848G>A	0.025 (6.41)	1 (0)
GALNT10	1	5: 153765901	c.967G>A	0.026 (3.93)	1 (0)
SLC25A16	2	10: 70246944	c.799C>T	0.031 (12.35)	1 (0)
PHOX2A	2	11: 71950885	c.763C>T	0.032 (12.36)	1 (0)
SECISBP2	2	9: 91954819	c.1253G>T	0.033 (7.32)	1 (0)
STRN3	2	14: 31371744	c.2135C>T	0.034 (5.39)	1 (0)
FOXK2	1	17: 80543821	c.1321C>T	0.037 (7.93)	0.17 (4.17)
SMYD5	1	2: 73450610	c.833G>A	0.067 (1.69)	0.025 (4.86)

(*Continued*)

Table 3. (Continued)

Gene		*Candidate variant (CV)*		*pVAAST* [d]	
Name	*Group*[a]	*Chromosome: position*[b]	HGVS annotation[c]	*CV only*	*Without CV*
SUSD5	1	3: 33216450	c.526C>T	0.21 (5.55)	0.0037 (14.67)

The genes are sorted according to the association p-value of the candidate variant. For each gene, the candidate variant is the one with the highest score.

[a]: The two groups correspond to:

1 –genes containing multiple rare variants contributing to pVAAST score

2 –genes containing one rare variant leading the association

[b]: position on the human reference genome hg19

[c]: annotation on the canonical transcript

[d]: p-values computed from 10^6 permutations (score)

https://doi.org/10.1371/journal.pone.0213387.t003

in affected members of the family and, its absence in unaffected members (Fig 4). Second, we investigated 188 RA patients (characteristics described in supplementary S1 Table) and 362 healthy parents from trio families using one affected member of family 3 as a positive control. No one of these samples showed the rare allele of this SNV.

Discussion

In this study, we identified a novel SNV associated with RA, *SUPT20H*: *c.73A>T(p.Lys25*)*, by performing WES in 9 multiplex RA pedigrees associated with *HLA-DRB1* SE risk alleles. This nonsense variant had a complete penetrance in a family with rather young age at RA onset ($mean_{onset\text{-}age}$ = 31 years old). In addition, we showed with the RVAT pVAAST, that this variant was significantly associated with RA ($pVAAST_{p\text{-}value} = 3.78^{*}10^{-3}$).

The *SUPT20H* gene, which has not been reported yet as a RA risk gene, is a member of the SAGA complex (Spt-Ada-Gcn5 acetyltransferase) gene family. It encodes for the p38 interacting protein (p38IP) which is predicted to contain a Nuclear Localization Signal (NLS), a PEST domain (Proline Glutamic acid Serine Threonine) and 2 serine rich domains [16]. The SNV identified in this study occurring in the NLS domain, the truncated protein that could result from this variation would not have any functional domain.

Previous studies reported p38IP interactions with 3 different proteins [16–18]. Indeed, the protein p38IP was shown to bind to and stabilize the protein GCN5 [17,19], member of the SAGA complex, and thus stabilize the complex itself [20]. In addition, *in vitro* analysis by

Table 4. Top 10 RA associated SNVs and their re-sequencing results.

Gene	*Chromosome*	*Position (bp)*	*Ref/Alt*	*rs number*[a]	*Family*	*#RA cases*	*#Unaffected*
SUPT20H	13	37622040	T/A	.	3	4/4	0/5
SLC9A6	X	135122270	C/T	.	6	4/4	1/3
TIMM44	19	7997576	G/T	rs139625465	10	2/2	1/3
NEK1	4	170506525	C/T	rs200161705	3	4/4	2/5
TMOD2	15	52060595	C/T	.	4	4/5	3/6
PCCA	13	100921021	G/A	.	3 17	4/4 0/2	2/5 1/6
EPB41L4A	5	111504714	G/A	rs200812889	4	2/5	2/6
GPR35	2	241569447	C/G	rs141249079	*NA*	*NA*	*NA*
INTS5	11	62415889	G/A	.	*NA*	*NA*	*NA*

[a]: rs number from dbsnp138

https://doi.org/10.1371/journal.pone.0213387.t004

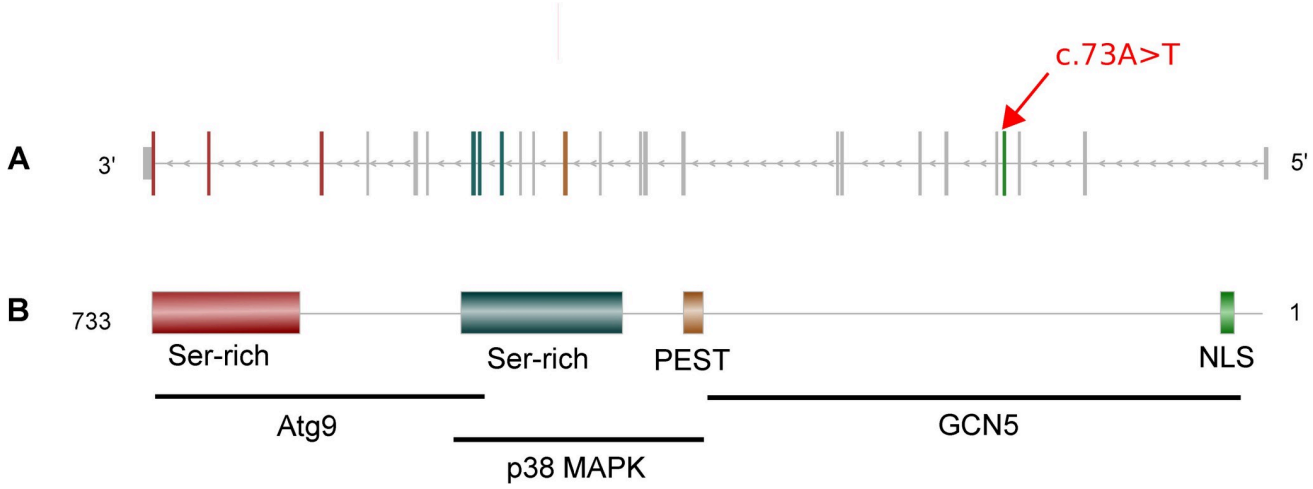

Fig 3. Gene SUPT20H and its product: p38IP. Abbreviations: NLS: Nuclear Localization signal, PEST: Proline (P) Glutamic acid (E) Serine (S) Threonine. (A) Gene model of *SUPT20H*. The red arrow indicates the exon including the nonsense variant identified by WES. The exons color code refers to the protein domain it encodes for. (B) On top, p38IP protein model of 733 amino-acid and on the bottom, the positions of p38IP interactors binding sites.

https://doi.org/10.1371/journal.pone.0213387.g003

Nagy and his colleagues [20] showed that this p38-GNC5 SAGA complex binds to the promotor of Endoplasmic Reticulum stress-induced genes, such as *GRP78*, enhancing their transcription. Interestingly, previous studies have noted the role of the *GRP78* gene in the pathogenesis of RA: the expression of this gene is increased in RA fibroblast-like synoviocytes (FLS) and it promotes their proliferation [21,22]. The protein p38IP, as indicated by its name, interacts also with p38 MAPK, a key protein for the development of RA [23]. And it has previously been reported that this interaction inhibits monocyte/macrophage differentiation [19].

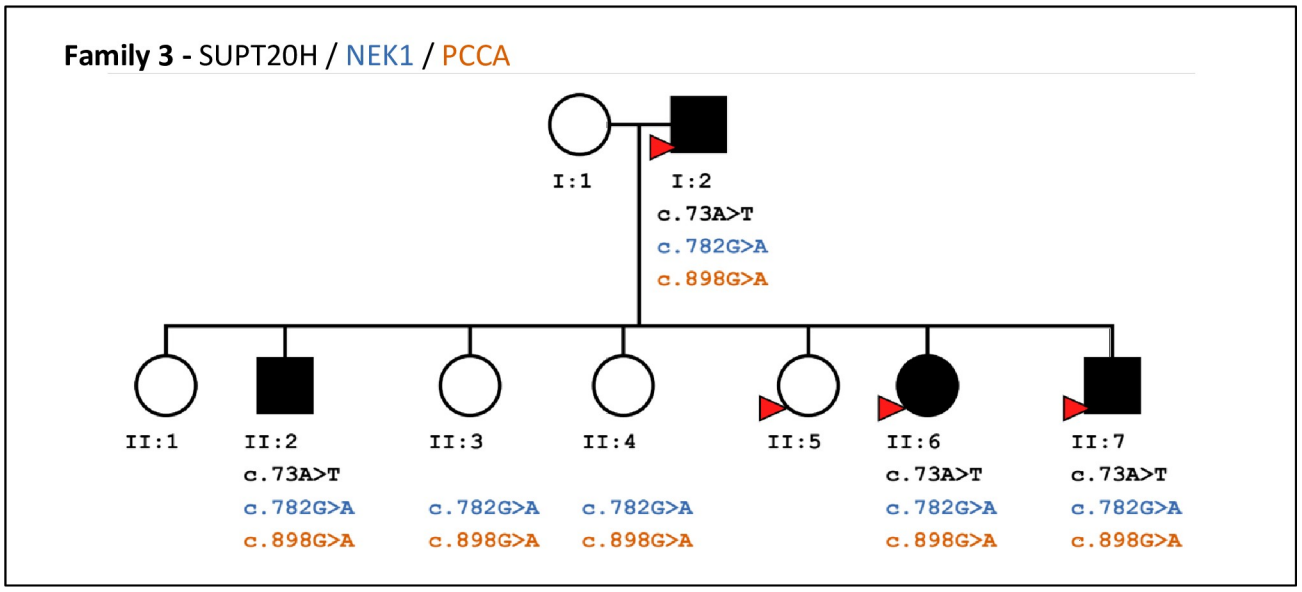

Fig 4. Representation of the pedigree 3 and the variants identified in his members. All the individuals represented here were part of the validation set but only the ones marked with a red arrow were part of the discovery set. Each of them is represented by a circle (if female) or a square (if male). Their filling color corresponds to their status: black if affected by RA or white if unaffected by the disease. The presence of a variant is represented by its HGVS id. The absence of this id indicates a homozygote genotype for the reference allele.

https://doi.org/10.1371/journal.pone.0213387.g004

Macrophages, present in a high concentration in inflamed synovial membrane and cartilage/pannus junction, contributes to the maintenance of the inflammation in RA patients [24,25]. Finally p38IP interacts with the protein Atg9 and is required for its translocation from trans-Golgi system to endosome during starvation induced macro-autophagy [16,26,27]. The protein p38 MAPK, which has higher affinity for p38IP than Atg9 inhibits this interaction and thus inhibits the autophagy process. This implication of p38IP in autophagy process is another argument that highlights the interest of studying the role of p38IP in the development of RA. Indeed, different studies reviewed by Dai and Hu demonstrated the role of autophagy in RA [28].

In addition to *SUPT20H* gene, we identified 42 genes significantly associated with RA (p-value < 0.05). We evaluated their association with the test implemented in pVAAST. This software combines a rare variant association test (RVAT) with a linkage score, thus it takes into account for the familial structure of our data. It allowed us to provide all the variants called in our samples without any *a priori* filtering based on biological knowledge, avoiding loss of information. Indeed, the RVAT incorporates a score evaluating for each variant the amino acid substitution impact and phylogenetic conservation information, allowing it to be robust in the presence of neutral and common variants [12]. Indeed, depending on the chosen test, variants with opposite effects can lead to a decreasing power for the association test [29].

Among these 42 genes, we observed a variable number of variants contributing to the association signal: 16 genes with more than one contributing variant and 26 with only one. In the first group of genes, the potentially pathogenic variants covered the rare to low frequency spectrum. Three of these genes (named *SMYD5*, *SUSD5* and *MNS1*), remained significantly associated with the disease without the candidate variants ($pVAAST_{p\text{-value}} < 0.036$). The fact that the signal for the other genes was not significant anymore, when tested without the candidate SNV, could be explained by the lower number of contributing variants identified in these genes that reduce the RVAT power compared to genes with more causal variants [30] (Table 2). In the second group, 8 genes (*CCDC189*, *CDKN2B*, *SLC9A6*, *PHOX2A*, *SLC25A16*, *TMOD2*, *INTS5* and *KCNA3*) contained only one variant. So, as expected, the score of the association test with all identified variants (familial data and control data) and with the candidate variant only were identical (see Table 1).

We can highlight two limitations in this association analysis. First, we added as controls the variants of 98 individuals extracted from the 1000 genomes project database. And, although we chose individuals from the same population as our pedigrees (European) and sequenced on the same platform (Illumina), two aspects could have introduced bias: the exome capturing kit and the variant calling workflow. The first aspect can introduce variants specific to one of the dataset because not targeted by the kit used for the other one. To partially overcome this issue, we filtered out variants from the 1000 genomes dataset present outside the boundaries of our targeted exomes. But, due to the wide range of capturing kits used to produce the data presents in the 1000 genomes database, we did not filter the variants according to the regions targeted by the 1000 genomes project. For the second aspect, it has previously been shown that despite a few variant specific to a variant calling software, the concordance between the called variants is high (92% between HaplotypeCaller [31], Samtools [32] and FreeBayes [33]) [34]. In addition, the PCA, performed on all the individuals included in the association test (Fig 2), shows that the observed genetic variability is mostly related to the intra-family specificity, not to the choice of the 98 CEU controls, comforting our choice of controls. Second, individuals of the discovery set have different clinical characteristics (such as sex, age at diagnostic, and ACPA status). These differences, if not controlled, could have introduced bias in the association analysis. Since, the sex was the sole information known for both discovery set and CEU controls, we were only able to control for this co-variable.

Finally, we previously suggested that rare variations present in the genome of RA cases in addition to *HLA-DRB1* SE could modulate the effect of this latter, leading to RA cases with different clinical characteristics. To test this hypothesis, we could investigate the interactions between *HLA-DRB1* and new RA risk genes, such as *SUPT20H*. For this analysis, we would need to perform a study including both individuals carrying and not carrying *HLA-DRB1* risk alleles.

Further work is also needed to identify new RA risk genes that may have been missed in this study because we used stringent criteria to select the candidate SNVs. Indeed, we removed variants: with incomplete penetrance, segregating in several families and not shared by all affected relatives of a given family.

In conclusion, we identified a new rare nonsense SNV, *SUPT20H*: *c.73A>T (p.Lys25*)*, associated with RA, by combining linkage and association analysis with pVAAST. Neither the gene nor the variant was previously associated with the disease. But the review of the literature about *SUPT20H* gene and its product, p38IP, supports the idea that this gene is involved in the pathogenesis of RA. Further work, with *in vitro* functional studies, needs to be done to evaluate the pathogenic effect of this new SNV and to validate its role in the different processes previously described in the context of RA. In addition, further studies need to be done to validate the other RA risk loci identified in this study, in particular *SMYD5*, *MNS1* and *SUSD5* genes in which we observed an aggregation of rare variants.

Methods

Participants

We studied 16 French RA multiplex families with at least 4 affected individuals per family carrying one or two *HLA-DRB1* shared epitope (SE) allele. Half of the families had only RA cases, whereas the other half had RA and/or other Autoimmune Diseases (AID: Lupus Erythematosus, Vitiligo, Sjögren syndrome, Hashimoto's thyroiditis) cases. Among the 110 recruited individuals within these families, 33 (30%) were reported as only affected by RA, 17 (15%) by RA and another AID, 15 (14%) by an AID different from RA and 45 (41%) were reported as unaffected. As discovery set, we selected 30 individuals belonging to 9 of the 16 previously described pedigrees, with a sufficient quantity of DNA for WES. This sample set consisted in 19 individuals with RA and 11 relatives not affected by RA (at least one per pedigree). All the remaining samples were included in the validation set.

Our study was approved by ethics committees of Hôpital Bicêtre and Hôpital Saint Louis (Paris, France; CPPRB 94–40). Everyone provided a written informed consent for the participation in the study.

For statistical analysis with the Rare Variant Association Test (RVAT), we added 98 healthy controls extracted from the 1000 genomes project database. We chose CEU individuals (Utah residents with Northern and Western European ancestry) for which whole exome sequences were generated using an Illumina platform.

Whole Exome Sequencing and variant calling

The exons were captured in the discovery set with Agilent SureSelect Human All Exon kit (V5) which target the exome of more than 20,000 genes. Then, we sequenced those regions on an Illumina HiSeq2000 platform and mapped the reads to the human reference genome hg19 [35] using BWA-MEM algorithm [36]. Finally, we marked and removed the duplicates with Picard toolkit [37]. The observed variance being low, we did not apply the recalibration to the WES data.

We called the variants using Haplotype Caller (HC) algorithm from the GATK suite [31,38] in the targeted regions plus 150 bp up and downstream. Then, we filtered out SNVs and small indels (maximum length of 50 bp) according to the following criteria: total read depth DP < 12, mapping quality MQ < 30, variant confidence QD < 2 and strand bias FS score > 25. We annotated genotypes with an individual DP < 10 as missing before filtering out variants with call-rate $< 95\%$ with Plink1.9 [39]. Finally, we removed variants located in segmental duplications [40] and repeated regions, such as described in RepeatMasker from UCSC [41].

Variant annotation and classification

We classified the remaining variants into rare variants (MAF $< 1\%$), low frequency variants ($1\% \leq$ MAF $< 5\%$) and common variants (MAF $\geq 5\%$) according to their frequency in public databases. To evaluate these frequencies, we worked on 4 datasets extracted from: (1) the 1000 Genomes Project (2015 August); (2) the Exome Aggregation Consortium project (ExAC); (3) the Exome Sequencing project (ESP6500–6500 exomes); (4) the Complete Genomics project (CG69–69 individuals). We extracted the allele frequency in population with European ancestry when the information was available.

To investigate the predicted effect of the genetic variants on proteins, we annotated them with CADD phred-like score [42] using ANNOVAR [43] and with variant effect defined in SNPeff [44]. The former is a score obtained from a model trained to separate evolutionary conserved variants, likely deleterious, from simulated variants, likely benign. And the latter evaluate the putative variant impact at the transcript level by using sequence ontology.

Principal Component Analysis of CEU controls and discovery set

To identify possible genetic population stratification between the discovery set and the 98 CEU controls, we performed a principal component analysis (PCA). For this analysis, we used a subset of variants located on autosomes classified as neutrals and frequents. We defined as neutral a variant annotated "LOW" by SNPEff and with a CADD phred-like score < 5. Those variants were considered frequent if their MAF was superior or equal to 5%. In addition, we processed the selected variants prior to performing the PCA analysis to remove those not in Hardy-Weinberg equilibrium (HW) and those in Linkage Disequilibrium (LD) with each other. We used Plink1.9 [39,45] to remove variants with $HW_{p\text{-value}} < 0.01$ and/or with LD ≥ 0.2. The Hardy-Weinberg equilibrium was evaluated on unaffected individuals only. We finally applied the PCA on these variations with the R package SNPRELATE [46].

Selection of RA risk candidate variants

We first selected the rare variants (MAF $\leq 1\%$) observed in only one family with an in-house python script. Further, we chose variants segregating with RA within families and being absent from healthy relatives.

Association analysis of selected genes with burden test

We tested the association of genes carrying at least one rare candidate variant using pVAAST [12]. This software extends the rare variant association test (RVAT) VAAST [47] to offer more power in the context of family-based studies. It provides a linkage-association score and a p-value for each evaluated gene and, a score for each variant carried by the gene to help the prioritization of these variants.

Considering the heterozygous nature of our candidate SNVs, we tested the candidate genes under a dominant model and set the maximum disease prevalence to 0.01, the world-wide prevalence of RA [48]. In addition, we provided to pVAAST a file containing for each variant included in the analysis the maximum frequency observed in the four public databases described above. To evaluate the significance of the test, we authorized genotyping error and did not restrain penetrance value for gene-drop simulations. We estimated p-values by allowing pVAAST to perform up-to 10^6 permutations.

We ran a first time pVAAST for each of our candidate genes by including all the variants, rare and frequent, identified in our discovery set and/or in the 98 controls from 1000 genomes project). The genes significantly associated with RA ($p\text{-value}_{run1} < 0.05$) were selected for a second run of pVAAST using the same parameters plus a control of the sex co-variable. Then, for each gene with $p\text{-value}_{run2} < 0.05$, we selected the variant with the highest score. We ran pVAAST on this variant alone (run 3) and then on the gene without this variant (run 4).

Validation in extended pedigrees of exome selected variants

The top 10 candidate variations, selected with the lowest p-values in the 3rd run of the pVAAST analysis, were re-sequenced in the validation set by using PGM System (Ion Personal Genome Machine). Primers were designed with the Ion AmpliSeq Designer to target the selected variants. We mapped the reads to the reference genome with BWA [36] and recalibrated base score with BQSR tool from GATK suite [31]. We then called variants in the strict limits of the targeted regions with HC and applied the same quality filters used for WES to the observed variants. For samples in both discovery and validation sets, we checked the concordance with whole exome data using VCFtools program package [49].

Genotyping of SUPT20H: *c.73A>T* in multiplex families and additional trios

The candidate rare variant was also genotyped using a custom assay with a FAM or VIC reporter Dye at the 5'end of each TaqMan MGB probe and a non-fluorescent quencher at the 3'end of each probe (Applied Biosystems, Foster City, CA, USA). Digital PCR (QX200 Droplet Digital PCR, Bio-Rad Laboratories, California, USA) was used to detect variant alleles. First, we analyzed members of the family in which the variant was characterized. Second, an independent sample of 188 trio families (one RA patient and his/her two parents from French European origin) was investigated to search for the variant, along with a positive and a negative control.

Supporting information

S1 Appendix. Pedigrees of the 16 families included in this study.
(XLSX)

S2 Appendix. Candidate gene variant calling file. VCF file containing all the high-quality variants identified in the 77 candidate genes.
(VCF)

S3 Appendix. Unfiltered pVAAST results.
(XLSX)

S1 Table. Characteristics of 188 RA index cases in which the SNV c.73A>T carried by SUPT20H was investigated. [a] Number of index cases / Number of index cases with data [b] previous and/or actual tobacco exposure (smokers and ex-smokers).

RF: Rheumatoid Factor.
ACPA: Anti-Cyclic Citrullinated Peptide Antibodies.
(DOCX)

Acknowledgments

We are grateful to RA patients and their families for participation in this study.

Author Contributions

Conceptualization: Maëva Veyssiere, Javier Perea, Elisabeth Petit-Teixeira, Valérie Chaudru.

Data curation: Anne Boland, Robert Olaso, Jean-François Deleuze, Valérie Chaudru.

Formal analysis: Maëva Veyssiere.

Funding acquisition: Jean-François Deleuze, Elisabeth Petit-Teixeira, Valérie Chaudru.

Investigation: Maëva Veyssiere, Laetitia Michou, Christophe Caloustian, François Cornelis, Elisabeth Petit-Teixeira.

Methodology: Maëva Veyssiere, Valérie Chaudru.

Project administration: Javier Perea, Elisabeth Petit-Teixeira, Valérie Chaudru.

Resources: Laetitia Michou, Anne Boland, Robert Olaso, Jean-François Deleuze, François Cornelis.

Software: Maëva Veyssiere.

Supervision: Javier Perea, Elisabeth Petit-Teixeira, Valérie Chaudru.

Validation: Maëva Veyssiere, Christophe Caloustian, Robert Olaso, Elisabeth Petit-Teixeira, Valérie Chaudru.

Visualization: Maëva Veyssiere.

Writing – original draft: Maëva Veyssiere.

Writing – review & editing: Maëva Veyssiere, Javier Perea, Laetitia Michou, Anne Boland, Christophe Caloustian, Robert Olaso, Jean-François Deleuze, François Cornelis, Elisabeth Petit-Teixeira, Valérie Chaudru.

References

1. Stastny P. Mixed lymphocyte cultures in rheumatoid arthritis. J Clin Invest. 1976; 57: 1148–1157. https://doi.org/10.1172/JCI108382 PMID: 1262462
2. Perricone C, Ceccarelli F, Valesini G. An overview on the genetic of rheumatoid arthritis: A never-ending story. Autoimmun Rev. 2011; 10: 599–608. https://doi.org/10.1016/j.autrev.2011.04.021 PMID: 21545847
3. Kurkó J, Besenyei T, Laki J, Glant TT, Mikecz K, Szekanecz Z. Genetics of rheumatoid arthritis—A comprehensive review. Clin Rev Allergy Immunol. 2013; 45: 170–179. https://doi.org/10.1007/s12016-012-8346-7 PMID: 23288628
4. Van der Woude Diane, Houwing-Duistermaat Jeanine J., Toes René E. M., Huizinga Tom W. J., Thomson Wendy, Worthington Jane, et al. Quantitative heritability of anti–citrullinated protein antibody–positive and anti–citrullinated protein antibody–negative rheumatoid arthritis. Arthritis Rheum. 2009; 60: 916–923. https://doi.org/10.1002/art.24385 PMID: 19333951
5. The contribution of H LA to rheumatoid arthritis—Deighton—1989—Clinical Genetics—Wiley Online Library [Internet]. [cited 19 Apr 2018]. Available: https://onlinelibrary-wiley-com.ezproxy.universite-paris-saclay.fr/doi/abs/10.1111/j.1399-0004.1989.tb03185.x

6. Vries RRP de, Woude D van der, Houwing JJ, Toes. Genetics of ACPA-positive rheumatoid arthritis: the beginning of the end? Ann Rheum Dis. 2011; 70: i51–i54. https://doi.org/10.1136/ard.2010.138040 PMID: 21339219
7. Diogo D, Kurreeman F, Stahl E a., Liao KP, Gupta N, Greenberg JD, et al. Rare, low-frequency, and common variants in the protein-coding sequence of biological candidate genes from GWASs contribute to risk of rheumatoid arthritis. Am J Hum Genet. 2013; 92: 15–27. https://doi.org/10.1016/j.ajhg.2012.11.012 PMID: 23261300
8. Bang S-Y, Na Y-J, Kim K, Joo YB, Park Y, Lee J, et al. Targeted exon sequencing fails to identify rare coding variants with large effect in rheumatoid arthritis. Arthritis Res Ther. 2014; 16: 447–447. https://doi.org/10.1186/s13075-014-0447-7 PMID: 25267259
9. Wang D, Li Y, Liu Y, He Y, Shi G. Expression of VSTM1-v2 Is Increased in Peripheral Blood Mononuclear Cells from Patients with Rheumatoid Arthritis and Is Correlated with Disease Activity. PLoS ONE. 2016;11. https://doi.org/10.1371/journal.pone.0146805 PMID: 26760041
10. Mitsunaga S, Hosomichi K, Okudaira Y, Nakaoka H, Suzuki Y, Kuwana M, et al. Aggregation of rare/low-frequency variants of the mitochondria respiratory chain-related proteins in rheumatoid arthritis patients. J Hum Genet. 2015; 60: 449–454. https://doi.org/10.1038/jhg.2015.50 PMID: 26016412
11. Okada Y, Diogo D, Greenberg JD, Mouassess F, Achkar WAL, Fulton RS, et al. Integration of Sequence Data from a Consanguineous Family with Genetic Data from an Outbred Population Identifies PLB1 as a Candidate Rheumatoid Arthritis Risk Gene. PLoS ONE. 2014; 9: e87645–e87645. https://doi.org/10.1371/journal.pone.0087645 PMID: 24520335
12. Hu H, Roach JC, Coon H, Guthery SL, Voelkerding KV, Margraf RL, et al. A unified test of linkage analysis and rare-variant association for analysis of pedigree sequence data. Nat Biotechnol. 2014; 32: 663–669. https://doi.org/10.1038/nbt.2895 PMID: 24837662
13. Oliver JE, Silman AJ. Why are women predisposed to autoimmune rheumatic diseases? Arthritis Res Ther. 2009; 11: 252. https://doi.org/10.1186/ar2825 PMID: 19863777
14. van Vollenhoven RF. Sex differences in rheumatoid arthritis: more than meets the eye . . . BMC Med. 2009; 7: 12. https://doi.org/10.1186/1741-7015-7-12 PMID: 19331649
15. Bainbridge MN, Wang M, Wu Y, Newsham I, Muzny DM, Jefferies JL, et al. Targeted enrichment beyond the consensus coding DNA sequence exome reveals exons with higher variant densities. Genome Biol. 2011; 12: R68. https://doi.org/10.1186/gb-2011-12-7-r68 PMID: 21787409
16. Webber JL, Tooze SA. Coordinated regulation of autophagy by p38α MAPK through mAtg9 and p38IP. EMBO J. 2010; 29: 27–40. https://doi.org/10.1038/emboj.2009.321 PMID: 19893488
17. Liu X, Xiao W, Wang X-D, Li Y-F, Han J, Li Y. The p38-interacting protein (p38IP) regulates G2/M progression by promoting α-tubulin acetylation via inhibiting ubiquitination-induced degradation of the acetyltransferase GCN5. J Biol Chem. 2013; 288: 36648–36661. https://doi.org/10.1074/jbc.M113.486910 PMID: 24220028
18. Zohn IE, Li Y, Skolnik EY, Anderson KV, Han J, Niswander L. p38 and a p38-Interacting Protein Are Critical for Downregulation of E-Cadherin during Mouse Gastrulation. Cell. 2006; 125: 957–969. https://doi.org/10.1016/j.cell.2006.03.048 PMID: 16751104
19. Yu X, Wang Q-L, Li Y-F, Wang X-D, Xu A, Li Y. A novel miR-200b-3p/p38IP pair regulates monocyte/macrophage differentiation. Cell Discov. 2016; 2: 15043. https://doi.org/10.1038/celldisc.2015.43 PMID: 27462440
20. Nagy Z, Riss A, Romier C, le Guezennec X, Dongre AR, Orpinell M, et al. The Human SPT20-Containing SAGA Complex Plays a Direct Role in the Regulation of Endoplasmic Reticulum Stress-Induced Genes. Mol Cell Biol. 2009; 29: 1649–1660. https://doi.org/10.1128/MCB.01076-08 PMID: 19114550
21. Yoo S-A, You S, Yoon H-J, Kim D-H, Kim H-S, Lee K, et al. A novel pathogenic role of the ER chaperone GRP78/BiP in rheumatoid arthritis. J Exp Med. 2012; 209: 871–886. https://doi.org/10.1084/jem.20111783 PMID: 22430489
22. Park JH, Gail MH, Weinberg CR, Carroll RJ, Chung CC, Wang Z. Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. Proc Natl Acad Sci U A. 2011;108. https://doi.org/10.1073/pnas.1114759108 PMID: 22003128
23. Clark AR, Dean JL. The p38 MAPK Pathway in Rheumatoid Arthritis: A Sideways Look. Open Rheumatol J. 2012; 6: 209–219. https://doi.org/10.2174/1874312901206010209 PMID: 23028406
24. Kinne RW, Bräuer R, Stuhlmüller B, Palombo-Kinne E, Burmester G-R. Macrophages in rheumatoid arthritis. Arthritis Res. 2000; 2: 189–202. https://doi.org/10.1186/ar86 PMID: 11094428
25. Kinne RW, Stuhlmüller B, Burmester G-R. Cells of the synovium in rheumatoid arthritis Macrophages. Arthritis Res Ther. 2007; 9: 224. https://doi.org/10.1186/ar2333 PMID: 18177511
26. Webber JL. Regulation of autophagy by p38α MAPK. Autophagy. 2010; 6: 292–293. https://doi.org/10.4161/auto.6.2.11128 PMID: 20087063

27. Webber JL, Tooze SA. New insights into the function of Atg9. FEBS Lett. 2010; 584: 1319–1326. https://doi.org/10.1016/j.febslet.2010.01.020 PMID: 20083107

28. Dai Y, Hu S. Recent insights into the role of autophagy in the pathogenesis of rheumatoid arthritis. Rheumatology. 2016; 55: 403–410. https://doi.org/10.1093/rheumatology/kev337 PMID: 26342228

29. Kosmicki JA, Churchhouse CL, Rivas MA, Neale BM. Discovery of rare variants for complex phenotypes. Hum Genet. 2016; 135: 625–634. https://doi.org/10.1007/s00439-016-1679-1 PMID: 27221085

30. Moutsianas L, Agarwala V, Fuchsberger C, Flannick J, Rivas MA, Gaulton KJ. The power of gene-based rare variant methods to detect disease-associated variation and test hypotheses about complex disease. PLoS Genet. 2015; 11. https://doi.org/10.1371/journal.pgen.1005165 PMID: 25906071

31. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinforma Ed Board Andreas Baxevanis Al. 2013; 11: 11.10.1–11.10.33. https://doi.org/10.1002/0471250953.bi1110s43

32. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011; 27: 2987–2993. https://doi.org/10.1093/bioinformatics/btr509 PMID: 21903627

33. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. ArXiv12073907 Q-Bio. 2012; Available: http://arxiv.org/abs/1207.3907

34. Hwang S, Kim E, Lee I, Marcotte EM. Systematic comparison of variant calling pipelines using gold standard personal exome variants. Sci Rep. 2015; 5: 17875. https://doi.org/10.1038/srep17875 PMID: 26639839

35. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001; 409: 860–921. https://doi.org/10.1038/35057062 PMID: 11237011

36. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv13033997 Q-Bio. 2013; Available: http://arxiv.org/abs/1303.3997

37. Picard Tools—By Broad Institute [Internet]. [cited 13 Nov 2017]. Available: http://broadinstitute.github.io/picard/

38. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010; 20: 1297–303. https://doi.org/10.1101/gr.107524.110 PMID: 20644199

39. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. 2015; 4. https://doi.org/10.1186/s13742-015-0047-8 PMID: 25722852

40. She X, Jiang Z, Clark RA, Liu G, Cheng Z, Tuzun E, et al. Shotgun sequence assembly and recent segmental duplications within the human genome. Nature. 2004; 431: 927. https://doi.org/10.1038/nature03062 PMID: 15496912

41. Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, et al. The UCSC Genome Browser database: 2014 update. Nucleic Acids Res. 2014; 42: D764–D770. https://doi.org/10.1093/nar/gkt1168 PMID: 24270787

42. 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: 310–315. https://doi.org/10.1038/ng.2892 PMID: 24487276

43. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010; 38: e164–e164. https://doi.org/10.1093/nar/gkq603 PMID: 20601685

44. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang 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). /01//; 6: 80–92. https://doi.org/10.4161/fly.19695 PMID: 22728672

45. Purcell S, Chang C. PLINK [Internet]. Available: www.cog-genomics.org/plink/1.9/

46. Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics. 2012; 28: 3326–3328. https://doi.org/10.1093/bioinformatics/bts606 PMID: 23060615

47. Hu H, Huff CD, Moore B, Flygare S, Reese MG, Yandell M. VAAST 2.0: Improved Variant Classification and Disease-Gene Identification Using a Conservation-Controlled Amino Acid Substitution Matrix. Genet Epidemiol. 2013; 37: 622–634. https://doi.org/10.1002/gepi.21743 PMID: 23836555

48. Silman AJ, Pearson JE. Epidemiology and genetics of rheumatoid arthritis. Arthritis Res Ther. 2002; 4: S265. https://doi.org/10.1186/ar578 PMID: 12110146

49. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011; 27: 2156–2158. https://doi.org/10.1093/bioinformatics/btr330 PMID: 21653522