博士論文

A Genome-Wide Association Study for Identifying the Susceptibility Genes to Interstitial Lung Disorder in Rheumatoid Arthritis Patients

（ゲノムワイド関連解析を用いた関節リウマチに併発する間質性肺病変における感受性遺伝子の探索）

豊岡　理人

Table of contents

Abbreviation

ACPA	anti-citrullinated protein antibody
bp	base-pair
chr	chromosome
CI	confidence interval
DI-ILD	drug-induced interstitial lung disorder
DMARDs	disease-modifying antirheumatic drugs
FGFR	fibroblast growth factor receptor
eQTL	expression quantitative trait locus
GWAS	genome-wide association study
HRCT	high-resolution computed tomography
HWE	Hardy-Weinberg equilibrium
HLA	human leukocyte antigen
IBD	identity by descent
IPF	idiopathic pulmonary fibrosis
ILD	interstitial lung disorder
kb	kilo base-pair

LCL	lymphoblastoid cell lines
LD	linkage disequilibrium
MAF	minor allele frequency
Mb	mega base-pair
MTX	Methotrexate
NRG	neuregulin
OR	odds ratio
PCA	principal component analysis
PDGF-β	platelet derived growth factor beta
PF	pulmonary fibrosis
Q-Q plot	quantile-quantile plot
RA	rheumatoid arthritis
RA-ILD	interstitial lung disorder in RA
SE	shared epitope
SNP	single nucleotide polymorphism
SP	surfactant protein
TGF-β	transforming growth factor-beta
TNF-α	Tumor necrosis factor alpha

Abstract

Rheumatoid arthritis (RA) is a chronic disorder characterized by systemic inflammation due to autoimmune dysfunction in synovial joints leading to destruction of articular cartilage and deformity of joints. RA patients develop various signs and symptoms including articular and extra-articular manifestations with progression of the disease. Interstitial lung disorder (ILD) is one of major extra-articular complications of the disease. Its prevalence rate is estimated to be 7.7-58.0% among RA patients. Interstitial lung disorder in RA (RA-ILD) shows poor prognosis, and median survival time upon diagnosis of ILD is only 2.6 years. To date, no data of GWAS has been reported for RA-ILD. One hundred and sixty nine unrelated Japanese RA patients with ILD and 294 Japanese RA patients without ILD were subjected to GWAS. The present study identified eight regions with *p*-values less than 5×10^{-6}: 5q35.2, 6q21.32, 7p14.3, 12q23.3, 12q24.11, 19q13.11, 19q13.12 and 22q13.1. A SNP in the region 19q13.11 yielded the lowest *p*-value of 5.98×10^{-7} by Cochran-Armitage trend test. *HLA-DRB1*04* is well-known as a susceptible allele of RA, nevertheless it was shown to function as protective for RA-ILD in the present study. Thus, sub-group analysis divided by presence/absence of *HLA-DRB1*04* were conducted to identify other genetic factors of this disease. From the result of genome-wide association tests between *HLA-DRB1*04* positive RA with ILD and RA without ILD, a strong association was observed at 7p21.3 with *p*-value of

4.2×10^{-7}. One of the flanking SNPs was reported to have an effect on a protein expression level of transforming growth factor-beta (*TGF-β*) which plays an important role in fibrogenic processes of the lung. Furthermore, from the result of genome-wide association tests between *HLA-DRB1*04* negative RA with and without ILD, a SNP in the region 19q13.11 showed an association with *p*-values of 2.72×10^{-6} under allelic model. The SNP with a lowest *p*-value in the region is positioned in the 2nd intron of a gene which belongs to neureglin gene family.

This study identified novel SNPs which might cause RA-ILD. Sub-group analysis divided by the presence/absence of *HLA-DRB1*04* revealed that those two groups had different genetic predispositions of RA-ILD.

本研究は本学倫理委員会により承認を得ております。(承認番号: G3208-(1))

新規疾患遺伝子の同定の可能性から、本学の知的財産部への届出を予定している為、学位論文での詳細な記載を免除願います。

1. Introduction

Rheumatoid arthritis (RA) is a chronic disorder characterized by systemic inflammation caused by autoimmune dysfunction in the synovial joints, leading to the destruction of articular cartilage and joint deformity. The prevalence of the disease currently stands at over 0.5% of the population in Japan. Patients afflicted with RA develop various signs and symptoms including articular and extra-articular manifestations accompanying disease progression. Extra-articular manifestations include pericarditis, pleuritis, rheumatoid nodules, Felty's syndrome, vasculitis, and interstitial lung disorder (ILD). ILD is one of the most common clinical manifestations of the lung among RA patients.

The initial report of RA patients with a pulmonary manifestation, based on observation by radiography, was published in 1948 [1]. A subsequent study suggested that ILD is a complication of RA, however, owing to the limitations imposed by a small sample size, the authors ultimately decided that the results were inconclusive. [2]. Following this, in 1965, the incidence of pulmonary lesions in RA patients was observed by chest radiography [3]. A study conducted in 1972 showed that 33% of the RA patients presented abnormalities that are consistent with interstitial fibrosis, as examined by chest radiography and diffusion capacity measurement, and the authors confined the association between RA and ILD [4]. Currently ILD in RA (RA related ILD; RA-ILD) is considered as one of the extra-articular

manifestations that can affect the prognosis and therapeutic approach [5, 6]. Additionally, it was also recognized that ILD resulted from complications due to the several drugs used in RA treatment, including disease-modifying antirheumatic drugs (DMARDs) and biological agents (drug-induced ILD; DI-ILD) [7, 8]. Thus, the pathogenesis of RA-ILD proved to be complex.

Although RA mainly occurs in female individuals, RA-ILD is frequently develops in male individuals, with a 2 : 1 male to female ratio among RA patients [9]. The prevalence of RA-ILD ranged widely from 7.7% to 58.0%, owing to the fact that previous reports employed different methods of clinical examination and inconsistent diagnostic criteria [10-12]. By using high-resolution computed tomography (HRCT) scans, 33% of RA patients in Australia demonstrated evidence of abnormalities consistent with ILD [12]. Another study conducted in Saudi Arabia demonstrated that 27.5% of RA patients presented abnormal HRCT findings in the lung [13]. A population-based cohort study revealed that the lifetime risk of developing ILD was 7.7% for RA patients and 0.9% for non-RA subjects, yielding a hazard ratio of 8.96 [10]. Annualized incidence of ILD in RA patients was 4.1/1000 and the 15-years cumulative incidence was 62.9/1000 in RA patients from an inception cohort with a 20-year follow-up period [6]. In Japan, the age-adjusted incidences of interstitial pneumonia among total, male and female patients were 1.06, 1.45 and 0.68 per 1,000 RA patients, respectively [14].

RA-ILD demonstrates a poor prognosis [9], accounting for 7.0% of all

RA-associated deaths and contributing to 13.0% of excess mortality among RA patients [10]. ILD is one of the main causes of death in RA patients both in Japan [15] and in USA. [10]. An American study also indicated that median survival time of RA-ILD was as brief as 2.6 years after its diagnosis [10].

Both genetic and environmental factors have been reported as risk factors of RA-ILD. Advanced age [10], gender (i.e., male) [15] and increased severity of joint involvements [10] have been shown to be risk factors for RA-ILD. Smoking was also indicated as the risk factor for both RA and RA-ILD. The correlation between RA and smoking was especially evident among heavy smokers (≥10 pack-years) and those with ongoing smoking history [16, 17]. Moreover, smoking was also associated with RA-ILD [18-20]. However, another study in Japanese population failed to corroborate the association between smoking history and RA-ILD by the logistic regression analysis [21]. An association between anti-citrullinated protein antibodies (ACPA) and RA-ILD has been reported, but is controversial. Four studies in Japanese [21], French [22], Greek [23] and Chinese [24] populations have indicated that an association exists between ACPA and RA-ILD, although other studies in Japan [25] and Korea [26] did not demonstrate this association.

Human leukocyte antigen (HLA) is known to be associated with RA. *HLA-DRB1*0101*, *HLA-DRB1*0401* and *HLA-DRB1*0404* are well known risk alleles

among RA in the Caucasian population [27]. The *HLA-DRB1*0405* allele helps predict a predominant risk of RA in Japanese population [28]. Shared epitope (SE), which is a conserved amino acid sequence in HLA-DRβ chain, has been reported as a risk factor for RA [27, 29]. Several studies have reported the association between extra-articular manifestations of RA with SE and *HLA-DR4*, a serological determinant encoded by some *HLA-DRB1* alleles [30, 31]. *HLA-DRB1*04:01*, *DRB1*04:05* [30-32] and *HLA-DRB1*15:02* [33] have been known as risk alleles of extra-articular manifestations of RA among European, East Asian and Japanese, respectively. In a previous report by my collaborators, *HLA-DRB1*04*, SE and *HLA-DQB1*04* were significantly associated with a decreased risk of RA-ILD [34]. In contrast, *HLA-DRB1*16*, *DR2* serological group and *HLA-DQB1*06* showed an increased risk of RA-ILD [34]. It is still unclear whether RA-ILD patients possess a different genetic background as compared with all RA patients.

Recent genome-wide linkage analysis [35] and various research studies, including genome-wide association studies (GWAS), revealed that *MUC5B* (rs35705950) and *TERT* (rs2736100) [36-39] single nucleotide polymorphisms (SNPs) were correlated with idiopathic pulmonary fibrosis (IPF), which shared common histological patterns of RA-ILD such as usual interstitial pneumonia. Surfactant protein (SP) is associated with familial ILD [40]. Tumor necrosis factor alpha (TNF-α) is known as a key factor to promote proliferation of

fibroblasts and to degrade extracellular matrix [41]. It triggers the expression of growth factors such as TGF-β, platelet derived growth factor beta (PDGF-β), chemokine and cytokines such as interleukin-1 alpha and interleukin-4 in IPF [42-44]. TGF-β plays a role in the induction of differentiation of fibroblasts into myofibroblasts, which are important constituents of the extracellular matrix in the lung fibrogenic processes [45]. PDGF-β is produced by macrophages, fibroblasts, epithelial and endothelial cells in the lung [42, 46]. Inhibition of PDGF tyrosine kinase receptor significantly attenuated the development of ILD in an experimental mouse model of pulmonary fibrosis (PF) [47]. Nevertheless, genetic background of RA-ILD in comparison with PF patients still requires further elucidation.

The ILD-promoting effect exerted by some therapeutics is an important issue in the consideration of RA treatment strategy. Methotrexate (MTX) is an anchor drug for the treatment of RA, though it bears the possible side effect of causing a pulmonary abnormality. Reported incidence of "methotrexate pneumonitis" in RA varied from 0.9% to 6.9% [48]; however,the morbidity and mortality reached up to 20.0%. *HLA-A*31:01* has been reported as a possible genetic factor of MTX induced ILD in the Japanese population [49]. Another therapeutic drug, leflunomide, has been suggested to increase the risk of RA-ILD in the setting of preexisting lung disease, with a potential impact on survival [50, 51].

In the past decade, GWAS has emerged as an effective genetic screening tool. GWAS

utilizes a large number of SNPs across the human genome as markers to reveal associations between specific polymorphisms and corresponding phenotypes. This method offers the advantage of statistical power compared to linkage studies. Easiness of sample collection is also a merit of GWAS, as compared with the affected sib-pair method and the transmission disequilibrium test that require a large number of familial samples. In fact, GWASs have contributed to the discovery of more than 12,000 susceptibility SNPs for over 1,000 diseases or traits [52]. GWASs have been applied to RA as well, and several novel genetic factors have been uncovered, such as *CCR6, NFKBIE, STAT4* and *TNFIP3A*, and in addition the latest meta-analysis reported that candidate SNPs of RA numbered up to 101 [53]. However, a comprehensive study using GWAS has yet to be conducted for RA-ILD. The present study performed GWAS to identify the susceptibility genes or SNPs that are associated with RA-ILD among Japanese.

2. Materials and Methods

2.1 GWAS of RA-ILD

2.1.1 Patients

This study was a collaborative effort between our laboratory and the Sagamihara National Hospital in Kanagawa, Japan, aimed at identifying common genetic variants associated with susceptibility to RA-ILD. 620 RA patients were recruited from 10 hospitals: Sagamihara National Hospital (Kanagawa), Shimoshizu National Hospital (Chiba), Hokkaido Medical Center (Hokkaido), Himeji Medical Center (Hyogo), Kurashiki Medical Center (Okayama), Morioka Hospital (Iwate), Ureshino Medical Center (Saga), Kyushu Medical Center (Fukuoka), Miyakonojo National Hospital (Miyazaki) and Beppu Medical Center (Ooita). All patients were considered to be native Japanese living in Japan. Institutional review boards at each hospital approved the study, and all participants gave their informed consent. A number of patients in this study had been described in a previous report by collaborators [34]. The current research was approved by Human Genome, Gene Analysis Research Ethics Committee of Graduate School of Medicine and Faculty of Medicine, The University of Tokyo. The approval code is G3208-(1).

. All patients fulfilled the revised American College of Rheumatology 1987 criteria for the classification of RA [54]. RA patients were examined for the diagnosis of ILD, based on

the findings by chest radiography or HRCT. Images were reviewed by two experienced physicians. The scans covered the whole lung with a slice thickness of 1-2 mm and a high spatial resolution reconstruction image algorithm. The findings of ILD included usual interstitial pneumonia, non-specific interstitial pneumonia, and ground-glass attenuation patterns. Those results were categorized from A to Z, on the basis of the Sagamihara Criteria as follows, A: findings of ILD were observed in HRCT images (length of shorter diameter of the lesion was more than 2 cm in any transverse section); B: findings of ILD were observed in conventional chest CT images (length of shorter diameter of the lesion was more than 2 cm in a transverse section); C: findings of ILD were observed in HRCT images (length of shorter diameter of the lesion was less than 2 cm in any transverse section); D: findings of ILD were observed in conventional chest CT images (length of shorter diameter of the lesion was less than 2 cm in any transverse section); E: findings of ILD were observed in chest radiograms; F: abnormalities were not observed in chest radiograms; G: abnormalities were not observed in conventional chest CT images; H: abnormalities were not observed in HRCT images; X: findings from lung HRCT images were predominantly other than ILD, including bronchiectasis, bronchiolitis, emphysema, organizing pneumonia, tuberculosis, and cancer; Y: findings from conventional chest CT images were predominantly other than ILD, if HRCT images were unavailable.; Z: findings from chest radiograms were predominantly other than

ILD, if CT images were unavailable. RA patients assigned to A to D were RA with ILD group and those to G and H were RA without ILD group. RA patients categorized in X to Z were excluded from the present study. The causality of MTX or disease-modifying antirheumatic drugs (DMARDs) to ILD is a well-known fact. Therefore, to avoid interaction between the therapeutic drug and RA-ILD, drug induced interstitial lung disorder (DI-ILD) patients were excluded from this study. DI-ILD was diagnosed by the following diagnostic criteria: In brief, acute or sub-acute exacerbation of ILD after treatment in RA patients, which was not caused by *Pneumocystis jirovecii* in RA patients were diagnosed as DI-ILD [55]. In this study, RA with ILD patients and RA patients without ILD after exclusion of DI-ILD were designated as ‘‘RA with ILD” and ‘‘RA without ILD’’, respectively.

2.1.2 Preparation of DNA for genome-wide SNP genotyping

The concentration of each extracted genomic DNA specimens from peripheral blood was measured using an optical density meter (Nanodrop ND-1000 spectrophotometer) and diluted to 10 ng/ul with EDTA.

2.1.3 Genome-wide SNP genotyping

Genome-wide SNP genotyping was conducted using the Affymetrix Axiom

Genome-Wide ASI 1 Array, according to the manufacturer's instructions. The Axiom Genome-Wide ASI 1 Array contains more than 600,000 SNP loci that were optimized to maximize the coverage for Asian populations based on the HapMap project data. Adjusted genomic DNA specimens were amplified and randomly fragmented into sequences of 25 to 125 base-pairs (bp). After purification, these were hybridized to the array. Each SNP was identified via multi-color ligation. After ligation, the assays were stained and imaged were acquired on the detector (GeneTitan MC Instruments) (**Figure 1**). All genotyping experiments were performed by experienced technical staff in the laboratory.

Genotype calling was performed using the automatic call system implemented in the Affymetrix Genotyping Console Software v4.0, which employed the Axiom GT1 genotype calling algorithm. It generated Dish QC, the recommended sample quality score for the Axiom arrays. The samples with Dish QC score less than 0.82 were considered to be of low quality [56] and therefore excluded. In some cases, this genotype calling may be assigned incorrectly due to batch effect or inaccurate information of prior probability provided by Affymetrix. Therefore, the scatter images belonging to SNPs of interest that were obtained from the results of genotyping were confirmed visually.

2.1.4 Quality control of SNPs and samples

Steps for quality control of SNPs (SNP QC) consisted of filtering out (1) SNPs of > 1% untyped SNPs among the tested samples; (2) SNPs with minor allele frequency (MAF) of < 5%; and (3) SNPs with *p*-value of $< 1 \times 10^{-3}$ as calculated by the Hardy-Weinberg equilibrium (HWE) test. Quality control of sample (sample QC) was applied by filtering out samples with > 10% untyped SNPs among all SNPs. Pairwise identity-by-descent (IBD) estimation was calculated to remove samples with unknown and unaware kinship or potential contamination of DNA. SNP QC and sample QC were performed using the PLINK software [57].

2.1.5 Principal component analysis

To address population stratification and remove outliers in this study, principal component analysis (PCA) was conducted using EIGENSOFT 4.2 [58, 59]. HapMap data were used as references. The PLINK formatted data of HapMap project phase 3 (hg18) was downloaded from its homepage [60]. HapMap reference data included 180 Caucasians who were Utah residents with northern and western European ancestry from the CEPH collection (CEU), 180 Yorubans from Nigeria (YRI), 91 Japanese from Tokyo (JPT), 90 Han Chinese from Beijing (CHB) and 100 Chinese from Metropolitan Denver, Colorado (CHD). First,

human genome assembly of genotyped data was converted from hg19 to hg18 in order to synchronize with the reference data using LiftOver [61]. Common SNPs were selected from the genotype data of the present study as well as the HapMap reference data. PCA was conducted for the group of five HapMap populations (CEU, YRI, CHB, CHD and JPT) and also for data from the present study, using genotype data of common SNPs. To confirm the genetic background, PCA was conducted again for the three East Asian populations (JPT, CHB and CHD) and for data from the present study. After calculation of eigenvector values, results using the first and second eigenvector values were generated by EIGENSOFT 4.2 [58, 59].

2.1.6 Genome-wide association analysis

An association test was performed under five models (allelic, dominant, recessive, genotypic model and Cochran Armitage trend test) for each SNP that had passed quality control protocols using PLINK software [57]. The allelic, genotypic, dominant, and recessive models were tested by chi-square test, if the expected number in each cell of the contingency table was greater than five. Fisher's exact test was conducted if one of the expected counts of the cells is less than five. If one of the 2-by-2 cells included 0, then Woolf's correction was applied to calculate the odds ratios. For multiple testing correction, the Bonferroni correction was adopted. The quantile-quantile (Q-Q) plot was depicted with the expected distribution under

the null hypothesis across observed allelic model association test statistics among SNPs, by IPGWAS software [62]. Population stratification may generate false-positive associations between SNPs and the corresponding phenotype. If the population stratification seem to be indicated from the samples, then the genomic inflation factor (λ) of the Q-Q plot would be much greater than 1.00. Genome-wide Manhattan plot was generated by Haploview [63]. Regional Manhattan plot was drawn by Locuszoom [64], which provided LD information derived from the 1000 Genomes Project and recombination fractions from HapMap data, and showed names and structures of gene around the region of interest.

2.1.7 Selection of candidate regions

Candidate regions for further analysis were selected by the lowest *p*-value of a SNP that was less than 1×10^{-6}, as indicated by the five models (allelic, dominant, recessive, genotypic model and Cochran-Armitage trend test).

2.1.8 Imputation analysis of SNPs

Imputation methods were used to estimate genotypes without typing to increase the chances of uncovering novel and significant associations by estimation of haplotypes using a genotyped panel, which included denser SNPs. Generally, stronger LD, lower MAF and higher

marker-density lead to a more accurate estimation of haplotypes and imputation of untyped loci. In this study, IMPUTE2 [65] was used to perform imputation analysis to predict the genotypes of untyped or missing SNPs, as IMPUTE2 offers the advantage of higher computational performance and lower error rates, compared to MACH v1.0.16, fastPHASE v.1.4.0 and BEAGLE v3.2 [66]. Haplotype data obtained from the 1000 Genomes Project (August 2009 release) were used as reference panels because the data of 1000 Genomes Project demonstrated higher density than the Axiom array and included data for the Japanese population. In order to apply IMPUTE2, GTOOL [67] was first used to convert the genotype file from PLINK format to IMPUTE2 input format. After imputing the haplotypes, IMPUTE2 generated the probability that showed the accuracy of the imputation for each imputed SNP. A 1- Mb window size and 0.9 of the threshold of the imputation probability, which were recommended by the developer, were applied for each candidate region of GWAS. After imputation, association tests for all models were performed again using PLINK 1.7 [57]. Regional association plots after imputation analysis were constructed using LocusZoom [64]. After imputation analysis, SNPs with >1% un-imputed genotype data and *p*-value of $<1 \times 10^{-3}$ by HWE test were eliminated. Filtering criteria for MAF was not applied such that associations between the disease and imputed SNPs that have low MAF would not be detected.

2.1.9 Association analysis between SNP and gene expression level

To examine putative functions of the SNPs of interest, GENEVAR was used to query existing GWAS databases associated with *cis*-gene expression levels (eQTL analysis) [68]. The software utilized four genetic variations and gene expression profiling data; three tissue types (adipose, lymphoblastoid cell lines (LCL) and skin) collected from 856 healthy female twins of the MuTHER study [69], lymphoblastoid cell lines from 726 HapMap3, including 8 populations [70]; three tissue types (adipose, LCL and skin) derived from a subset of ~160 MuTHER healthy female twins [71]; and three cell types (fibroblast, LCL and T-cell) derived from the umbilical cords of 75 Geneva GenCord individuals) [72]. Possible association between the genotypes of the SNPs of interest and the expression levels of transcripts were examined. To perform the eQTL analysis for the candidate SNPs and genes, the database of HapMap3 JPT derived from lymphoblastoid cell lines was used for the application. eQTL analysis was conducted using the following conditions: Spearman's rank correlation coefficient for correlation and regression test were examined between candidate SNPs and the expression levels of the gene that were within 1-Mb distance to the SNPs and the *p*-value threshold was less than 0.001.

HaploReg v2 [73] was utilized to append functional annotations of SNPs regarding

sequence conservation across mammals, the number of cell types that showed a chromatin state of promoter histone marks or enhancer histone marks, and the number of cell types that demonstrated DNase hypersensitivity. Taken together, these data provided estimations that may determine if the SNPs of interest are regulatory SNPs. This software also provided functional annotations for flanking SNPs or small indels of the candidate SNPs using LD information from the 1000 Genomes Project.

2.1.10 Statistical analysis of association by gender and by age

To examine the association by gender, chi-square test or Fisher's exact test was conducted according to an expected value of five. Shapiro-Wilk test was conducted to ascertain a normal distribution of the age in both RA with ILD and RA without ILD. If rejected by the Shapiro-Wilk test, the Wilcoxon rank sum test was performed. Otherwise, the Student's *t*-test was used. Shapiro-Wilk test, Wilcoxon rank sum test and the Student's *t*-test were all conducted among R [74].

2.1.11 *HLA* alleles of subjects

The information of *HLA* alleles was provided by my collaborator, Dr. Hiroshi Furukawa, and included data about the alleles of six *HLA* genes (*HLA-A, HLA-B, HLA-C,*

HLA-DRB1, *HLA-DQB1* and *HLA-DPB1*).

2.1.12 Association test of *HLA* alleles

Association analysis of *HLA* was performed using the two-digit allele of the *HLA-DRB1* gene between RA with ILD and RA without ILD by Fisher's exact test. If one of cells in a 2-by-2 table included zero, Woolf's correction was applied to calculate the odds ratio. Bonferroni correction was adopted to adjust for multiple testing by multiplying the number of alleles of the *HLA-DRB1* in both RA with ILD and RA without ILD.

2.1.13 Confirmation of the association results by retyping

To confirm the result of the genome-wide SNP genotyping, validation was required using different disciplines to determine the genotypes. TaqMan assay was conducted according to the manufacturer's protocol. Amplification and polymerization reactions were carried out on 384-well plates, and fluorescence was measured using the Roche LightCycler 480 system.

2.2 Sub-group analysis divided by presence/absence of *HLA-DRB1*04*

To control the effect of *HLA-DRB1*04* which is known as a risk factor for RA, a

sub-group analysis was performed. Subjects were divided into two groups according to the presence or absence of *HLA-DRB1*04*. Association tests for all genotyped SNPs were conducted for these two groups. Similar statistical analysis, regional imputation analysis and association analysis between candidate SNPs and gene expression levels were conducted.

2.3 Comparison of the reported associations in RA and PF with association results of the RA-ILD GWAS

To compare the genetic background between RA-ILD and RA, and between RA-ILD and PF, the association results of reported RA and PF susceptibility SNPs were compared with those of RA-ILD.

A total of 101 RA susceptibility SNPs were reported using a 3-stage meta-analysis for trans-ethnic populations. Twenty RA susceptibility SNPs with *p*-values of less than 5×10^{-6} in the Asian population in their stage 1 meta-analysis, were selected from the previous study [53].

Eleven PF [38] susceptibility SNPs with reported *p*-values that reached genome-wide significance level (*p*-value $< 5 \times 10^{-8}$), and rs35705950 in *MUC5B* by genome-wide linkage analysis [35], were also selected.

The SNPs in high LD with r^2 of $\geqq$ 0.8 in Asian population of the 1000 Genomes

Project with selected SNPs were also selected and defined as proxy SNPs. For those selected SNPs that were untyped or missing in the present study, proxy SNPs were used instead. If more than two proxy SNPs were present, one of the proxy SNPs was selected using the following criteria: lower *p*-value, higher r^2 value and nearer physical position. The susceptibility SNPs of the present study were also compared with RA and PF susceptibility SNPs using *p*-values and ORs (Odds Ratios).

3. Results

3.1 GWAS of RA-ILD

3.1.1 Quality control of SNPs and samples

In this thesis, genome-wide association study was performed to compare the allele frequencies in RA patients with ILD to those in RA patients without ILD. Genome-wide SNP genotyping was performed using the Affymetrix Axiom platform on 620 recruited patients. The genotyping data of six of the samples were not obtained due to scan error or low Dish QC. The remaining genotype data were subjected to quality control for sample QC and SNP QC. After excluding 116,313 SNPs with SNP of >1% untyped SNPs among the tested samples, 156,323 SNPs with MAF of <5% and 435 SNPs with *p*-value in RA without ILD of $<1 \times 10^{-3}$ calculated by HWE test, 359,782 SNPs passed the SNP QC. One sample was excluded by sample QC due to low quality data.

To avoid unknown or unintended inclusion of relatives in both RA with ILD and RA without ILD, pairwise IBD values for each pair were calculated using the PLINK software [57]. Two genotyped data were excluded from RA without ILD group and were not subjected to subsequent analyses. No relatives were included in the following analyses.

3.1.2 Principal component analysis

PCA was conducted to exclude outliers with different genetic backgrounds from the majority of study samples. A total of 190,035 SNPs were used for PCA with five populations (JPT, CHB, CHD, CEU and YRI) and RA patients after applying SNP QC. From the combined population results, East Asian HapMap populations and RA patients contributed a single cluster, and CEU and YRI populations constituted distinct clusters (**Figure 2 A**). A total of 208,320 SNPs were used for PCA in detail with East Asian HapMap populations (JPT, CHB, and CHD) and RA patients. The PCA of the East Asian HapMap population and RA patients showed two clusters, one consisting of RA patients and JPT data, and the other was clearly composed mainly of CHB and CHD population. Only one individual from the RA patients was located at the border of the Chinese population cluster, which indicated that the individual possessed a different genetic background from the Japanese population (**Figure 2 B**). Therefore, this individual was excluded from subsequent analysis.

3.1.3 Association results of RA with ILD compared with RA without ILD

After SNP QC, sample QC, PCA and IBD ascertainment, genome-wide association tests were conducted using 169 patients that have RA with ILD and 294 patients with RA without ILD. The male gender contributed to a higher risk factor of RA with ILD compared to RA without ILD (**Table 1**).

To verify the population stratification, a Q-Q plot was constructed (**Figure 3 A**). No population stratification was observed because the genomic inflation factor was modest (λ = 1.02). After excluding the *HLA* region [75], λ-value decreased from 1.02 to 1.01 (**Figure 3 B**). Genome-wide Manhattan plots were drawn according to the chromosomal positions of individual SNPs (x axis) and the negative logarithm of *p*-value (y axis) under the Cochran-Armitage trend test and dominant model (**Figure 4** and **Figure 5**).

The association test between RA with ILD and RA without ILD identified 11 SNPs with minimum *p*-values of less than 5×10^{-6} according any of the five models (**Table 2**). No SNP showed a significant *p*-value as calculated by Bonferroni correction (*p*-value = 0.05 / 359,782 = 1.39×10^{-7}). The region 19q13.11 and 12q24.11 showed *p*-values of less than 1×10^{-6}. Therefore, SNPs within these regions were applied to the subsequent analysis, including regional imputation analysis, eQTL analysis by GENEVAR and assessment of functional annotations of SNPs in the region by Haploreg v2. In addition, the SNP marker_H was also a

point of focus because it was located in the intragenic region.

The SNP marker_A1, showed the lowest *p*-value (MAF: 0.27 / 0.14 and *p*-value: 5.98 × 10^{-7} calculated by Cochran-Armitage trend test) (**Figure 6 A**). The nearest gene, which encodes a zinc finger protein, was located 150 kb upstream of the SNP. From a regional imputation analysis, 13 SNPs showed lower *p*-values than that of the genotyped SNP and the lowest *p*-value was 1.87 × 10^{-7} (**Figure 6 B**). One SNP, marker_a14, which was in high LD (r^2= 0.82) with the SNP marker_A1, was shown to exhibit a promoter chromatin state and was located in DNase hypersensitivity site (**Table 3**). No *cis*-eQTL association for both this SNP and the flanking SNPs with high LD ($r^2 > 0.8$) was found.

The SNP marker_B in the region 12q24.11 showed an association (MAF: 0.30/0.46, *p*-value: 6.86 × 10^{-7}, OR: 0.37 and 95%CI: 0.25-0.55 under dominant model) (**Figure 7 A**). The gene containing this SNP in 11th intron encodes a transporter-like protein that is expressed in the entire brain. According to the result of the regional imputation analysis, none of the SNPs showed a stronger association than the SNP marker_B (**Figure 7 B**). The genotyped and imputed SNPs were not located in any of the exons of this gene. An eQTL analysis could not detect any association. This SNP and flanking SNPs exhibited enhancer histone marks and DNase hypersensitivity sites by HaploReg v2 only in two cell types (data not shown).

The SNP marker_H located in the 5th intron of a gene that codes for a

phosphodiesterase enzyme in the region 7p14.3, showed an association (MAF: 0.13/0.26, *p*-value: 4.38×10^{-6}, OR: 0.44 and 95%CI: 0.31-0.63 under allelic model) (**Figure 8 A**). From the regional imputation analysis, seven SNPs showed lower *p*-values than that of the SNP marker_H and the lowest *p*-value was 2.01×10^{-6} (**Figure 8 B**). Genotyped SNPs and imputed SNPs that showed relatively strong associations were not exonic. From Haploreg v2, the region around the SNP marker_H was shown to have enhancer histone marks in three cell types and DNase hypersensitivity in one cell types. (data not shown).

3.2 Allele frequencies of *HLA-DRB1* in the RA patients

To exclude the effect of HLA class II region, a sub-group analysis was performed. Prior to conducting a sub-group analysis, HLA association analysis was carried out to uncover susceptibility of HLA alleles. The *HLA-DRB1*04* allele showed the strongest and most protective association with RA with ILD (corrected *p*-value: 2.11×10^{-4}, OR: 0.56 and 95%CI: 0.42-0.75). On the other hand, the *HLA-DRB1*15* allele indicated an increasing risk of RA with ILD with corrected *p*-value of 6.20×10^{-3}, OR of 1.73 and 95% CI of 1.23 to 2.43 (**Table 4**).

3.3 Sub-group analysis divided by presence/absence of *HLA-DRB1*04*

3.3.1 Association results of *HLA-DRB1*04* positive RA with ILD compared to RA without ILD

By focusing on *HLA-DRB1*04* positive patients among those having RA with ILD, relevant associations were reevaluated. Exclusively for the patients who possessed the *HLA-DRB1*04* allele, association tests were conducted for 85 *HLA-DRB1*04* positive RA patients with ILD and 197 *HLA-DRB1*04* positive RA patients without ILD.

The Q-Q plot showed that no population stratification was found with genomic inflation factor of 1.01 (**Figure 9**). Manhattan plot showed that one clear peak existed on

chromosome 7p21.3, which could not be detected before division (Figure 10).

A total of 9 SNPs in the 7p21.3 showed low *p*-values of less than 1×10^{-6}. The SNP marker_I1 in the region 7p21.3 showed the strongest association (MAF: 0.41/0.2, *p*-value: 4.20×10^{-7}, OR: 2.71 and 95%CI: 1.83-4.00 under allelic model) (**Table 5** and **Figure 11 A**). This SNP was located 400 kb upstream from a gene. Imputation analysis showed that no SNPs exhibited stronger association than the SNP (**Figure 11 B**). The result of eQTL analysis did not show any *cis*-eQTL association with any genes. Analysis with HaploReg v2 demonstrated that the region around the SNP marker_I1 did not have any enhancer histone marks, promoter histone markes or DNase hypersensitivity sites. (data not shown). The flanking SNPs marker_I3, which was in high LD ($r^2 = 0.97$ in Asian population), was reported as a protein quantitative locus of TGF-β1 [76].

3.3.2 Association results of *HLA-DRB1*04* negative RA with ILD compared to RA without ILD

After selecting the samples that did not contain the *HLA-DRB1*04* allele, association tests were conducted for 84 *HLA-DRB1*04* negative RA patients with ILD and 97 *HLA-DRB1*04* negative RA patients without ILD. No SNP was revealed to be significant after the Bonferroni correction.

The Q-Q plot produced a genomic inflation factor of 1.01, which did not involve any population stratification (**Figure 12**). Manhattan plot showed that one cluster existed in the region of 10q23.1 (**Figure 13**).

The SNP marker_K1 in the region 10q23.1 located on the second intron showed an association (MAF: 0.19/0.04, *p*-value: 2.72×10^{-6}, OR: 5.77 and 95%CI: 2.95-12.89) (**Table 6** and **Figure 14 A**). The gene containing the SNP marker_K1 is a member of the neuregulin gene family, which encodes ligands for the transmembrane tyrosine kinase receptors. From a regional imputation analysis, one SNP showed lower *p*-values (1.80×10^{-6}) than the genotyped SNP (**Figure 14 B**). None of the genes showed *cis*-eQTL effects of the SNP marker_K1. The SNP marker_K1 and flanking region appeared to possess enhancer histone marks in two cell types and DNase hypersensitivity sites only in one cell type, as revealed by HaploReg v2 (data not shown).

3.4 Validation analysis of results of the RA-ILD GWAS results

Validation analysis was conducted to confirm the accuracy of genotyping by GWAS using TaqMan assay. Minimum *p*-values among 5 models (allelic, dominant, recessive, genotypic and Cochran-Armitage trend test) of the SNPs marker_A1, marker_B and marker_E located in regions of 19q13.11, 12q24.11 and 6p21.32, respectively, from RA-ILD GWAS results, were verified by TaqMan assay with minimum *p*-values of 7.66×10^{-7}, 2.95×10^{-7} and 5.47×10^{-6}, respectively (**Table 7**). The SNP marker_I6, which demonstrated association in the results of *HLA-DRB1*04* positive RA with ILD compared to RA without ILD, was verified by TaqMan assay with minimum *p*-value of 2.93×10^{-6}. The SNP marker_K1 in 10q23.1, which had the lowest *p*-value in the association results of *HLA-DRB1*04* negative RA with ILD compared to *HLA-DRB1*04* negative RA without ILD, was verified with minimum *p*-value of 4.74×10^{-6}.

3.5 Comparison of the reported associations of RA and PF with association results of the RA-ILD GWAS

To reveal possible overlap of susceptibility SNPs to RA or PF with those of RA-ILD, reported RA and PF susceptibility SNPs were selected and compared with results derived from the present study.

Firstly, association results of RA-ILD and reported RA susceptibility SNPs were compared. A total of 101 RA susceptibility SNPs were listed from a recent meta-analysis of trans-ethnic populations [53]. Twenty RA susceptibility SNPs and those of proxy SNPs were selected from the previous study and compared with the association results of RA-ILD GWAS (**Table 8**). Nine of 20 RA susceptibility SNPs were not available in the present genotyped SNPs. Ten of 20 RA susceptibility SNPs did not show an association with RA-ILD. The remaining one SNP rs9268839 located near the *HLA-DRB1* gene only showed a moderate association, with a *p*-value of 1.60×10^{-3}. The SNP rs9268839 showed OR of >1 in the previous RA meta-analysis, but had an OR of <1 in the present study. Furthermore, the susceptibility SNPs and proxy SNPs of the present study for *HLA-DRB1*04* positive/negative patients did not overlap with RA susceptibility SNPs.

Secondly, 11 PF susceptibility SNPs, of which *p*-values reached genome-wide significance level (p-value $< 5 \times 10^{-8}$) in the reported GWAS [38], and rs35705950 in the

MUC5B gene as identified by genome-wide linkage analysis [35] were selected to compare to the association results of the present study (**Table 9**). Seven of the 12 PF susceptibility SNPs were not available in the present data. The remaining five PF susceptibility SNPs or those of proxy SNPs did not show an association with RA-ILD. Moreover, the susceptibility SNPs of the present study and those of proxy SNPs did not overlap with PF susceptibility SNPs.

4. Discussion

A genome-wide association study was conducted on a group of RA-ILD patients. DI-ILD RA patients were excluded in order to specifically target and detect RA-ILD susceptibility SNPs instead of drug susceptibility SNPs. Subjects of the male gender appeared to be at higher risk for RA with ILD, compared to RA without ILD (**Table 1**), consistent with previous reports [10]. The genome-wide association tests were conducted by employing 359,782 SNPs that have passed quality control protocols, including sample QC, SNP QC, PCA and IBD assessments. Initially, 169 RA patients with ILD and 294 RA patients without ILD were compared. No population stratification was detected, as determined by calculation of genomic inflation factor after examination of the Q-Q plot. This present GWAS identified 11 SNPs in eight regions with minimum *p*-value of less than 5×10^{-6}; 5q35.2, 6q21.32, 7p14.3, 12q23.3, 12q24.11, 19q13.11, 19q13.12 and 22q13.1.

The SNP marker_A1 in the region 19q13.11 showed the strongest association with *p*-value of 5.98×10^{-7} under the Cochran-Armitage trend test. The nearest gene that encodes a zinc finger protein was located 150 kb upstream from this SNP. To date no biological function has been reported to be associated with this gene. A regional imputation analysis showed that one SNP demonstrated stronger association with a *p*-value of 1.87×10^{-7} (**Figure 6**). The flanking SNP marker_a14 in high LD with the SNP marker_A1 was strongly predicted to

have a possible regulatory function by using HaploReg v2 (**Table 3**). However, no *cis*-eQTL association was detected. The possibility of mis-genotyping was relatively low for this SNP because neighboring SNPs in high LD also showed low *p*-values. The associated biological functions remained to be elucidated.

The SNP marker_B located at 12q24.11 showed the second lowest *p*-value. Imputation analysis could not detect any SNPs with stronger associations. The gene containing the SNP encodes evolutionarily conserved synaptic vesicle protein [77]. Mouse homolog of this gene is expressed in the murine central nervous system [78]. For humans, this gene appeared to be associated with uremia [79]. Another study indicated that a copy number variation of this gene was observed in sporadic, early-onset Alzheimer disease [80]. In addition, structural similarity of this gene to the drug transporter SLC 22 family has been suggested [81]. This raises the possibility and potential that therapeutic agents could affected RA-ILD onset. Gene expression data were reported in mice [82] and in the human lung [83]. Moreover, a literature search seems to suggest that the relationship between this gene and RA, ILD or PF have not yet been investigated.

The SNP marker_H in the gene located in 7p14.3 displayed an association. This gene belongs to the cyclic nucleotide phosphodiesterases gene family, which is responsible for catalyzing the hydrolysis of cAMP and cyclic guanosine monophosphates [84]. Gene

expression in human tissues is low in lung and high in heart [85, 86]. Phosphodiesterase-inhibiting drugs can suppress TGF-β-induced differentiation of lung fibroblast into myofibroblast, which is a histophysiological feature of ILD and chronic obstructive pulmonary disease [45]. This gene is relevant to fibroblast growth factor receptor (FGFR) signaling pathway. FGFR is associated with idiopathic pulmonary arterial hypertension which is complicated to IPF [87]. Basic fibroblast growth factor, a potent mitogenic factor for smooth muscle cells, myofibroblasts, and fibroblasts, is critical for IPF [88]. This finding suggested that IPF and RA-ILD share some genetic factor through this gene and/or FGFR signaling pathway.

HLA analysis using *HLA-DRB1* allele was conducted to reveal the association between RA with ILD and RA without ILD. *HLA-DRB1*04* showed a protective OR. This result was partially consistent with a previous report, owing to the fact that some samples were redundant [34]. This result was contrary to other previous reports, which suggested that *HLA-DRB1*04* was a risk factor for RA in the Caucasian population [89], and *HLA-DRB1*0405* was a risk factor for RA in the Japanese population [34]. This study implied that RA with ILD and RA without ILD possessed different genetic backgrounds. Thus, sub-group analysis was conducted to adjust for the effect of *HLA-DRB1*04*.

Association tests between *HLA-DRB1*04* positive RA with ILD and RA without ILD

uncovered an association to the SNP marker_I1 in 7p21.3. The flanking SNP marker_I3 (p-value = 8.26×10^{-7}) is reported as a protein quantitative trait locus that significantly regulates TGF-β1 protein expression [76]. The *TGF-β* gene has previously described as a risk factor for interstitial lung disease in RA patients [42]. Fibroblast, epithelial cells, endothelial cells, dendritic cells and macrophages produce *TGF-β* [43]. *TGF-β* contributes to the differentiation of fibroblasts into myofibroblasts, which are the main constituents of the extracellular matrix in lung fibrogenic processes [47]. Thus, this region may be implicated in the process and lead to RA-ILD.

Association tests of *HLA-DRB1*04* negative RA with ILD as compared to RA without ILD have detected an association to the SNP marker_K1 in the second intron in 10q23.1. The gene containing the marker, is a member of the neuregulin (*NRG*) gene family and encodes ligands for the transmembrane tyrosine kinase receptors. Expression of this gene was high in human lung [86], and it has been reported to demonstrate an association with schizophrenia and schizoaffective disorder in a Chinese population [90]. Another study suggested that this gene was associated with schizophrenia and bipolar disorder in Australia [91]. It was hypothesized that this gene influences neuroblast proliferation, migration, and differentiation through the ErbB4 signaling pathway. Moreover, the ErbB signalling pathway is known to demonstrate an association with acute pulmonary inflammation of lung interstitial

cells such as smooth muscle cells [92]. Another study described that NRG and its transmembrane receptor ErbB4 function as a transcriptional cofactor for the expression of surfactant protein B in the murine fetal lung [93]. Thus, this SNP might affect RA susceptibility through its function of changing the surface tension function. This region was detected by sub-group analysis of *HLA-DRB1*04* negative patients. The relationship between this gene and the specific *HLA-DRB1*04* allele remains unclear. Further analysis would therefore be needed to elucidate the interaction between the ErbB pathway or this gene and the specific *HLA-DRB1* alleles.

Comparison with the susceptibility SNP from the present study and reported RA susceptibility SNPs revealed no overlap of associations. This was a plausible result, because all of the subjects were RA patients and shared risks for RA. No susceptibility SNP of PF from previous study showed an association in the present study. Half of the susceptibility SNPs reported in RA and PF research turned out to be unavailable. Thus, further analysis is required to reveal any overlap between susceptibility SNPs of PF and RA-ILD.

The present study has several limitations. First, RA-ILD is a heterogeneous disease and misdiagnoses are possible. DI-ILD patients were excluded so that susceptible SNPs to drug responses were not inappropriately detected. Acute pneumonitis patients were excluded by definition of DI-ILD; however, chronic drug-induced ILD patients were included in

RA-ILD. Nevertheless, this may lead to a reduction in statistical power. Secondly, associations demonstrated in this study could not achieve the genome-wide significance level (p-value $< 5 \times 10^{-8}$) due to the limited sample size. If the sample size of both RA with ILD and RA without ILD can be increased by 23%, allelic p-value would be able to reach the genome-wide significant level. After the addition of more samples, further analytical study such as GWAS or replication study would be necessary in order to verify the results of the present study. Thirdly, the GWAS that was carried out in this study was based on the hypothesis of "common disease common variant". Thus, the utilized platforms for GWAS were designed using common variants. Rare mutations that cause a common disease could not be detected by GWAS. Imputation analysis, however, can compensate for this weakness because this method allows for the imputation of SNPs that were not genotyped and had lower MAF, as determined using denser reference data.

5. Conclusion

A Genome-wide association study was conducted for 169 RA patients with ILD and 294 RA patients without ILD. This GWAS identified 2 SNPs with minimum *p*-values of less than 1×10^{-6}. The SNP marker_A1 in the region 19q13.11 demonstrated the strongest association. The flanking SNP marker_a14 in high LD with the SNP marker_A1 was strongly predicted as possessing a possible regulatory function. The SNP marker_B located at 12q24.11 showed the second lowest *p*-value. The gene containing the SNP was reported to encode an evolutionarily conserved synaptic vesicle protein. The biological functions represented by the SNPs to RA-ILD still remain unclear.

Sub-group analysis was conducted to adjust for the effects of *HLA-DRB1*04*. Association tests between *HLA-DRB1*04* positive RA with ILD and RA without ILD revealed an association of the SNP marker_I1. The flanking SNP marker_I3 of the SNP marker_I1 is considered a protein quantitative trait locus that regulates the TGF-β1 protein expression. Association tests of *HLA-DRB1*04* negative RA with ILD compared to RA without ILD detected an association with the SNP marker_K1. It was revealed that the gene containing the marker was a member of the NRG gene family.

Comparison with the results from previously reported RA GWAS and that from the present study revealed no overlap of an association. In addition, no susceptibility SNP of PF

showed an association in the present study.

The present study has limitations on heterogeneity of RA-ILD and sample size. For further analysis, it is required to increase sample size and to performe a replication study.

Acknowledgements

My heartfelt appreciation goes to Professor Katsushi Tokunaga (Department of Human Genetics, Graduate School of Medicine, University of Tokyo), for supervising my study and providing me an opportunity to complete my substantial project in three years. I am thankful to Associate Professor Akihiko Mabuchi, whose knowledge, experience and suggestions have helped me and enabled me to develop an understanding of the subject. I am very thankful to assistant professor Taku Miyagawa, Dr. Hiroko Miyadera, Dr. Yuki Hitomi, Dr. Nao Nishida, Dr. Hiromi Sawai, Dr. Yosuke Ohmae and Dr. Khor Seik Soon for their enlightening and insightful suggestions to this study. I would like to extend my gratitude to Dr. Shigeto Tohma (Sagamihara Hospital), Professor Naoyuki Tsuchiya and Dr. Jun Ohhashi (Tsukuba University). A special thank is dedicate to Dr. Hiroshi Furukawa (Sagamihara National Hospital) for his guidance in interstitial lung disorder and rheumatoid arthritis. I am indebted to my many of colleagues in our laboratory.

References

1. P. Ellman and R. E. Ball, Rheumatoid disease with joint and pulmonary manifestations, *Br. Med. J.*, vol. 4583: pp. 816–20, 1948.
2. A. Aronoff, E. Bywaters, and G. Fearnley, Lung lesions in rheumatoid arthritis, *Br. Med. J.*, vol. 2: pp. 228–32, 1955.
3. B. Stack and I. Grant, Rheumatoid interstitial lung disease, *Br. J. Dis. Chest*, vol. 59: pp. 202-11, 1965.
4. M. S. Popper, M. L. Bogdonoff, and R. L. Hughes, Interstitial rheumatoid lung disease. A reassessment and review of the literature, *Chest*, vol. 62, no. 3: pp. 243–50, 1972.
5. A. Young and G. Koduri, Extra-articular manifestations and complications of rheumatoid arthritis, *Best Pract. Res. Clin. Rheumatol.*, vol. 21: pp. 907–27, 2007.
6. G. Koduri, S. Norton, A. Young, N. Cox, P. Davies, J. Devlin, J. Dixey, A. Gough, P. Prouse, J. Winfield, and P. Williams, Interstitial lung disease has a poor prognosis in rheumatoid arthritis: results from an inception cohort, *Rheumatology (Oxford)*, vol. 49: pp. 1483–9, 2010.
7. S. Inokuma, Leflunomide-induced interstitial pneumonitis might be a representative of disease-modifying antirheumatic drug-induced lung injury, *Expert Opin. Drug Saf.*, vol. 10:

pp. 603–11, 2011.

8. R. Perez-Alvarez, M. Perez-de-Lis, C. Diaz-Lagares, J. M. Pego-Reigosa, S. Retamozo, A. Bove, P. Brito-Zeron, X. Bosch, and M. Ramos-Casals, Interstitial lung disease induced or exacerbated by TNF-targeted therapies: analysis of 122 cases, *Semin. Arthritis Rheum.*, vol. 41: pp. 256–64, 2011.
9. A. L. Olson, J. J. Swigris, D. B. Sprunger, A. Fischer, E. R. Fernandez-Perez, J. Solomon, J. Murphy, M. Cohen, G. Raghu, and K. K. Brown, Rheumatoid arthritis-interstitial lung disease-associated mortality, *Am. J. Respir. Crit. Care Med.*, vol. 183: pp. 372–8, 2011.
10. T. Bongartz, C. Nannini, Y. F. Medina-Velasquez, S. J. Achenbach, C. S. Crowson, J. H. Ryu, R. Vassallo, S. E. Gabriel, and E. L. Matteson, Incidence and mortality of interstitial lung disease in rheumatoid arthritis: A population-based study, *Arthritis Rheum.*, vol. 62: pp. 1583–91, 2010.
11. J. K. Dawson, H. E. Fewins, J. Desmond, M. P. Lynch, and D. R. Graham, Fibrosing alveolitis in patients with rheumatoid arthritis as assessed by high resolution computed tomography, chest radiography, and pulmonary function tests, *Thorax*, vol. 56: pp. 622–7, 2001.
12. E. Gabbay, R. Tarala, R. Will, G. Carroll, B. Adler, D. Cameron, and F. R. Lake, Interstitial lung disease in recent onset rheumatoid arthritis, *Am. J. Respir. Crit. Care Med.*, vol. 156:

pp. 528–35, 1997.

13. H. M. Habib, A. a Eisa, W. R. Arafat, and M. a Marie, Pulmonary involvement in early rheumatoid arthritis patients, *Clin. Rheumatol.*, vol. 30: pp. 217–21, 2011.

14. K. Shidara, D. Hoshi, E. Inoue, T. Yamada, A. Nakajima, A. Taniguchi, M. Hara, S. Momohara, N. Kamatani, and H. Yamanaka, Incidence of and risk factors for interstitial pneumonia in patients with rheumatoid arthritis in a large Japanese observational cohort, IORRA, *Mod. Rheumatol.*, vol. 20: pp. 280–6, 2010.

15. A Nakajima, E. Inoue, E. Tanaka, G. Singh, E. Sato, D. Hoshi, K. Shidara, M. Hara, S. Momohara, a Taniguchi, N. Kamatani, and H. Yamanaka, Mortality and cause of death in Japanese patients with rheumatoid arthritis based on a large observational cohort, IORRA, *Scand. J. Rheumatol.*, vol. 39: pp. 360–7, 2010.

16. T. R. Mikuls, H. Sayles, F. Yu, T. Levan, K. a Gould, G. M. Thiele, D. L. Conn, B. L. Jonas, L. F. Callahan, E. Smith, R. Brasington, L. W. Moreland, R. J. Reynolds, and S. L. Bridges, Associations of cigarette smoking with rheumatoid arthritis in African Americans, *Arthritis Rheum.*, vol. 62: pp. 3560–8, 2010.

17. D. Hutchinson, L. Shepstone, R. Moots, J. T. Lear, and M. P. Lynch, Heavy cigarette smoking is strongly associated with rheumatoid arthritis (RA), particularly in patients without a family history of RA, *Ann. Rheum. Dis.*, vol. 60: pp. 223–7, 2001.

18. Y. Tsuchiya, N. Takayanagi, H. Sugiura, Y. Miyahara, D. Tokunaga, Y. Kawabata, and Y. Sugita, Lung diseases directly associated with rheumatoid arthritis and their relationship to outcome, *Eur. Respir. J.*, vol. 37: pp. 1411–7, 2011.

19. B. Balbi, V. Cottin, S. Singh, W. De Wever, F. J. F. Herth, and C. Robalo Cordeiro, Smoking-related lung diseases: a clinical perspective, *Eur. Respir. J.*, vol. 35: pp. 231–3, 2010.

20. B. R. Gochuico, N. a Avila, C. K. Chow, L. J. Novero, H.-P. Wu, P. Ren, S. D. MacDonald, W. D. Travis, M. P. Stylianou, and I. O. Rosas, Progressive preclinical interstitial lung disease in rheumatoid arthritis, *Arch. Intern. Med.*, vol. 168: pp. 159–66, 2008.

21. S. Mori, Y. Koga, and M. Sugimoto, Different risk factors between interstitial lung disease and airway disease in rheumatoid arthritis, *Respir. Med.*, vol. 106: pp. 1591–9, 2012.

22. F. Aubart, B. Crestani, P. Nicaise-Roland, F. Tubach, C. Bollet, K. Dawidowicz, E. Quintin, G. Hayem, E. Palazzo, O. Meyer, S. Chollet-Martin, and P. Dieudé, High levels of anti-cyclic citrullinated peptide autoantibodies are associated with co-occurrence of pulmonary diseases with rheumatoid arthritis, *J. Rheumatol.*, vol. 38: pp. 979–82, 2011.

23. I. Alexiou, A. Germenis, A. Koutroumpas, A. Kontogianni, K. Theodoridou, and L. I. Sakkas, Anti-cyclic citrullinated peptide-2 (CCP2) autoantibodies and extra-articular manifestations in Greek patients with rheumatoid arthritis, *Clin. Rheumatol.*, vol. 27: pp.

511–3, 2008.

24. Y. Yin, D. Liang, L. Zhao, Y. Li, W. Liu, Y. Ren, Y. Li, X. Zeng, F. Zhang, F. Tang, G. Shan, and X. Zhang, Anti-cyclic citrullinated Peptide antibody is associated with interstitial lung disease in patients with rheumatoid arthritis, *PLoS One*, vol. 9: p. e92449, 2014.
25. N. Inui, N. Enomoto, T. Suda, Y. Kageyama, H. Watanabe, and K. Chida, Anti-cyclic citrullinated peptide antibodies in lung diseases associated with rheumatoid arthritis, *Clin. Biochem.*, vol. 41: pp. 1074–7, 2008.
26. L. Jearn and T. Kim, Level of anticitrullinated peptide/protein antibody is not associated with lung diseases in rheumatoid arthritis, *J. Rheumatol.*, vol. 39:pp. 1493-94, 2012.
27. T. I. Evans, J. Han, R. Singh, and G. Moxley, The genotypic distribution of shared-epitope DRB1 alleles suggests a recessive mode of inheritance of the rheumatoid arthritis disease-susceptibility gene, Arthritis Rheum., vol. 3: pp. 1754–61, 1995.
28. S. Wakitani, N. Murata, Y. Toda, R. Ogawa, T. Kaneshige, Y. Nishimura, and T. Ochi, The relationship between HLA-DRB1 alleles and disease subsets of rheumatoid arthritis in Japanese, *Br. J. Rheumatol.*, vol. 36：pp. 630–6, 1997.
29. P. K. Gregersen, J. Silver, and R. J. Winchester, The shared epitope hypothesis. An approach to understanding the molecular genetics of susceptibility to rheumatoid arthritis, Arthritis Rheum., vol. 30: pp. 1205–13, 1987.

30. C. Turesson, C. M. Weyand, and E. L. Matteson, Genetics of rheumatoid arthritis: Is there a pattern predicting extraarticular manifestations?, *Arthritis Rheum*, vol. 51: pp. 853–63, 2004.

31. J. S. S. Lanchbury, E. E. M. Jaeger, D. M. Sansom, M. a. Hall, P. Wordsworth, J. Stedeford, J. I. Bell, and G. S. Panayi, Strong primary selection for the Dw4 subtype of DR4 accounts for the HLA-DQw7 association with Felty's syndrome, *Hum. Immunol.*, vol. 32: pp. 56–64, 1991.

32. H. Y. Kim, J. K. Min, H. I. Yang, S. H. Park, Y. S. Hong, W. H. Jee, S. H. Lee, C. S. Cho, T. G. Kim, and H. Han, The impact of HLA-DRB1*0405 on disease severity in Korean patients with seropositive rheumatoid arthritis, *Rheumatology*, vol. 36: pp. 440–3, 1997.

33. K. Migita, T. Nakamura, T. Koga, and K. Eguchi, HLA-DRB1 alleles and rheumatoid arthritis-related pulmonary fibrosis, *J. Rheumatol.*, vol. 37: pp. 1–4, 2010.

34. H. Furukawa, S. Oka, K. Shimada, S. Sugii, J. Ohashi, T. Matsui, T. Ikenaka, H. Nakayama, A. Hashimoto, H. Takaoka, Y. Arinuma, Y. Okazaki, H. Futami, A. Komiya, N. Fukui, T. Nakamura, K. Migita, A. Suda, S. Nagaoka, N. Tsuchiya, and S. Tohma, Association of human leukocyte antigen with interstitial lung disease in rheumatoid arthritis: a protective role for shared epitope, *PLoS One*, vol. 7: p. e33133, 2012.

35. M. A. Seibold, A. L. Wise, M. C. Speer, M. P. Steele, K. K. Brown, J. E. Loyd, T. E.

Fingerlin, W. Zhang, G. Gudmundsson, S. D. Groshong, C. M. Evans, S. Garantziotis, K. B. Adler, B. F. Dickey, R. M. du Bois, I. V Yang, A. Herron, D. Kervitsky, J. L. Talbert, C. Markin, J. Park, A. L. Crews, S. H. Slifer, S. Auerbach, M. G. Roy, J. Lin, C. E. Hennessy, M. I. Schwarz, and D. A. Schwartz, A common MUC5B promoter polymorphism and pulmonary fibrosis, *N. Engl. J. Med.*, vol. 364: pp. 1503–12, 2011.

36. G. M. Hunninghake, H. Hatabu, Y. Okajima, W. Gao, J. Dupuis, J. C. Latourelle, M. Nishino, T. Araki, O. E. Zazueta, S. Kurugol, J. C. Ross, R. San José Estépar, E. Murphy, M. P. Steele, J. E. Loyd, M. I. Schwarz, T. E. Fingerlin, I. O. Rosas, G. R. Washko, G. T. O'Connor, and D. a Schwartz, MUC5B promoter polymorphism and interstitial lung abnormalities, *N. Engl. J. Med.*, vol. 368: pp. 2192–200, 2013.

37. R. Wei, C. Li, M. Zhang, Y. L. Jones-Hall, J. L. Myers, I. Noth, and W. Liu, Association between MUC5B and TERT polymorphisms and different interstitial lung disease phenotypes, *Transl. Res.*, vol. 163:pp. 494-502, 2013.

38. T. E. Fingerlin, E. Murphy, W. Zhang, A. L. Peljto, K. K. Brown, M. P. Steele, J. E. Loyd, G. P. Cosgrove, D. Lynch, S. Groshong, H. R. Collard, P. J. Wolters, W. Z. Bradford, K. Kossen, S. D. Seiwert, R. M. du Bois, C. K. Garcia, M. S. Devine, G. Gudmundsson, H. J. Isaksson, N. Kaminski, Y. Zhang, K. F. Gibson, L. H. Lancaster, J. D. Cogan, W. R. Mason, T. M. Maher, P. L. Molyneaux, A. U. Wells, M. F. Moffatt, M. Selman, A. Pardo, D. S. Kim, J. D.

Crapo, B. J. Make, E. a Regan, D. S. Walek, J. J. Daniel, Y. Kamatani, D. Zelenika, K. Smith, D. McKean, B. S. Pedersen, J. Talbert, R. N. Kidd, C. R. Markin, K. B. Beckman, M. Lathrop, M. I. Schwarz, and D. a Schwartz, Genome-wide association study identifies multiple susceptibility loci for pulmonary fibrosis, *Nat. Genet.*, vol. 45: pp. 613–20, 2013.

39. T. Mushiroda, S. Wattanapokayakit, a Takahashi, T. Nukiwa, S. Kudoh, T. Ogura, H. Taniguchi, M. Kubo, N. Kamatani, and Y. Nakamura, A genome-wide association study identifies an association of a common variant in TERT with susceptibility to idiopathic pulmonary fibrosis, *J. Med. Genet.*, vol. 45: pp. 654–6, 2008.

40. L. M. Nogee, A. E. Dunbar, S. E. Wert, F. Askin, A. Hamvas, and J. A. Whitsett, A Mutation in the Surfactant Protein C Gene Associated with Familial Interstitial Lung Disease, *N. Engl. J. Med.*, vol. 344, no. 8, pp. 573–9, 2001. .

41. L. K. a Lundblad, J. Thompson-Figueroa, T. Leclair, M. J. Sullivan, M. E. Poynter, C. G. Irvin, and J. H. T. Bates, Tumor necrosis factor-alpha overexpression in lung disease: a single cause behind a complex phenotype, *Am. J. Respir. Crit. Care Med.*, vol. 171: pp. 1363–70, 2005.

42. D. P. Ascherman, Interstitial lung disease in rheumatoid arthritis, *Curr. Rheumatol. Rep.*, vol. 12: pp. 363–9, 2010.

43. C. Agostini and C. Gurrieri, Chemokine/cytokine cocktail in idiopathic pulmonary fibrosis,

Proc. Am. Thorac. Soc., vol. 3: pp. 357–63, 2006.

44. M. Vasakova, I. Striz, a Slavcev, S. Jandova, J. Dutka, M. Terl, L. Kolesar, and J. Sulc, Correlation of IL-1alpha and IL-4 gene polymorphisms and clinical parameters in idiopathic pulmonary fibrosis, *Scand. J. Immunol.*, vol. 65: pp. 265–70, 2007.

45. T. R. Dunkern, D. Feurstein, G. a Rossi, F. Sabatini, and A. Hatzelmann, Inhibition of TGF-beta induced lung fibroblast to myofibroblast conversion by phosphodiesterase inhibiting drugs and activators of soluble guanylyl cyclase, *Eur. J. Pharmacol.*, vol. 572: pp. 12–22, 2007.

46. J. M. Vignaud, M. Allam, N. Martinet, M. Pech, F. Plenat, and Y. Martinet, Presence of platelet-derived growth factor in normal and fibrotic lung is specifically associated with interstitial macrophages, while both interstitial macrophages and alveolar epithelial cells express the c-sis proto-oncogene, *Am. J. Respir. Cell Mol. Biol.*, vol. 5: pp. 531–8, 1991.

47. A. Abdollahi, M. Li, G. Ping, C. Plathow, S. Domhan, F. Kiessling, L. B. Lee, G. McMahon, H.-J. Gröne, K. E. Lipson, and P. E. Huber, Inhibition of platelet-derived growth factor signaling attenuates pulmonary fibrosis, *J. Exp. Med.*, vol. 201: pp. 925–35, 2005.

48. V. Saravanan and C. Kelly, Drug-related pulmonary problems in patients with rheumatoid arthritis, *Rheumatology (Oxford)*, vol. 45: pp. 787–9, 2006.

49. H. Furukawa, S. Oka, K. Shimada, N. Tsuchiya, and S. Tohma, HLA-A*31:01 and

methotrexate-induced interstitial lung disease in Japanese rheumatoid arthritis patients: a multidrug hypersensitivity marker?, *Ann. Rheum. Dis.*, vol. 72: pp. 153–5, 2013.

50. B. Chikura, S. Lane, and J. K. Dawson, Clinical expression of leflunomide-induced pneumonitis, *Rheumatology (Oxford)*, vol. 48: pp. 1065–8, 2009.

51. T. Sawada, S. Inokuma, T. Sato, T. Otsuka, Y. Saeki, T. Takeuchi, T. Matsuda, T. Takemura, and A. Sagawa, Leflunomide-induced interstitial lung disease: prevalence and risk factors in Japanese patients with rheumatoid arthritis, *Rheumatology (Oxford)*, vol. 48: pp. 1069–72, 2009.

52. D. Welter, J. MacArthur, J. Morales, T. Burdett, P. Hall, H. Junkins, A. Klemm, P. Flicek, T. Manolio, L. Hindorff, and H. Parkinson, The NHGRI GWAS Catalog, a curated resource of SNP-trait associations, *Nucleic Acids Res.*, vol. 42: pp. D1001–6, 2014.

53. Y. Okada, D. Wu, G. Trynka, T. Raj, C. Terao, K. Ikari, Y. Kochi, K. Ohmura, A. Suzuki, S. Yoshida, R. R. Graham, A. Manoharan, W. Ortmann, T. Bhangale, J. C. Denny, R. J. Carroll, A. E. Eyler, J. D. Greenberg, J. M. Kremer, D. a. Pappas, L. Jiang, J. Yin, L. Ye, D.-F. Su, J. Yang, G. Xie, E. Keystone, H.-J. Westra, T. Esko, A. Metspalu, X. Zhou, N. Gupta, D. Mirel, E. a. Stahl, D. Diogo, J. Cui, K. Liao, M. H. Guo, K. Myouzen, T. Kawaguchi, M. J. H. Coenen, P. L. C. M. van Riel, M. a. F. J. van de Laar, H. J. Guchelaar, T. W. J. Huizinga, P. Dieudé, X. Mariette, S. Louis Bridges Jr, A. Zhernakova, R. E. M. Toes, P. P. Tak, C.

Miceli-Richard, S. Y. Bang, H. S. Lee, J. Martin, M. a. Gonzalez-Gay, L. Rodriguez-Rodriguez, S. Rantapää-Dahlqvist, L. Ärlestig, H. K. Choi, Y. Kamatani, P. Galan, M. Lathrop, S. Eyre, J. Bowes, A. Barton, N. de Vries, L. W. Moreland, L. a. Criswell, E. W. Karlson, A. Taniguchi, R. Yamada, M. Kubo, J. S. Liu, S. C. Bae, J. Worthington, L. Padyukov, L. Klareskog, P. K. Gregersen, S. Raychaudhuri, B. E. Stranger, P. L. De Jager, L. Franke, P. M. Visscher, M. a. Brown, H. Yamanaka, T. Mimori, A. Takahashi, H. Xu, T. W. Behrens, K. a. Siminovitch, S. Momohara, F. Matsuda, K. Yamamoto, and R. M. Plenge, Genetics of rheumatoid arthritis contributes to biology and drug discovery, *Nature*, vol. 506: pp. 376–81, 2013.

54. F. C. Arnett, S. M. Edworthy, D. a Bloch, D. J. McShane, J. F. Fries, N. S. Cooper, L. a Healey, S. R. Kaplan, M. H. Liang, and H. S. Luthra, The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis, *Arthritis Rheum.*, vol. 31: pp. 315–24, 1988.

55. H. Kameda, H. Tokuda, F. Sakai, T. Johkoh, S. Mori, Y. Yoshida, N. Takayanagi, H. Taki, Y. Hasegawa, K. Hatta, H. Yamanaka, M. Dohi, S. Hashimoto, H. Yamada, S. Kawai, T. Takeuchi, K. Tateda, and H. Goto, Clinical and radiological features of acute-onset diffuse interstitial lung diseases in patients with rheumatoid arthritis receiving treatment with biological agents: importance of pneumocystis pneumonia in Japan revealed by a

multicenter study, *Intern. Med.*, vol. 50: pp. 305–13, 2011.

56. T. J. Hoffmann, M. N. Kvale, S. E. Hesselson, Y. Zhan, C. Aquino, Y. Cao, S. Cawley, E. Chung, S. Connell, J. Eshragh, M. Ewing, J. Gollub, M. Henderson, E. Hubbell, C. Iribarren, J. Kaufman, R. Z. Lao, Y. Lu, D. Ludwig, G. K. Mathauda, W. McGuire, G. Mei, S. Miles, M. M. Purdy, C. Quesenberry, D. Ranatunga, S. Rowell, M. Sadler, M. H. Shapero, L. Shen, T. R. Shenoy, D. Smethurst, S. K. Van den Eeden, L. Walter, E. Wan, R. Wearley, T. Webster, C. C. Wen, L. Weng, R. A. Whitmer, A. Williams, S. C. Wong, C. Zau, A. Finn, C. Schaefer, P.-Y. Kwok, and N. Risch, Next generation genome-wide association tool: design and coverage of a high-throughput European-optimized SNP array, *Genomics*, vol. 98： pp. 79–89, 2011.

57. S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, and P. C. Sham, PLINK: a tool set for whole-genome association and population-based linkage analyses, *Am. J. Hum. Genet.*, vol. 81: pp. 559–75, 2007.

58. N. Patterson, A. L. Price, and D. Reich, Population structure and eigenanalysis, *PLoS Genet.*, vol. 2: p. e190, 2006.

59. A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich, Principal components analysis corrects for stratification in genome-wide association

studies, *Nat Genet*, vol. 38: pp. 904–9, 2006.

60. ftp://ftp.ncbi.nlm.nih.gov/hapmap/genotypes/2009-01_phaseIII/plink_format/

61. P. A. Fujita, B. Rhead, A. S. Zweig, A. S. Hinrichs, D. Karolchik, M. S. Cline, M. Goldman, G. P. Barber, H. Clawson, A. Coelho, M. Diekhans, T. R. Dreszer, B. M. Giardine, R. A. Harte, J. Hillman-Jackson, F. Hsu, V. Kirkup, R. M. Kuhn, K. Learned, C. H. Li, L. R. Meyer, A. Pohl, B. J. Raney, K. R. Rosenbloom, K. E. Smith, D. Haussler, and W. J. Kent, The UCSC Genome Browser database: update 2011, *Nucleic Acids Res.*, vol. 39: pp. D876–82, 2011.

62. Y. H. Fan and Y. Q. Song, IPGWAS: an integrated pipeline for rational quality control and association analysis of genome-wide genetic studies, *Biochem. Biophys. Res. Commun.*, vol. 422: pp. 363–8, 2012.

63. J. C. Barrett, B. Fry, J. Maller, and M. J. Daly, Haploview: analysis and visualization of LD and haplotype maps, *Bioinformatics*, vol. 21: pp. 263–5, 2005.

64. R. J. Pruim, R. P. Welch, S. Sanna, T. M. Teslovich, P. S. Chines, T. P. Gliedt, M. Boehnke, G. R. Abecasis, and C. J. Willer, LocusZoom: regional visualization of genome-wide association scan results, *Bioinformatics*, vol. 26: pp. 2336–7, 2010.

65. B. N. Howie, P. Donnelly, and J. Marchini, A flexible and accurate genotype imputation method for the next generation of genome-wide association studies, *PLoS Genet.*, vol. 5, no.

6, p. e1000529, 2009.

66. J. Marchini and B. Howie, Genotype imputation for genome-wide association studies, Nat. Rev. Genet., vol. 11: pp. 499–511, 2010.

67. http://www.well.ox.ac.uk/~cfreeman/software/gwas/gtool.html (cited 2014-04-20)

68. T. P. Yang, C. Beazley, S. B. Montgomery, A. S. Dimas, M. Gutierrez-Arcelus, B. E. Stranger, P. Deloukas, and E. T. Dermitzakis, Genevar: a database and Java application for the analysis and visualization of SNP-gene associations in eQTL studies, Bioinformatics, vol. 26: pp. 2474–6, 2010.

69. A. C. Nica, L. Parts, D. Glass, J. Nisbet, A. Barrett, M. Sekowska, M. Travers, S. Potter, E. Grundberg, K. Small, A. K. Hedman, V. Bataille, J. Tzenova Bell, G. Surdulescu, A. S. Dimas, C. Ingle, F. O. Nestle, P. di Meglio, J. L. Min, A. Wilk, C. J. Hammond, N. Hassanali, T.-P. Yang, S. B. Montgomery, S. O'Rahilly, C. M. Lindgren, K. T. Zondervan, N. Soranzo, I. Barroso, R. Durbin, K. Ahmadi, P. Deloukas, M. I. McCarthy, E. T. Dermitzakis, and T. D. Spector, The architecture of gene regulatory variation across multiple human tissues: the MuTHER study, *PLoS Genet.*, vol. 7: p. e1002003, 2011.

70. B. E. Stranger, S. B. Montgomery, A. S. Dimas, L. Parts, O. Stegle, C. E. Ingle, M. Sekowska, G. D. Smith, D. Evans, M. Gutierrez-Arcelus, A. Price, T. Raj, J. Nisbett, A. C. Nica, C. Beazley, R. Durbin, P. Deloukas, and E. T. Dermitzakis, Patterns of cis regulatory

variation in diverse human populations, *PLoS Genet.*, vol. 8: p. e1002639, 2012.

71. E. Grundberg, K. S. Small, Å. K. Hedman, A. C. Nica, A. Buil, S. Keildson, J. T. Bell, T.-P. Yang, E. Meduri, A. Barrett, J. Nisbett, M. Sekowska, A. Wilk, S.-Y. Shin, D. Glass, M. Travers, J. L. Min, S. Ring, K. Ho, G. Thorleifsson, A. Kong, U. Thorsteindottir, C. Ainali, A. S. Dimas, N. Hassanali, C. Ingle, D. Knowles, M. Krestyaninova, C. E. Lowe, P. Di Meglio, S. B. Montgomery, L. Parts, S. Potter, G. Surdulescu, L. Tsaprouni, S. Tsoka, V. Bataille, R. Durbin, F. O. Nestle, S. O'Rahilly, N. Soranzo, C. M. Lindgren, K. T. Zondervan, K. R. Ahmadi, E. E. Schadt, K. Stefansson, G. D. Smith, M. I. McCarthy, P. Deloukas, E. T. Dermitzakis, and T. D. Spector, Mapping cis- and trans-regulatory effects across multiple tissues in twins, *Nat. Genet.*, vol. 44: pp. 1084–9, 2012.

72. A. S. Dimas, S. Deutsch, B. E. Stranger, S. B. Montgomery, C. Borel, H. Attar-Cohen, C. Ingle, C. Beazley, M. Gutierrez Arcelus, M. Sekowska, M. Gagnebin, J. Nisbett, P. Deloukas, E. T. Dermitzakis, and S. E. Antonarakis, Common regulatory variation impacts gene expression in a cell type-dependent manner, *Science*, vol. 325: pp. 1246–50, 2009.

73. L. D. Ward and M. Kellis, HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants, *Nucleic Acids Res.*, vol. 40: pp. D930–4, 2012.

74. R Core Team, R: A language and environment for statistical computing, 2014.

75. R. Horton, L. Wilming, V. Rand, R. C. Lovering, E. a Bruford, V. K. Khodiyar, M. J. Lush, S. Povey, C. C. Talbot, M. W. Wright, H. M. Wain, J. Trowsdale, A. Ziegler, and S. Beck, Gene map of the extended human MHC, *Nat. Rev. Genet.*, vol. 5: pp. 889–99, 2004.

76. D. Melzer, J. R. B. Perry, D. Hernandez, A.-M. Corsi, K. Stevens, I. Rafferty, F. Lauretani, A. Murray, J. R. Gibbs, G. Paolisso, S. Rafiq, J. Simon-Sanchez, H. Lango, S. Scholz, M. N. Weedon, S. Arepalli, N. Rice, N. Washecka, A. Hurst, A. Britton, W. Henley, J. van de Leemput, R. Li, A. B. Newman, G. Tranah, T. Harris, V. Panicker, C. Dayan, A. Bennett, M. I. McCarthy, A. Ruokonen, M.-R. Jarvelin, J. Guralnik, S. Bandinelli, T. M. Frayling, A. Singleton, and L. Ferrucci, A genome-wide association study identifies protein quantitative trait loci (pQTLs), *PLoS Genet.*, vol. 4: p. e1000072, 2008.

77. R. Janz, K. Hofmann, and T. C. Su, SVOP, an evolutionarily conserved synaptic vesicle protein, suggests novel transport functions of synaptic vesicles, vol. 18: pp. 9269–81, 1998.

78. E. Y. Cho, C. J. Lee, K. S. Son, Y. J. Kim, and S. J. Kim, Characterization of mouse synaptic vesicle-2-associated protein (Msvop) specifically expressed in the mouse central nervous system, *Gene*, vol. 429:pp. 44–8, 2009.

79. H. Y. Wang, C. Y. Lin, C. C. Chien, W. C. Kan, Y. F. Tian, P. C. Liao, H. Y. Wu, and S. B. Su, Impact of uremic environment on peritoneum: a proteomic view, *J. Proteomics*, vol. 75: pp. 2053–63, 2012.

80. A. Rovelet-Lecrux, S. Legallic, D. Wallon, J.-M. Flaman, O. Martinaud, S. Bombois, A. Rollin-Sillaire, A. Michon, I. Le Ber, J. Pariente, M. Puel, C. Paquet, B. Croisile, C. Thomas-Antérion, M. Vercelletto, R. Lévy, T. Frébourg, D. Hannequin, and D. Campion, A genome-wide study reveals rare CNVs exclusive to extreme phenotypes of Alzheimer disease, *Eur. J. Hum. Genet.*, vol. 20: pp. 613–7, 2012.

81. J. A. Jacobsson, T. Haitina, J. Lindblom, and R. Fredriksson, Identification of six putative human transporters with structural similarity to the drug transporter SLC22 family, *Genomics*, vol. 90: pp. 595–609, 2007.

82. A. Visel, C. Thaller, and G. Eichele, GenePaint.org: an atlas of gene expression patterns in the mouse embryo, *Nucleic Acids Res.*, vol. 32: pp. D552–6, 2004.

83. http://www.ncbi.nlm.nih.gov/geo/tools/profileGraph.cgi?ID=GDS4279:229818_at (cited 2014-08-21)

84. D. R. Repaskes, J. V Swinnen, S. C. Jin, J. Van Wyk, and M. Conti, A Polymerase Chain Reaction Strategy to Identify and Clone Cyclic Nucleotide Phosphodiesterase cDNAs, *J. Biol. Chem.*, vol. 267: pp. 18683-8, 1992.

85. http://www.ncbi.nlm.nih.gov/geo/tools/profileGraph.cgi?ID=GDS4279:236344_at (cited 2014-08-21)

86. RefExA (http://157.82.78.238/refexa/main_search.jsp) (cited 2014-04-20)

87. R. Rajkumar, K. Konishi, T. J. Richards, D. C. Ishizawar, A. C. Wiechert, N. Kaminski, and F. Ahmad, Genomewide RNA expression profiling in lung identifies distinct signatures in idiopathic pulmonary arterial hypertension and secondary pulmonary hypertension., *Am. J. Physiol. Heart Circ. Physiol.*, vol. 298, pp. H1235–48, 2010.

88. Y. Inoue, T. E. King, E. Barker, E. Daniloff, and L. S. Newman, Basic fibroblast growth factor and its receptors in idiopathic pulmonary fibrosis and lymphangioleiomyomatosis, *Am. J. Respir. Crit. Care Med.*, vol. 166: pp. 765–73, 2002.

89. D. Jaraquemada, W. Ollier, and J. Awad, HLA and rheumatoid arthritis: a combined analysis of 440 British patients, *Ann. Rheum. Dis.*, vol. 45: pp. 627–36, 1986.

90. Y. C. Wang, J. Y. Chen, M. L. Chen, C. H. Chen, I. C. Lai, T. T. Chen, C. J. Hong, S. J. Tsai, and Y. J. Liou, Neuregulin 3 genetic variations and susceptibility to schizophrenia in a Chinese population, *Biol. Psychiatry*, vol. 64: pp. 1093–6, 2008.

91. S. Boer, M. Berk, and B. Dean, Levels of neuregulin 1 and 3 proteins in Brodmann's area 46 from subjects with schizophrenia and bipolar disorder, *Neurosci. Lett.*, vol. 466: pp. 27–9, 2009.

92. D. Dreymueller, C. Martin, J. Schumacher, E. Groth, J. K. Boehm, L. K. Reiss, S. Uhlig, and A. Ludwig, Smooth muscle cells relay acute pulmonary inflammation via distinct ADAM17/ErbB axes, *J. Immunol.*, vol. 192: pp. 722–31, 2014.

93. K. Zscheppang, T. Dörk, A. Schmiedl, F. E. Jones, and C. E. L. Dammann, Neuregulin receptor ErbB4 functions as a transcriptional cofactor for the expression of surfactant protein B in the fetal lung, *Am. J. Respir. Cell Mol. Biol.*, vol. 45: pp. 761–7, 2011.

Tables

Table 1 Male sex and older age showed the significant association with RA with ILD compared to RA without ILD.

		RA with ILD	RA without ILD	*p* -value
Sex				
	Male/female n (%)	60/109 (35.5/64.5)	44/250 (15.0/85.0)	4.21×10^{-7}
Age				
	Average y.o. (SD)	69.9 (8.9)	61.8 (11.0)	4.22×10^{-15}

The association of male sex was tested by chi-square test. Average age was calculated by Wilcoxon rank sum test.
Male sex and age indicated the association of RA-ILD with *p*-value of 4.21×10^{-7} and 4.22×10^{-15}, respectively.

Table 2 List of SNPs with *p*-value of < 5.0×10^{-6} from the result of association tests between RA with ILD and RA without ILD

Chromosomal region	SNP name	Allele	MAF		Allelic		Dominant		Recessive		Genotypic	Trend	location
			w ILD	w/o ILD	*p*-value	OR (95%CI)	*p*-value	OR (95%CI)	*p*-value	OR (95%CI)	*p*-value	*p*-value	
19q13.11	marker_A1	AC	0.27	0.14	8.34×10^{-7}	2.29 (1.64-3.19)	1.33×10^{-6}	2.64 (1.78-3.93)	1.45×10^{-2}	3.60 (1.21-10.71)	3.70×10^{-6}	**5.98×10^{-7}**	intergenic
12q24.11	marker_B	AG	0.30	0.46	2.49×10^{-6}	0.51 (0.39-0.68)	**6.86×10^{-7}**	0.37 (0.25-0.55)	1.48×10^{-2}	0.50 (0.29-0.88)	3.03×10^{-6}	2.30×10^{-6}	intragenic
12q23.3	marker_C	TC	0.32	0.49	**1.26×10^{-6}**	0.51 (0.38-0.67)	1.50×10^{-4}	0.47 (0.32-0.70)	6.25×10^{-5}	0.33 (0.19-0.58)	1.36×10^{-5}	2.44×10^{-6}	intergenic
22q13.1	marker_D1	AG	0.51	0.38	1.46×10^{-4}	1.69 (1.29-2.22)	8.62×10^{-2}	1.45 (0.95-2.19)	**1.59×10^{-6}**	3.17 (1.95-5.14)	9.82×10^{-6}	1.33×10^{-4}	intergenic
6p21.32	marker_E	CT	0.49	0.33	**1.70×10^{-6}**	1.96 (1.49-2.57)	2.83×10^{-4}	2.13 (1.41-3.21)	2.96×10^{-5}	2.88 (1.73-4.79)	1.07×10^{-5}	2.71×10^{-6}	intergenic
19q13.11	marker_A2	TC	0.23	0.11	2.32×10^{-6}	2.35 (1.64-3.37)	6.63×10^{-6}	2.63 (1.74-3.97)	4.24×10^{-2}	4.15 (1.06-16.27)	8.34×10^{-6}	**1.81×10^{-6}**	intergenic
	marker_A3	AG	0.30	0.17	4.71×10^{-6}	2.08 (1.52-2.84)	6.91×10^{-5}	2.19 (1.49-3.22)	2.75×10^{-4}	5.60 (2.00-15.68)	9.63×10^{-6}	**3.17×10^{-6}**	intergenic
19q13.12	marker_F	GA	0.57	0.42	2.84×10^{-5}	1.78 (1.36-2.34)	3.09×10^{-2}	1.63 (1.05-2.52)	**3.30×10^{-6}**	2.78 (1.79-4.30)	1.72×10^{-5}	4.05×10^{-5}	intergenic
22q13.1	marker_D2	AG	0.51	0.38	1.48×10^{-4}	1.69 (1.29-2.22)	6.54×10^{-2}	1.49 (0.98-2.26)	**3.56×10^{-6}**	3.03 (1.88-4.90)	2.02×10^{-5}	1.32×10^{-4}	intergenic
5q35.2	marker_G	CT	0.23	0.38	**4.09×10^{-6}**	0.50 (0.37-0.67)	4.13×10^{-4}	0.51 (0.34-0.74)	4.90×10^{-5}	0.20 (0.09-0.47)	1.83×10^{-5}	5.95×10^{-6}	intergenic
7p14.3	marker_H	CT	0.13	0.26	**4.38×10^{-6}**	0.44 (0.31-0.63)	3.88×10^{-5}	0.42 (0.28-0.64)	3.62×10^{-3}	0.21 (0.07-0.69)	3.44×10^{-5}	9.17×10^{-6}	intragenic

MAF: minor allele frequency, w ILD: RA with ILD, w/o ILD: RA without ILD, OR: odds ratio and 95%CI: 95% confidence interval. Bold font highlights the most significant *p*-value under five models; allelic, dominant, recessive, genotypic model and Cochran-Armitage trend test. Two regions showed strong association, with minimum *p*-values of less than 1×10^{-6}.

Table 3 List of flanking SNPs of the SNP marker_A1 with functional annotations provided by HaploReg v.2

Variants	LD (r^2)	Allele	ASN MAF	Conserved	number of cell types		
					Promoter histone mark	Enhancer histone mark	DNase Hypersensitive Site
marker_a1	0.88	T/C	0.22				
marker_a2	0.98	C/T	0.22				4
marker_A1	1	C/A	0.22				
marker_a3	1	A/C	0.22				
marker_a4	1	A/T	0.22				
marker_a5	1	C/T	0.22				
marker_a6	0.92	A/T	0.2				
marker_a7	0.92	A/T	0.2				
marker_a8	0.92	GAA/G	0.2				
marker_a9	0.92	G/A	0.2				
marker_a10	0.9	C/T	0.2				
marker_a11	0.86	G/A	0.21				5
marker_a12	0.88	A/C	0.2				
marker_a13	0.89	A/C	0.2				7
marker_a14	0.82	G/C	0.19	G, S	4		63
marker_a15	0.84	G/A	0.2				
marker_a16	0.83	C/T	0.2				2
marker_a17	0.83	T/G	0.2				2
marker_a18	0.81	T/C	0.2	S			1
marker_a19	0.82	A/AGG	0.2				
marker_a20	0.82	G/A	0.2				

LD (r^2): r-squared value between the SNP marker_A1 and other variants and ASN MAF: minor allele frequency in Asian population of 1000 genomes project.

If a SNP was detected as a conserved element across mammals, G and/or S were displayed in the 5th column by which to use data. (G: Genomic Evolutionary Rate Profiling and S: SiPhy-omega)

Table 4 Results of association test of *HLA-DRB1*04* alleles

Alleles	RA with ILD (%)	RA without ILD (%)	OR (95%CI)	*P* -value	Corrected *p* -value
*HLA-DRB1*01*	36 (10.7)	43 (7.3)	1.51 (0.95-2.4)	2.12×10^{-2}	NS
*HLA-DRB1*04*	98 (29.0)	248 (42.2)	0.56 (0.42-0.75)	1.76×10^{-5}	2.11×10^{-4}
*HLA-DRB1*07*	1 (0.3)	2 (0.3)	0.87 (0.08-9.62)	4.42×10^{-1}	NS
*HLA-DRB1*08*	10 (3.0)	34 (5.8)	0.50 (0.24-1.02)	1.91×10^{-2}	NS
*HLA-DRB1*09*	50 (14.8)	94 (16.0)	0.91 (0.63-1.32)	6.73×10^{-2}	NS
*HLA-DRB1*10*	4 (1.2)	3 (0.5)	2.34 (0.52-10.5)	1.59×10^{-1}	NS
*HLA-DRB1*11*	2 (0.6)	8 (1.4)	0.43 (0.09-2.04)	1.58×10^{-1}	NS
*HLA-DRB1*12*	17 (5.0)	23 (3.9)	1.30 (0.68-2.47)	9.44×10^{-2}	NS
*HLA-DRB1*13*	16 (4.7)	18 (3.1)	1.57 (0.79-3.13)	6.10×10^{-2}	NS
*HLA-DRB1*14*	22 (6.5)	27 (4.6)	1.45 (0.81-2.58)	5.45×10^{-2}	NS
*HLA-DRB1*15*	78 (23.1)	87 (14.8)	1.73 (1.23-2.43)	5.17×10^{-4}	6.20×10^{-3}
*HLA-DRB1*16*	4 (1.2)	1 (0.2)	7.03 (0.78-63.16)	5.60×10^{-2}	NS

OR: odds ratio, 95%CI: 95% confidence interval and NS: not significant. *P*-values of HLA alleles were calculated by Fisher's exact test. Corrected *p*-value was lead after multiplying the number of *HLA-DRB1* alleles. If corrected *p*-value was over 0.05, it was denoted "NS". *HLA-DRB1*04* and *HLA-DRB1*15* showed the significant association after the correction for multiple testing.

Table 5 List of SNPs with *p*-value of < 5.0×10^{-6} from the result of association tests for *HLA-DRB1*04* positive patients

Chromosomal region	SNP name	Allele	MAF		Allelic		Dominant		Recessive		Genotypic	Trend	location
			w ILD	w/o ILD	*p*-value	OR (95%CI)	*p*-value	OR (95%CI)	*p*-value	OR (95%CI)	*p*-value	*p*-value	
7p21.3	marker_I1	GT	0.41	0.2	**4.20×10^{-7}**	2.71 (1.83-4.00)	1.02×10^{-6}	3.76 (2.18-6.49)	4.77×10^{-3}	3.47 (1.40-8.57)	2.43×10^{-6}	4.59×10^{-7}	intergenic
	marker_I2	GA	0.41	0.21	**6.01×10^{-7}**	2.65 (1.80-3.91)	$2.17 \times 1^{0-6}$	3.58 (2.09-6.13)	4.02×10^{-3}	3.38 (1.42-8.05)	4.71×10^{-6}	8.79×10^{-7}	intergenic
	marker_I3	CT	0.41	0.21	**8.26×10^{-7}**	2.63 (1.78-3.88)	2.17×10^{-6}	3.58 (2.09-6.13)	5.07×10^{-3}	3.44 (1.39-8.50)	4.78×10^{-6}	8.73×10^{-7}	intergenic
	marker_I4	CG	0.41	0.21	**1.22×10^{-6}**	2.60 (1.76-3.84)	3.73×10^{-6}	3.49 (2.03-5.99)	4.77×10^{-3}	3.47 (1.40-8.57)	7.45×10^{-6}	1.32×10^{-6}	intergenic
	marker_I5	CT	0.41	0.21	**1.43×10^{-6}**	2.57 (1.74-3.79)	2.57×10^{-6}	3.55 (2.07-6.08)	9.76×10^{-3}	3.06 (1.27-7.39)	7.81×10^{-6}	1.73×10^{-6}	intergenic
	marker_I6	CT	0.41	0.21	**1.47×10^{-6}**	2.58 (1.75-3.81)	4.65×10^{-6}	3.44 (2.00-5.90)	4.57×10^{-3}	3.49 (1.41-8.62)	8.73×10^{-6}	1.52×10^{-6}	intergenic
	marker_I7	GT	0.41	0.21	**1.47×10^{-6}**	2.58 (1.75-3.81)	3.16×10^{-6}	3.51 (2.05-6.03)	8.52×10^{-3}	3.12 (1.29-7.54)	8.93×10^{-6}	1.88×10^{-6}	intergenic
	marker_I8	TA	0.41	0.22	**1.78×10^{-6}**	2.54 (1.72-3.73)	4.74×10^{-6}	3.43 (2.00-5.87)	7.33×10^{-3}	3.06 (1.31-7.13)	1.25×10^{-5}	2.60×10^{-6}	intergenic
	marker_I9	GA	0.41	0.21	**2.09×10^{-6}**	2.54 (1.72-3.76)	4.39×10^{-6}	3.46 (2.01-5.94)	9.23×10^{-3}	3.09 (1.28-7.46)	1.23×10^{-5}	2.58×10^{-6}	intergenic
7q36.1	marker_J1	CT	0.27	0.48	**4.40×10^{-6}**	0.41 (0.28-0.61)	6.09×10^{-5}	0.35 (0.21-0.59)	5.86×10^{-4}	0.21 (0.09-0.55)	2.62×10^{-5}	4.42×10^{-6}	intragenic
	marker_J2	G-	0.27	0.48	**4.51×10^{-6}**	0.41 (0.28-0.61)	4.59×10^{-5}	0.34 (0.20-0.58)	7.31×10^{-4}	0.22 (0.09-0.57)	2.34×10^{-5}	4.00×10^{-6}	intragenic

MAF: minor allele frequency, w ILD: RA with ILD, w/o ILD: RA without ILD, OR: odds ratio and 95%CI: 95% confidence interval. Bold font highlights the most significant *p*-value under five models; allelic, dominant, recessive, genotypic model and Cochran-Armitage trend test.

Table 6 List of SNPs with *p*-value of < 5.0×10^{-6} from the result of association tests for *HLA-DRB1*04* negative patients

Chromosomal region	SNP name	Allele	MAF		Allelic		Dominant		Recessive		Genotypic	Trend	location
			w ILD	w/o ILD	*p*-value	OR (95%CI)	*p*-value	OR (95%CI)	*p*-value	OR (95%CI)	*p*-value	*p*-value	
10q23.1	marker_K1	CG	0.19	0.04	**2.72×10^{-6}**	5.77 (2.59-12.89)	1.26×10^{-5}	5.98 (2.55-14.02)	4.35×10^{-2}	11.04 (0.59-208.12)	1.61×10^{-5}	7.13×10^{-6}	intergenic
	marker_K2	TC	0.19	0.04	**2.72×10^{-6}**	5.77 (2.59-12.89)	1.26×10^{-5}	5.98 (2.55-14.02)	4.35×10^{-2}	11.04 (0.59-208.12)	1.61×10^{-5}	7.13×10^{-6}	intergenic
	marker_K3	TA	0.19	0.04	**3.38×10^{-6}**	5.69 (2.55-12.70)	1.39×10^{-5}	5.87 (2.51-13.75)	4.46×10^{-2}	10.91 (0.58-205.51)	1.91×10^{-5}	8.75×10^{-6}	intergenic
	marker_K4	AG	0.19	0.04	**3.38×10^{-6}**	5.69 (2.55-12.70)	1.39×10^{-5}	5.87 (2.51-13.75)	4.46×10^{-2}	10.91 (0.58-205.51)	1.91×10^{-5}	8.75×10^{-6}	intergenic
	marker_K5	GC	0.19	0.04	**3.38×10^{-6}**	5.69 (2.55-12.70)	1.39×10^{-5}	5.87 (2.51-13.75)	4.46×10^{-2}	10.91 (0.58-205.51)	1.91×10^{-5}	8.75×10^{-6}	intergenic
	marker_K6	AG	0.19	0.04	**4.00×10^{-6}**	5.63 (2.52-12.56)	1.50×10^{-5}	5.80 (2.48-13.60)	4.56×10^{-2}	10.79 (0.58-203.42)	1.91×10^{-5}	1.02×10^{-5}	intergenic
	marker_K7	AG	0.19	0.04	**4.00×10^{-6}**	5.63 (2.52-12.56)	1.50×10^{-5}	5.80 (2.48-13.60)	4.56×10^{-2}	10.79 (0.58-203.42)	1.91×10^{-5}	1.02×10^{-5}	intergenic
3p22.1	marker_L	CT	0.57	0.39	4.22×10^{-4}	2.13 (1.40-3.23)	**4.05×10^{-6}**	5.85 (2.63-13.03)	2.90×10^{-1}	1.46 (0.73-2.93)	2.17×10^{-5}	4.01×10^{-4}	intergenic
10q23.1	marker_K8	GA	0.20	0.04	**4.74×10^{-6}**	5.56 (2.49-12.42)	1.68×10^{-5}	5.73 (2.44-13.45)	3.15×10^{-2}	10.68 (0.57-201.32)	1.80×10^{-4}	1.18×10^{-5}	intergenic

MAF: minor allele frequency, w ILD: RA with ILD, w/o ILD: RA without ILD, OR: odds ratio and 95%CI: 95% confidence interval.
Bold font highlights the most significant *p*-value under five models: allelic, dominant, recessive, genotypic model and Cochran-Armitage trend test.

Table 7 Validation analysis of the SNPs from GWAS by TaqMan assay showed consistent results of associations

Chromosomal region	SNP name	Allele	MAF		Allelic		Dominant		Recessive		Genotypic	Trend
			w ILD	w/o ILD	*p*-value	OR (95%CI)	*p*-value	OR (95%CI)	*p*-value	OR (95%CI)	*p*-value	*p*-value
6p21.32	marker_E	CT	0.49	0.34	**5.47×10^{-6}**	1.88 (1.43-2.47)	2.77×10^{-4}	2.13 (1.41-3.22)	2.18×10^{-4}	2.5 (1.52-4.09)	4.70×10^{-5}	8.90×10^{-6}
7p21.3	marker_I6	GT	0.40	0.21	**2.93×10^{-6}**	2.53 (1.7-3.75)	1.19×10^{-5}	3.26 (1.9-5.59)	4.68×10^{-3}	3.47 (1.4-8.6)	2.06×10^{-5}	3.58×10^{-6}
10q23.1	marker_K1	CG	0.20	0.04	**4.74×10^{-6}**	5.56 (2.49-12.42)	1.68×10^{-5}	5.73 (2.44-13.45)	3.15×10^{-2}	10.68 (0.57-201.32)	1.80×10^{-4}	1.18×10^{-5}
12q24.11	marker_B	AG	0.31	0.47	2.10×10^{-6}	0.51 (0.38-0.67)	**2.95×10^{-7}**	0.36 (0.24-0.53)	2.50×10^{-2}	0.54 (0.31-0.93)	1.69×10^{-6}	2.60×10^{-6}
19q13.11	marker_A1	AC	0.27	0.14	9.75×10^{-7}	2.28 (1.63-3.19)	1.78×10^{-6}	2.62 (1.75-3.9)	1.35×10^{-2}	3.64 (1.22-10.85)	4.76×10^{-6}	**7.66×10^{-7}**

MAF: minor allele frequency, w ILD: RA with ILD, w/o ILD: RA without ILD, OR: odds ratio and 95%CI: 95% confidence interval. Bold font highlights the most significant *p*-value under five models: allelic, dominant, recessive, genotypic model and Cochran-Armitage trend test.

Table 8 List of reported association of RA susceptibility SNPs and those of RA-ILD GWAS results

chr	Variant	Gene	RA				RA-ILD					
			Alleles	RAF	OR(95%CI)	*p* -value	Variant	LD(r^2)	Alleles	AF	OR(95%CI)	*p* -value
1	rs227163	*TNFRSF9*	CT	0.33	1.11 (1.08-1.16)	5.00×10^{-6}	←	←	←	0.43	0.95 (0.73-1.25)	7.05×10^{-1}
	rs2301888	*PADI4*	GA	0.42	1.19 (1.14-1.25)	8.90×10^{-13}	rs11203367	0.93	CT	0.57	1.1 (0.84-1.44)	5.33×10^{-1}
	chr1:161644258	*FCGR2B*	CG	0.21	1.15 (1.08-1.22)	3.30×10^{-6}				n.a.		
2	rs1858037	*SPRED2*	TA	0.22	1.19 (1.12-1.26)	7.40×10^{-7}	rs7559283	0.9	CT	0.17	0.87 (0.60-1.25)	4.34×10^{-1}
	rs11889341	*STAT4*	TC	0.33	1.16 (1.10-1.22)	6.30×10^{-9}	rs10168266	0.87	TC	0.32	1.16 (0.87-1.54)	3.43×10^{-1}
6	rs9268839	*HLA-DRB1*	GA	0.39	1.9 (1.81-1.99)	3.50×10^{-134}	←	←	←	0.63	0.64 (0.49-0.85)	1.60×10^{-3}
	rs2233424	*NFKBIE*	TC	0.16	1.24 (1.17-1.31)	9.30×10^{-13}				n.a.		
	rs7752903	*TNFAIP3*	GT	0.07	1.34 (1.23-1.46)	1.40×10^{-10}	←	←	←	0.09	1.2 (0.75-1.90)	4.61×10^{-1}
	rs1571878	*CCR6*	CT	0.45	1.28 (1.22-1.35)	1.10×10^{-20}				n.a.		
8	rs2736337	*BLK*	CT	0.7	1.15 (1.08-1.21)	2.00×10^{-6}	rs1478901	0.94	GC	0.71	1.14 (0.84-1.53)	4.19×10^{-1}
10	rs71508903	*ARID5B*	TC	0.18	1.18 (1.12-1.25)	4.00×10^{-9}				n.a.		
	rs6479800	*RTKN2*	CG	0.09	1.22 (1.13-1.32)	6.90×10^{-7}	rs7100297	0.84	TC	0.12	1.19 (0.80-1.78)	4.10×10^{-1}
	rs726288	*SFTPD*	TC	0.21	1.22 (1.14-1.31)	8.20×10^{-8}				n.a.		
11	rs73013527	*ETS1*	CT	0.7	1.16 (1.09-1.23)	1.20×10^{-6}	rs4245081	0.83	CT	0.71	0.91 (0.68-1.22)	5.23×10^{-1}
14	rs3783782	*PRKCH*	AG	0.26	1.14 (1.09-1.19)	9.80×10^{-8}				n.a.		
	rs2582532	*PLD4-AHNAK2*	CT	0.72	1.18 (1.11-1.25)	4.40×10^{-6}				n.a.		
18	rs8083786	*PTPN2*	GA	0.32	1.18 (1.13-1.24)	2.40×10^{-10}	rs7234029	0.98	GA	0.38	1.07 (0.82-1.41)	6.36×10^{-1}
	rs2469434	*CD226*	CT	0.35	1.11 (1.07-1.15)	1.80×10^{-6}	rs1788097	0.83	TC	0.47	1.05 (0.80-1.37)	7.68×10^{-1}
22	rs909685	*SYNGR1*	AT	0.79	1.22 (1.12-1.33)	3.80×10^{-6}				n.a.		
X	chrX:78464616	*P2RY10*	AC	0.45	1.16 (1.09-1.22)	4.40×10^{-7}				n.a.		

RAF: risk allele frequency, OR: odds ratio, 95%CI: 95% confidence interval, AF: allele frequency and n.a.: not available. An arrow denoted that RA and RA-ILD have same variant. RA susceptibility SNPs were provided by previous paper and have a *p*-value of less than 5×10^{-6} in Asian population. (Okada *et al.* [53])

Table 9 List of reported association of PF susceptibility SNPs and those of RA-ILD GWAS results

chr	Variant	Gene	PF				RA-ILD					
			Minor allele	MAF	OR (95%CI)	*p* -value	Variant	LD (r^2)	Allele	AF	OR (95%CI)	*p* -value
3	rs6793295	*LRRC34*	C	0.32	1.3 (1.19–1.42)	3.20×10^{-7}	←	←	CT	0.74	0.89 (0.66-1.21)	4.69×10^{-1}
4	rs2609255	*FAM13A*	G	0.26	1.29 (1.18–1.42)	5.27×10^{-6}	rs2609262	0.9	AG	0.45	1.21 (0.92-1.58)	1.73×10^{-1}
5	rs2736100	*TERT*	C	0.43	0.73 (0.67–0.79)	7.60×10^{-14}				n.a.		
6	rs2076295	*DSP*	G	0.54	1.43 (1.32–1.55)	1.14×10^{-16}	rs3778337	0.99	GA	0.5	0.87 (0.67-1.15)	3.29×10^{-1}
7	rs4727443	20kb 5' of *AZGP1*	A	0.46	1.3 (1.20–1.41)	6.72×10^{-9}	rs6465759	0.82	CT	0.36	1.35 (1.02-1.78)	3.62×10^{-2}
10	rs11191865	*OBFC1*	G	0.45	0.8 (0.74–0.87)	2.82×10^{-7}				n.a.		
11	rs35705950	*MUC5B*	T	0.38	8.3 (5.8-11.9)	4.60×10^{-31}				n.a.		
	rs7934606	*MUC2*	T	0.52	1.52 (1.40–1.65)	5.46×10^{-22}				n.a.		
13	rs1278769	*ATP11A*	A	0.2	0.79 (0.72–0.88)	9.11×10^{-7}				n.a.		
15	rs2034650	*IVD*	G	0.42	0.77 (0.71–0.84)	1.86×10^{-9}	rs1001528	0.96	AG	0.14	1.05 (0.72-1.54)	7.93×10^{-1}
17	rs1981997	*MAPT*	A	0.17	0.71 (0.64-0.78)	2.52×10^{-8}				n.a.		
19	rs12610495	*DPP9*	G	0.34	1.29 (1.18–1.41)	9.57×10^{-9}				n.a.		

MAF: minor allele frequency, OR: odds ratio, 95%CI: 95% confidence interval, AF: allele frequency and n.a.: not available. An arrow denoted that PF and RA-ILD have same variation. The SNP rs35705950 was from genome-wide linkage analysis (Seibold *et al.* [35]) and the others were from GWAS (Fingerlin *et al.* [38]).

Figures

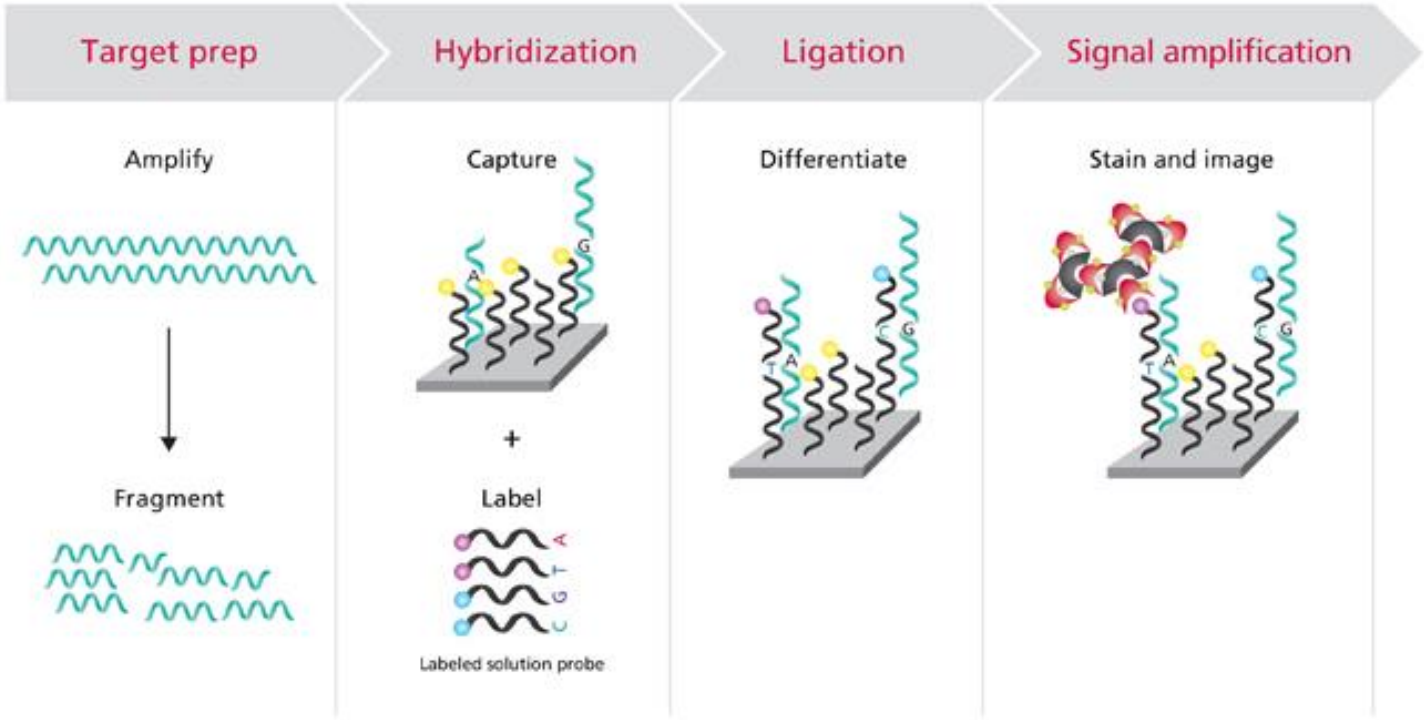
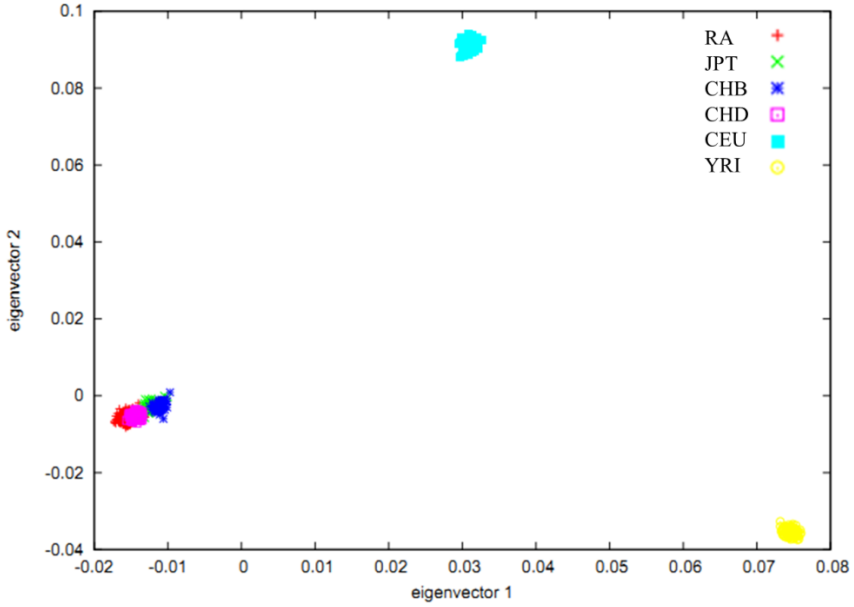


Figure 1 Schematic presentation of overview assay in Affymetrix Axiom Genome-Wide Population-Optimized Human Array. Figure is from http://www.affymetrix.com/

A. RA samples and five HapMap populations



B. RA samples and three East Asian HapMap populations

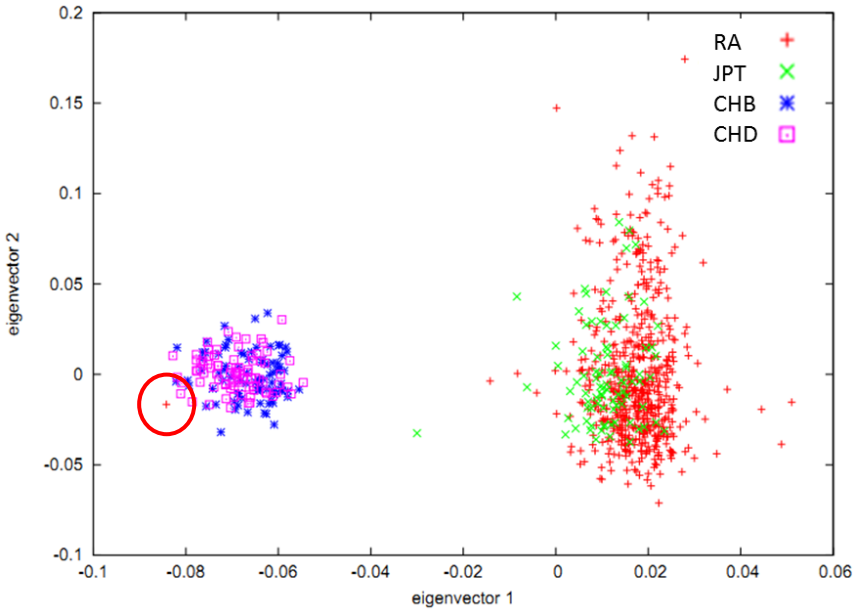


Figure 2 Principal component analysis of RA samples and reference populations; (A) RA samples and five reference populations (JPT, CHB, CHD, YRI and CEU) and (B) RA samples and three East Asian populations (JPT, CHB and CHD). One RA sample (red circle) was supposed to be an outlier from Japanese population.

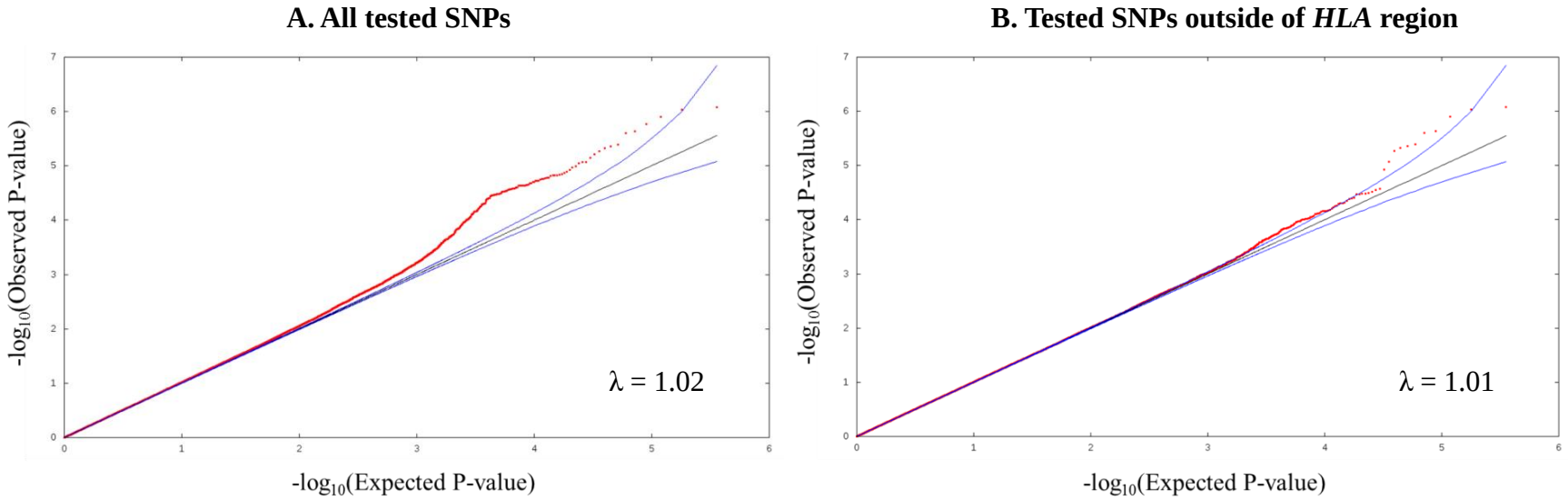


Figure 3 Quantile-quantile plots of SNPs subjected to the association test between RA with ILD and RA without ILD using (A) all tested SNPs and (B) that of SNPs outside the HLA region. Red dots showed the distribution of the negative common logarithm of expected *p*-values (x axis) compared to those of observed *p*-value (y axis). A black line shows a straight line, whose inclination is equivalent to 1. Blue lines indicate 95% confidence interval. No population stratifications were observed because of having λ-values of 1.02. λ-value decreased from 1.02 to 1.01 after selecting tested SNPs outside of *HLA* region.

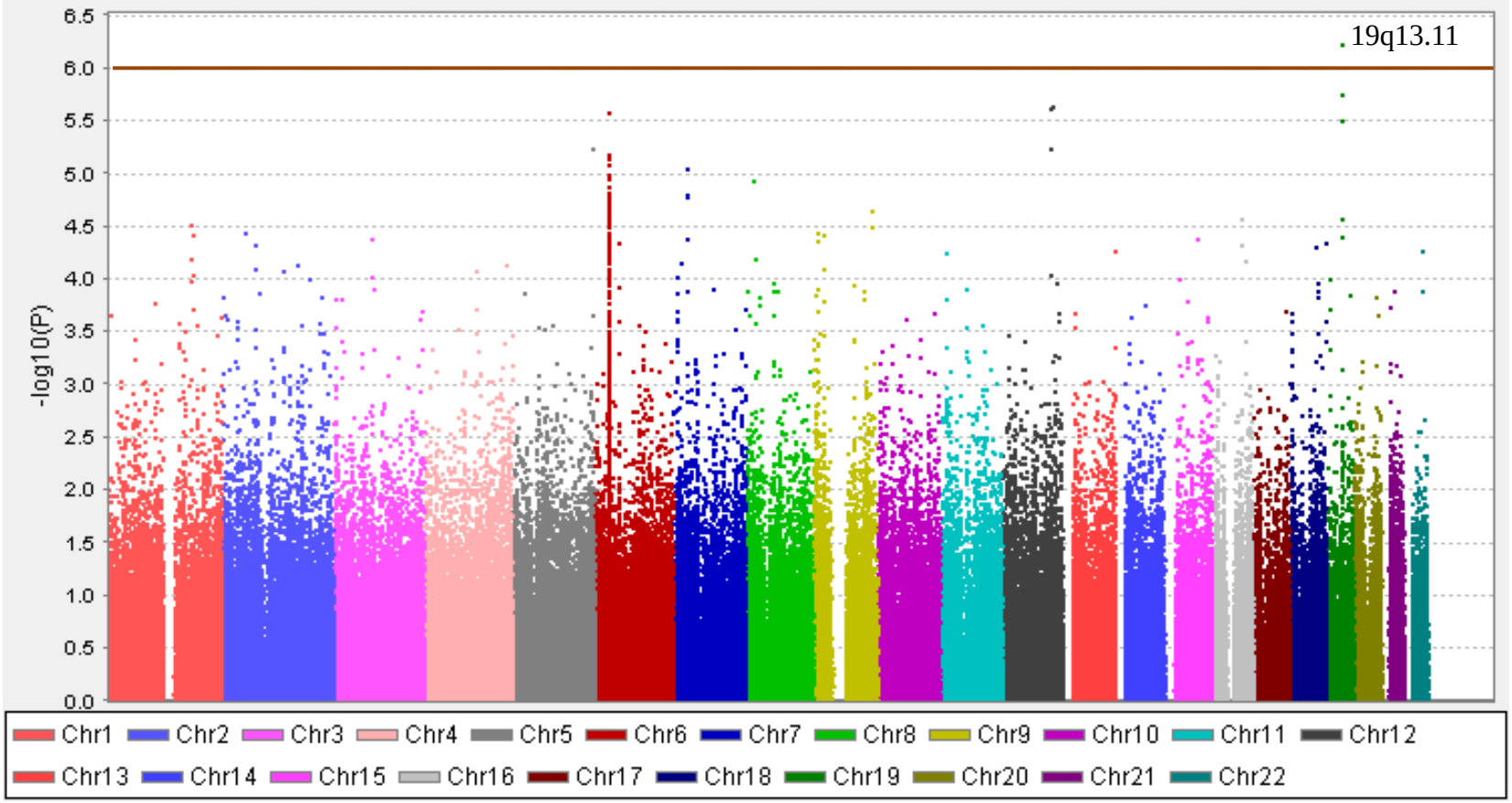


Figure 4 GWAS results of 359,782 SNPs with 169 RA with ILD and 294 RA without ILD patients calculated by Cochran-Armitage trend test. For each plot, the negative common logarithm of *p*-value (y axis) of the SNP was calculated by Cochran-Armitage trend test and plotted according to its chromosomal position (x axis). The brown line corresponded to *p*-value of 1×10^{-6}. The region 19q13.11 indicated the lowest *p*-value of 5.98×10^{-7}.

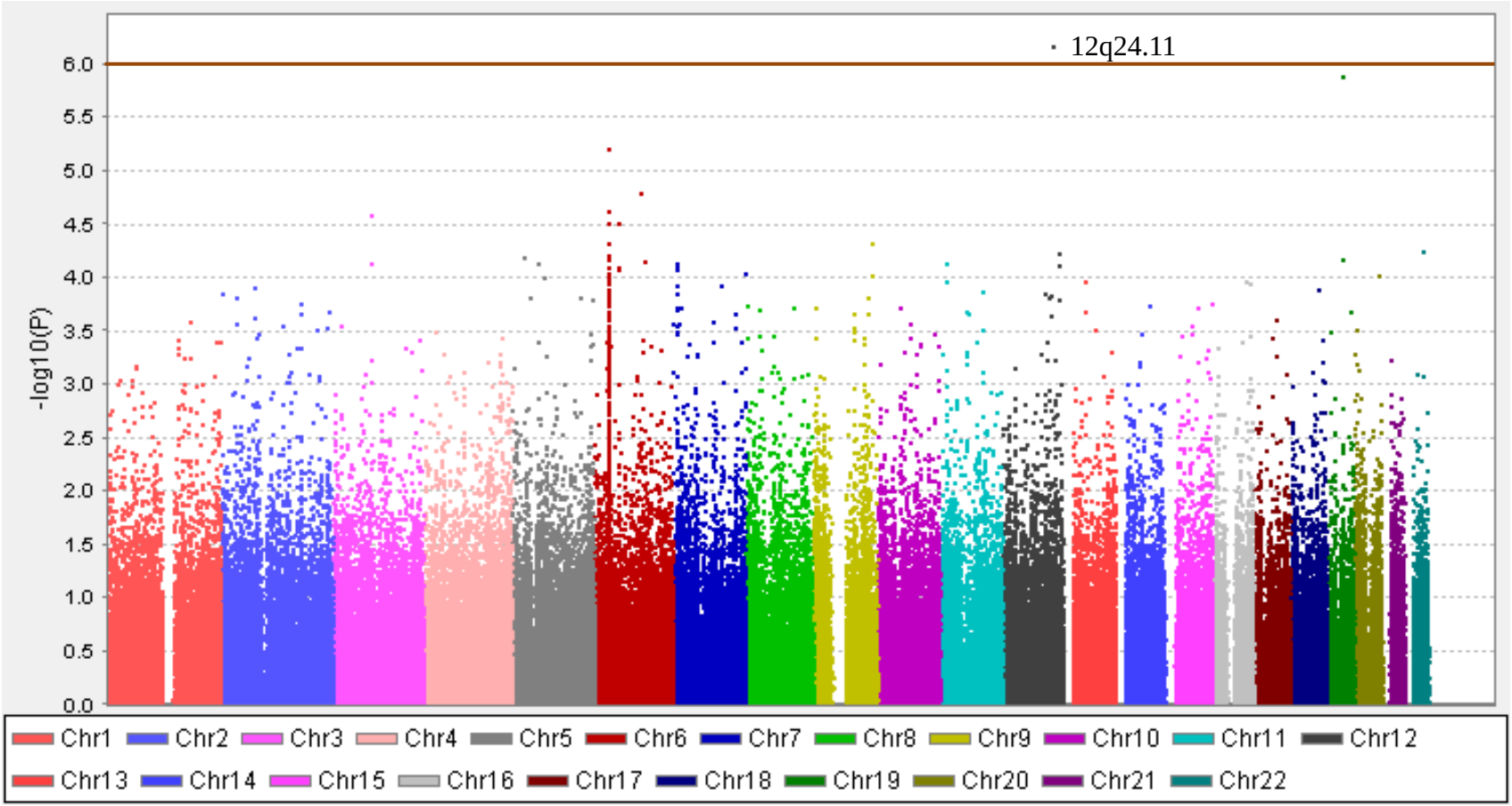


Figure 5 GWAS results of 359,782 SNPs with 169 RA with ILD and 294 RA without ILD patients under dominant model. For each plot, the negative common logarithm of *p*-value (y axis) of the SNP was calculated by dominant model and plotted according to its chromosomal position (x axis). The brown line corresponded to *p*-value of 1×10^{-6}. A SNP in the region 12q24.11 indicated a *p*-value of less than 1×10^{-6}.

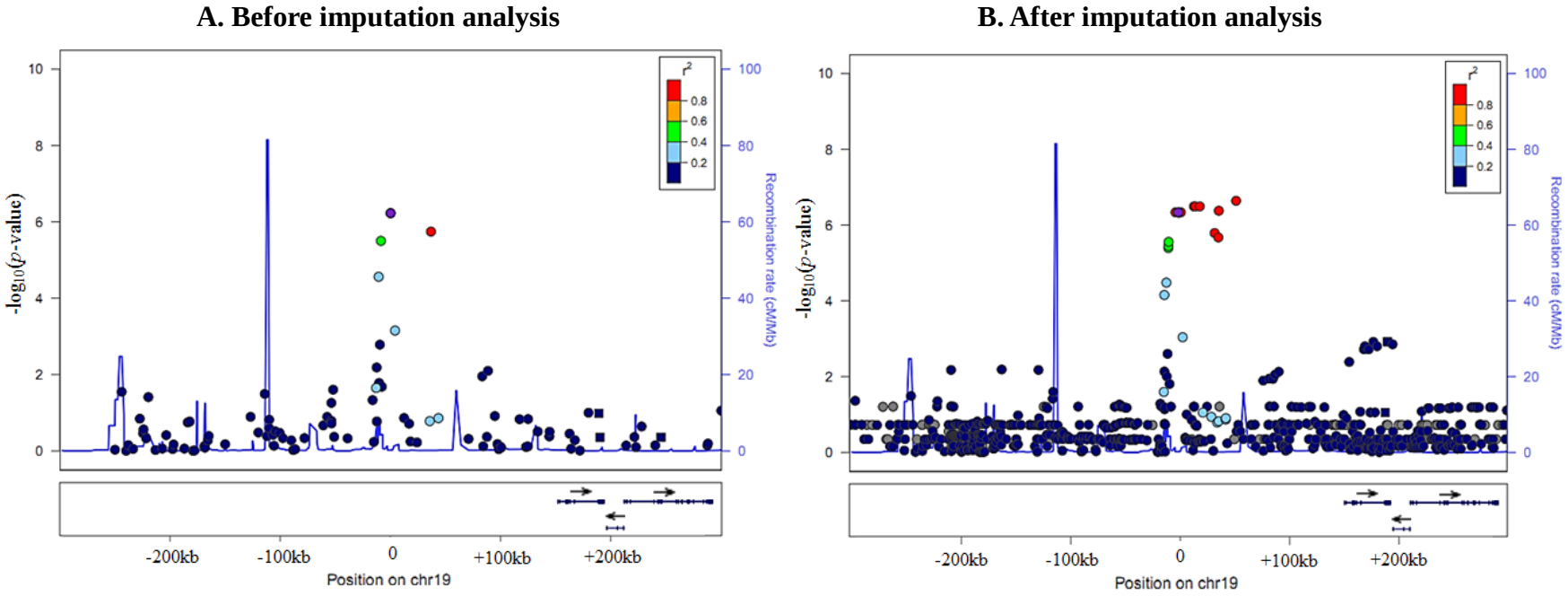


Figure 6 Regional Manhattan plots for the SNP marker_A1 in 19q13.11 (A) before imputation analysis and (B) after imputation analysis. The negative common logarithm of *p*-value (y axis) for each SNP in the region is shown according to its chromosomal positions (x axis). The SNP marker_A1 is shown in purple. Blue line behind the plots indicated the recombination rate. Colors of each plot reflected the r-square value between the SNP marker_A1 and the other plotted SNPs (blue: $0 < r^2 \leqq 0.2$, light blue: $0.2 < r^2 \leqq 0.4$, green: $0.4 < r^2 \leqq 0.6$, orange: $0.6 < r^2 \leqq 0.8$ and red: $r^2 \geqq 0.8$). SNPs are functionally annotated as synonymous SNP or UTR (square) and no annotation (circle). Blue lines in the lower box show the structure of genes. Arrows in the lower box represented the coding strand of genes. 13 SNPs showed lower *p*-values than that of the SNP marker_A1 and the lowest *p*-value was 1.87×10^{-7}.

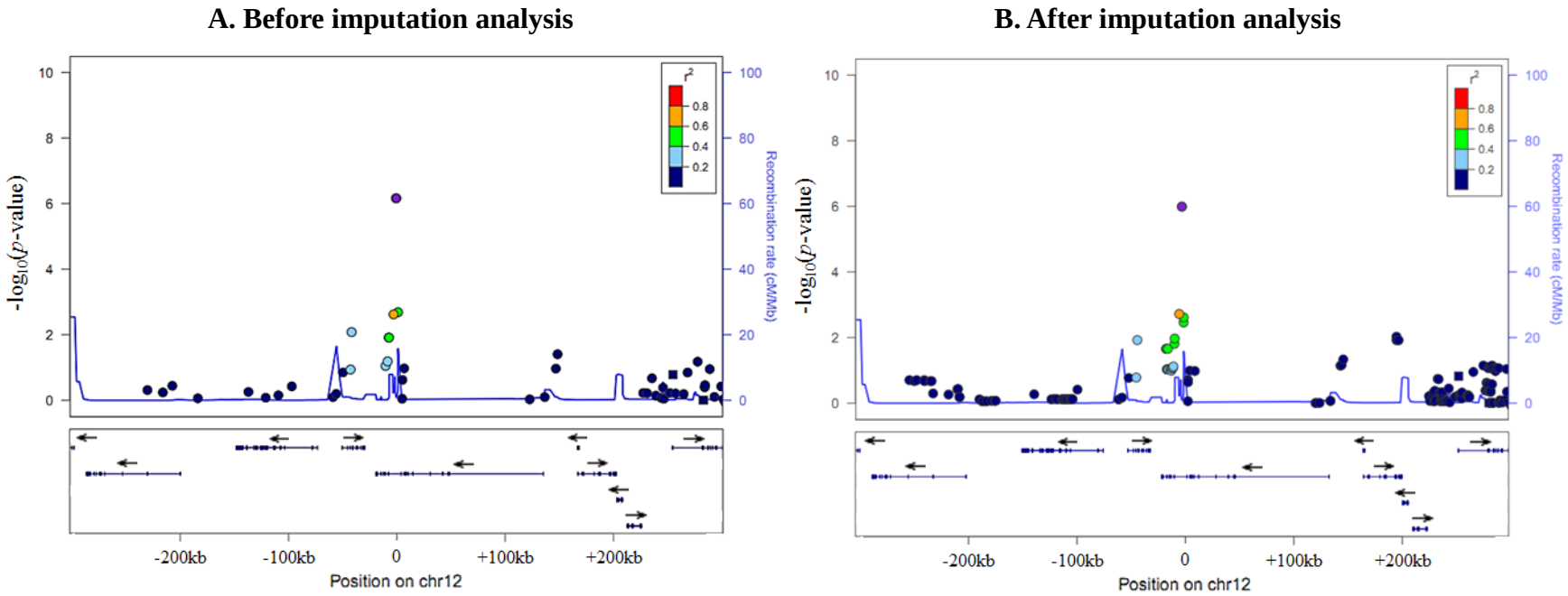


Figure 7 Regional Manhattan plots for the SNP marker_B in 12q24.11 (A) before imputation analysis and (B) after imputation analysis. The negative common logarithm of *p*-values (y axis) for each SNP in the region is shown according to its chromosomal position (x axis). The SNP marker_B is shown in purple. Blue line behind the plots indicated the recombination rate from HapMap. Colors of each plot reflected the r-square value between the SNP marker_B and the other plotted SNP**s** (blue: $0 < r^2 \leqq 0.2$, light blue: $0.2 < r^2 \leqq 0.4$, green: $0.4 < r^2 \leqq 0.6$, orange: $0.6 < r^2 \leqq 0.8$ and red: $r^2 \geqq 0.8$). SNPs are functionally annotated as synonymous SNP or UTR (square) and no annotation (circle). Blue lines in the lower box show the structure of genes. Arrows in the lower box represented the coding strand of genes. No SNP showed stronger association than the SNP marker_B**.**

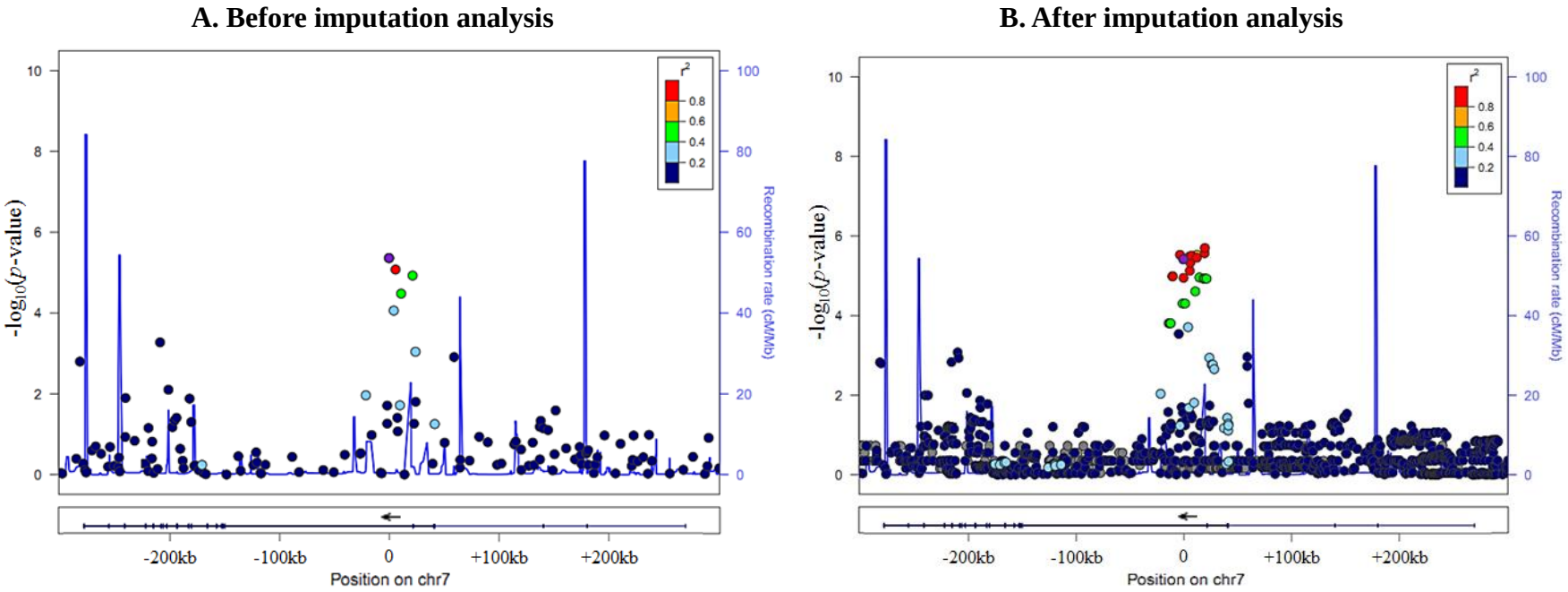


Figure 8 Regional Manhattan plots for the SNP marker_H in 7p14.3 (A) before imputation analysis and (B) after imputation analysis. The negative common logarithm of *p*-value (y axis) for each SNP in the region is shown according to its chromosomal position (x axis). The SNP marker_H is shown in purple. Blue line behind the plots indicated the recombination rate. Colors of each plot reflected the r-square value between the SNP marker_H and the other plotted SNPs (blue: $0 < r^2 \leqq 0.2$, light blue: $0.2 < r^2 \leqq 0.4$, green: $0.4 < r^2 \leqq 0.6$, orange: $0.6 < r^2 \leqq 0.8$ and red: $r^2 \geqq 0.8$). SNPs are functionally annotated as synonymous SNP or UTR (square) and no annotation (circle). Blue lines in the lower box showed the structure of the gene. An arrow in the lower box represented the coding strand of genes. Seven SNPs showed lower *p*-values than that of the SNP marker_H and the lowest *p*-value was 2.01×10^{-6}.

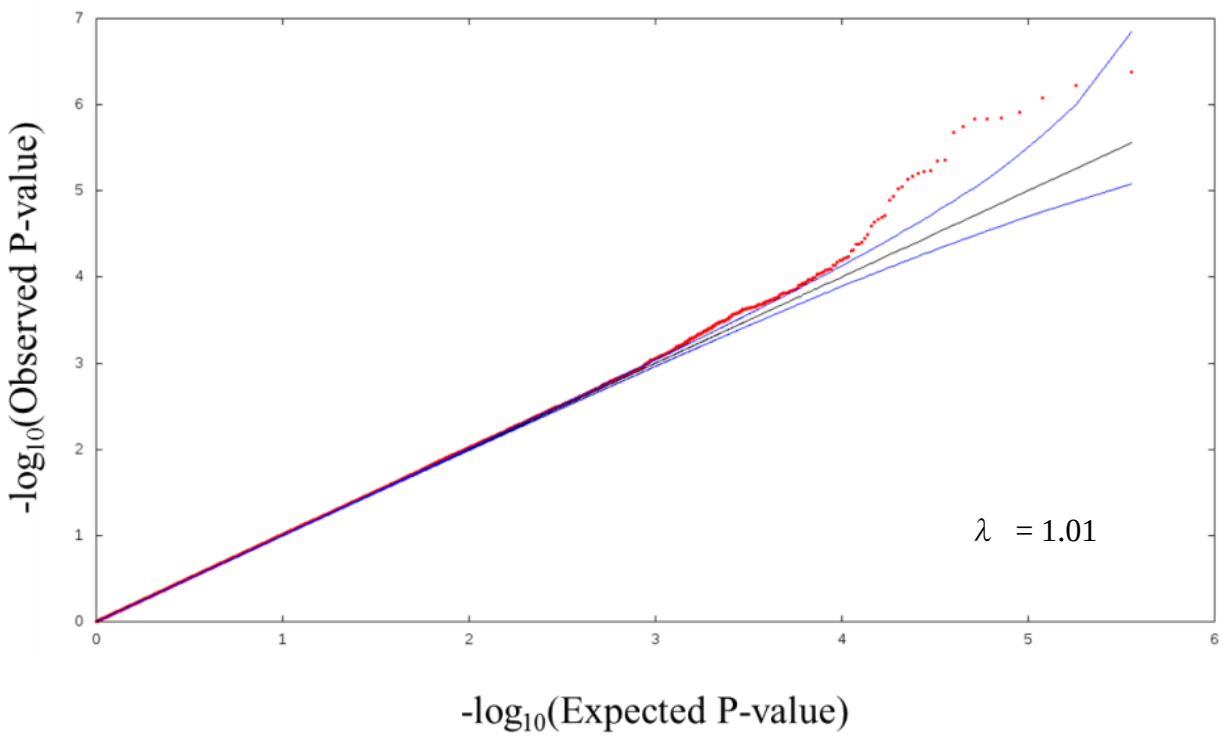


Figure 9 Quantile-quantile plots of the association test between RA with ILD and RA without ILD patients for *HLA-DRB1*04* positive patients. Red dots showed the distribution of the negative common logarithm of expected *p*-values (x axis) compared to those of observed *p*-value (y axis). A black line shows a straight line, which inclination is equivalent to 1. Blue lines indicate 95% confidence interval. The Q-Q plots showed λ-value of 1.01 and no population stratification was observed.

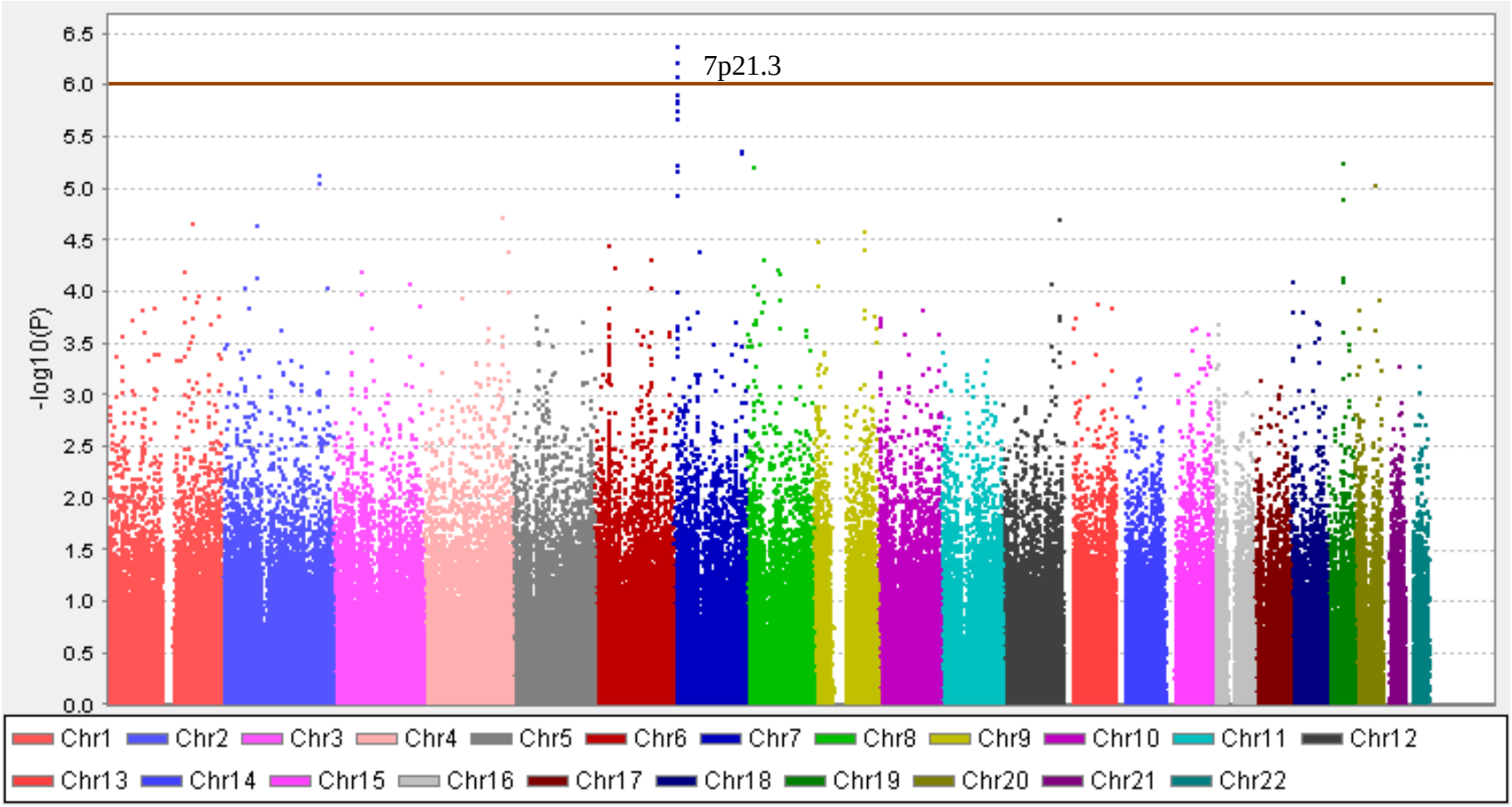


Figure 10 GWAS results of 359,782 SNPs with 85 RA with ILD and 197 RA without ILD patients under an allelic model for *HLA-DRB1*04* positive patients. For each plot, the negative common logarithm of *p*-value (y axis) of the SNPs was shown according to its chromosomal position (x axis). The brown line corresponds to *p*-value of 1×10^{-6}. The region 7p21.3 showed a clear peak and strong association.

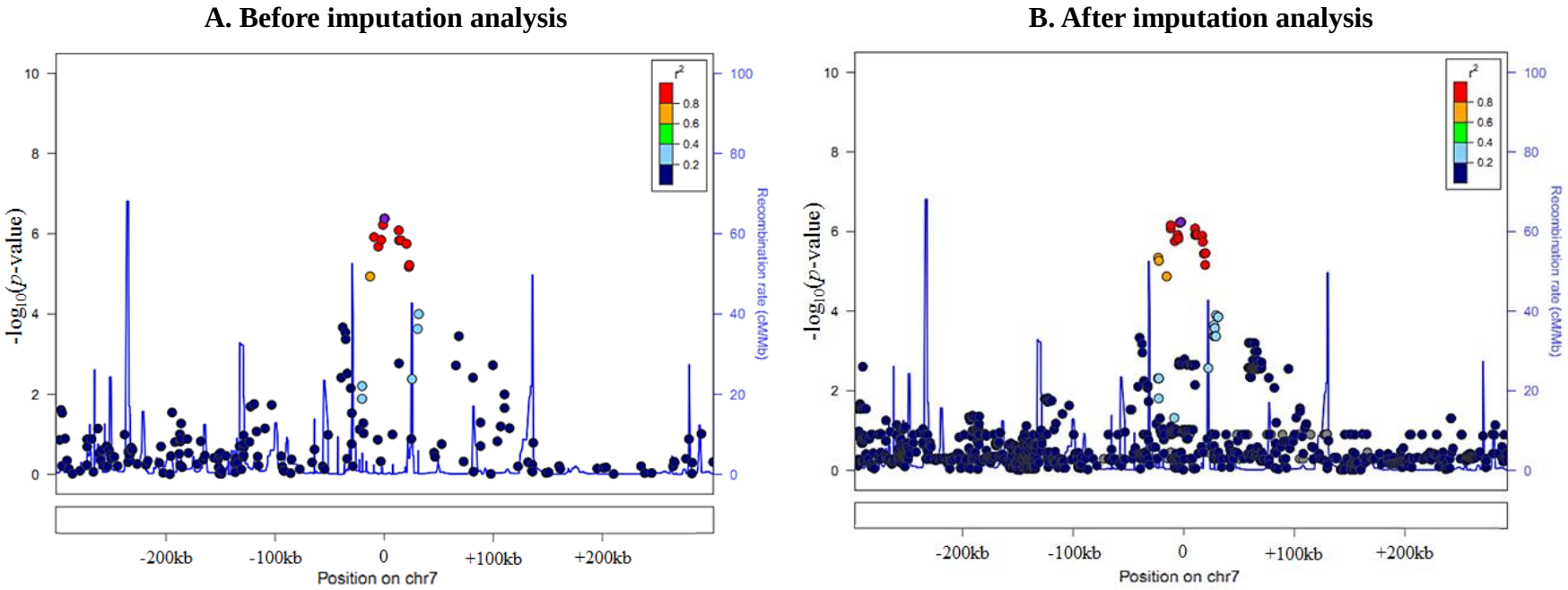


Figure 11 Regional Manhattan plots for the SNP marker_I1 in 7p21.3 (A) before imputation analysis and (B) after imputation analysis. The negative common logarithm of *p*-value (y axis) for each SNP in the region is shown according to its chromosomal position (x axis). The SNP marker_I1 is shown in purple. Blue line behind the plots indicated the recombination rate. Colors of each plot reflected the r-square value between the SNP marker_I1 and the other plotted SNPs (blue: $0 < r^2 \leqq 0.2$, light blue: $0.2 < r^2 \leqq 0.4$, green: $0.4 < r^2 \leqq 0.6$, orange: $0.6 < r^2 \leqq 0.8$ and red: $r^2 \geqq 0.8$). Imputation analysis showed that no SNP had stronger association than the SNP marker_I1.

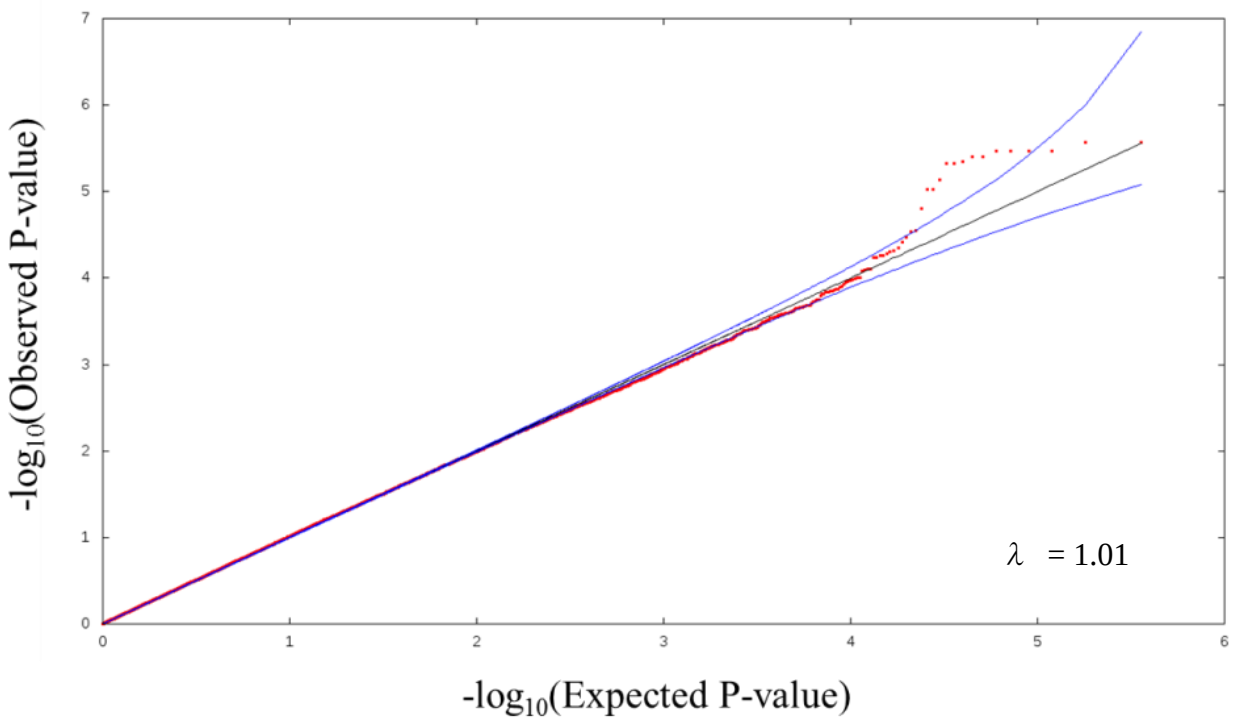


Figure 12 Quantile-quantile plots of the association test between RA with ILD and RA without ILD patients for *HLA-DRB1*04* negative patients. Red dots showed the distribution of the negative common logarithm of expected *p*-values (x axis) compared to those of observed *p*-value (y axis). A black line shows a straight line, which inclination is equivalent to 1. Blue lines indicate 95% confidence interval. The Q-Q plots did not indicate inflation due to population stratification: λ-values of 1.01.

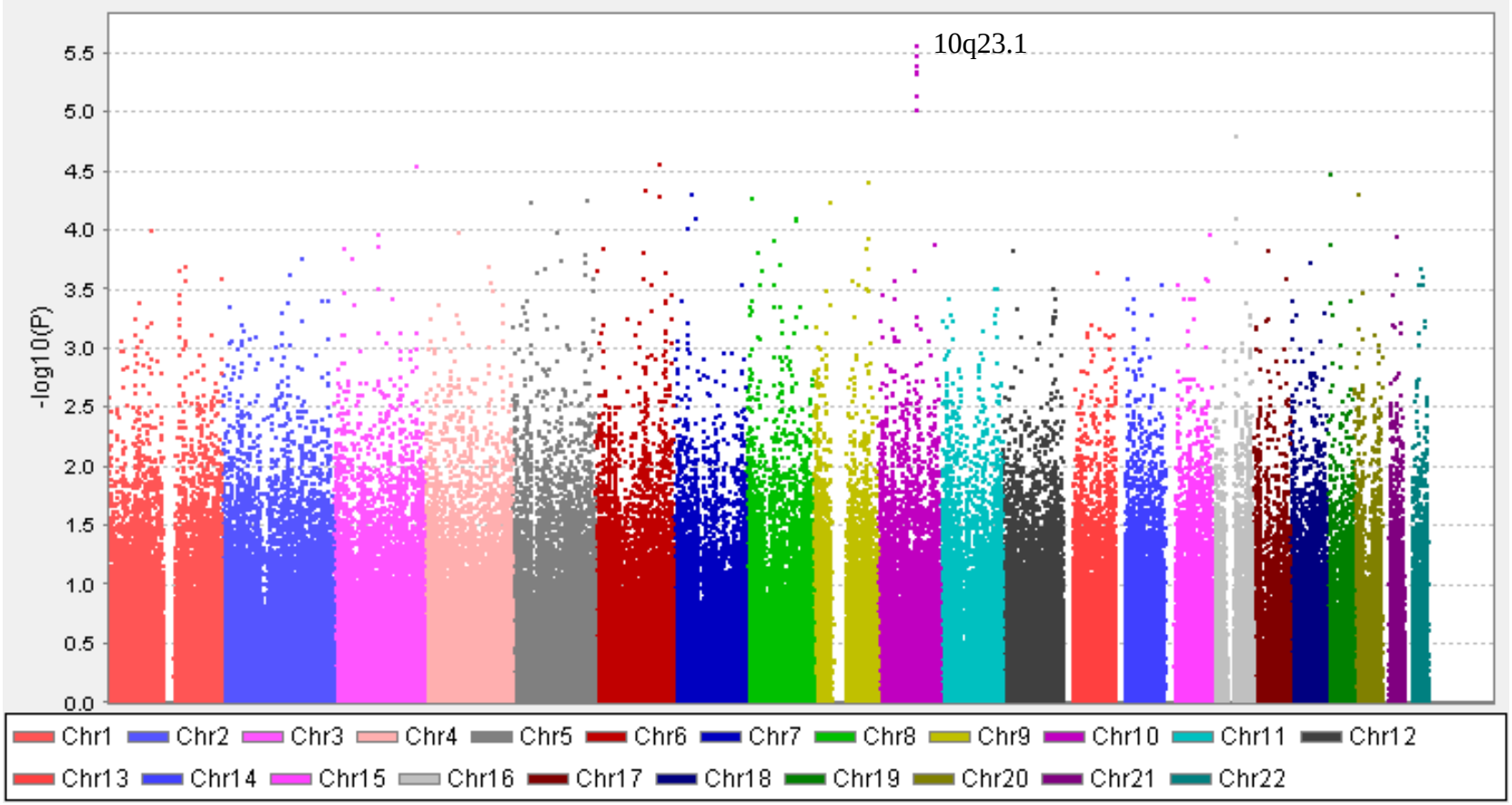


Figure 13 GWAS results of 359,782 SNPs with 84 RA with ILD and 97 RA without ILD patients under an allelic model for *HLA-DRB1*04* negative patients. For each plot, the negative common logarithm of *p*-value (y axis) of the SNP was shown according to its chromosomal position (x axis). The region 10q23.1 showed an association although the number of samples was decreased.

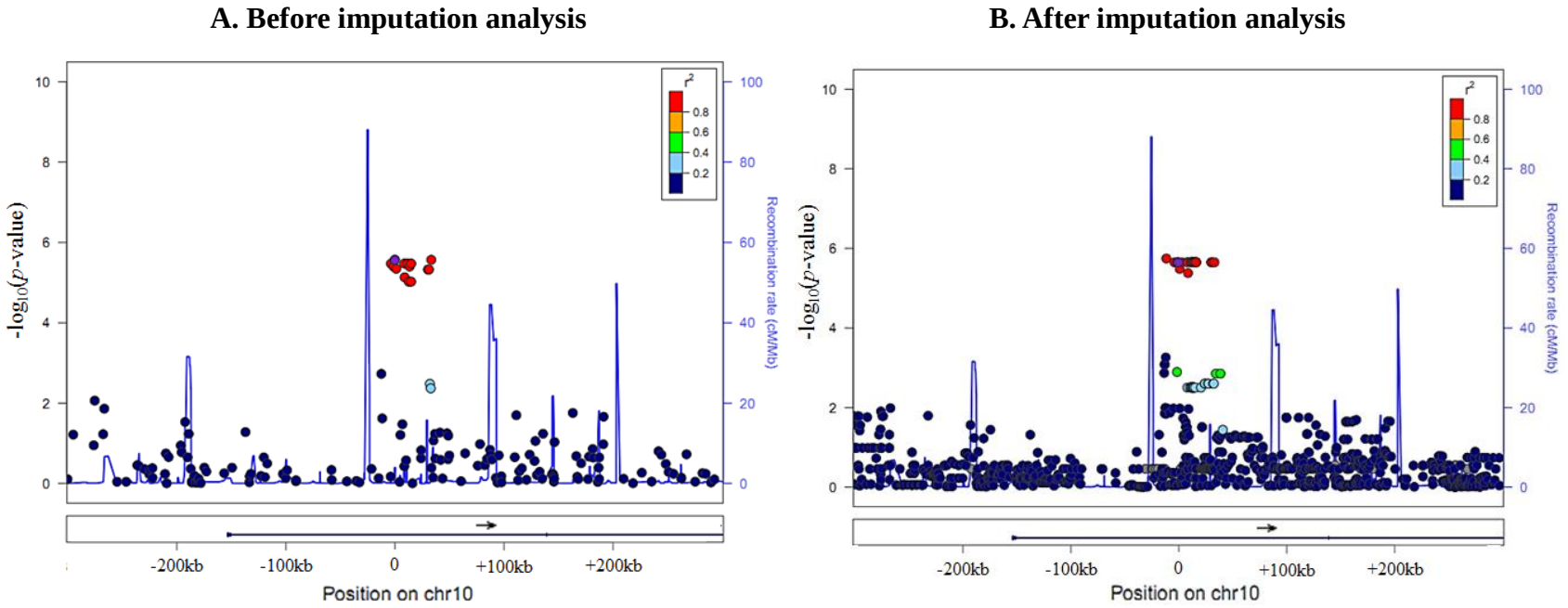


Figure 14 Regional Manhattan plots for marker_K1 (A) before imputation analysis and (B) after imputation analysis. The negative common logarithm of *p*-value (y axis) for each SNP in the region is shown according to its chromosomal position (x axis). The SNP marker_K1 is shown in purple. Blue line behind the plots indicated the recombination rate. Colors of each plot reflected the r-square value between the SNP marker_K1 and the other plotted SNP**s** (blue: $0 < r^2 \leqq 0.2$, light blue: $0.2 < r^2 \leqq 0.4$, green: $0.4 < r^2 \leqq 0.6$, orange: $0.6 < r^2 \leqq 0.8$ and red: $r^2 \geqq 0.8$). Blue lines in the lower box show the structure of the gene. An arrow in the lower box represented the coding strand of genes. One SNP showed lower *p*-value than that of the SNP marker_K1 and its *p*-value was 1.80×10^{-6}