2018

Genetic Influences On Rheumatoid Arthritis In Global Populations

Vincent Albert Laufer

University of Alabama at Birmingham

GENETIC INFLUENCES ON RHEUMATOID ARTHRITIS IN GLOBAL POPULATIONS

by

VINCENT A. LAUFER

S. LOUIS BRIDGES, MD, PHD, COMMITTEE CHAIR
SEAN DAVIS, MD, PHD
ROBERT P. KIMBERLY, MD
ROBINNA G. LORENZ, MD, PHD
HEMANT K. TIWARI, PHD

A DISSERTATION

Submitted to the graduate faculty of the University of Alabama at Birmingham,
in partial fulfillment of the degree of
Doctor of Philosophy

BIRMINGHAM, ALABAMA

2018

GENETIC INFLUENCES ON RHEUMATOID ARTHRITIS IN
GLOBAL POPULATIONS

VINCENT A. LAUFER

CELL, MOLECULAR AND DEVELOPMENTAL BIOLOGY

ABSTRACT

Rheumatoid arthritis (RA) is a complex disease having numerous genetic and environmental risk factors the interplay of which produces RA pathobiology. While the sheer number of genetic and environmental risk factors complicates understanding of disease biology, understanding has progressed far enough for insight into the most likely mechanisms for the development of the disease. Modern studies of the genetics of RA are massively parallel, enabling researchers to systematically interrogate variants throughout the human genome for associations in genome-wide association studies or GWAS.

Such studies have been carried out in European and Asian cohorts many times, and the most recent RA meta-analyses includes tens of thousands of genotypes from these populations. By contrast, there is a paucity of genotyping data available in individuals of African ancestry with RA. In the studies that follow, we attempt to address this disparity by presenting the largest genetic studies in African-Americans to date.

Following these association studies, we employ fine-mapping methods, which operate on association results and output a (short) list of candidate pathogenic. Recent studies have become increasingly sophisticated in their approaches to this. One such approach is trans-ethnic fine-mapping, which uses differences in the association pattern and LD between variants in the same risk locus across multiple global populations. By examining these together, fine-mapping algorithms can estimate which variants are the most likely to be the pathogenic variants. In the present studies, we carry out fine-mapping studies using aggregated data that draws on >100,000 RA patients and controls from 3 global ancestries.

Thus, our studies had 3 chief aims. First, we aimed to discover novel associations with RA that have not been found before in other ethnicities. Second, we endeavored to validate in African-Americans known associations identified in genetic studies of RA in Asians and Europeans with RA. Last, we employed trans-ethnic fine-mapping algorithms to isolate candidate causal variants in the loci we identified. Pursuant to the first aim, we find 3 novel associations with RA in the *CSMD3*, *GPC5*, and *RBFOX1* loci that appear to be unique to individuals of African ancestry. Second, we replicate 28 genetic risk loci discovered in other populations, and present evidence that 4 such loci are unlikely to replicate. Last, we identify several new candidate pathogenic variants, including several that may have relevance for precision medicine.

The findings in these studies have far-ranging implications for the design of future genetic studies, in particular for those that hope to cost-effectively identify functional variants that produce RA, which is necessary before mechanistic studies of how genetic variants can produce disease risk can begin. Therefore, the present study can serve as a basis for future studies into the genetics of RA in global populations.

Keywords: Rheumatoid Arthritis, Genetics, GWAS, meta-analysis, trans-ethnic, fine-mapping, complex disease, post-GWAS, African-American

ACKNOWLEDGEMENTS

I owe my deepest thanks to my mentor, Lou Bridges, for his patient support. Lou – I know that in many regards I must not have been the easiest student to mentor. Thank you for consistently making time for me, too, among the dozens of faculty and staff whose needs you attend to. Thank you for allowing me the time I needed to learn how to analyze genetic data well despite the external demands placed on you to publish these studies quickly. Thank you for being as much a scientific mentor and as a role model: you show all of us in the rheumatology community how to proceed with integrity even in the highly competitive environment of modern research. I sincerely hope that the analysis I performed more fully realizes the research goals you envisioned than what would have resulted without my involvement, and that this is some recompense for all the time and expense you invested in my development.

I also thank the members of my PhD advisory committee Drs. Lorenz, Kimberly, Davis, and Tiwari, for the guidance you provided me not only on the experiments conducted but also on necessary steps to develop as a scientist more generally.

In addition, I need to acknowledge three individuals in particular who were not on my advisory committee, but without whom it would have been difficult to progress. I chose a PhD that demanded a great deal of statistics and computer programming without having had formal training in either. I fear that without the aid of Dr. Richard J. Reynolds IV and Dr. Carl D. Langefeld my development as a statistician would have been too slow, and my PhD studies would have suffered as a result. I would also like to thank Dr. Langefeld for being the first of my advisors to draw my attention to the importance of differences between populations and starting my thinking on how that might be a productive area for research. Similarly, I feel that without the intensive help in computer programming I received from Mr. Curtis Hendrickson, I would not have progressed fast enough to complete my studies adequately. The three of you helped more than you know.

I also thank Mr. Randy Seay for help with numerous tasks relating to advancement in the graduate school, scheduling courses, and handling sometimes urgent administrative tasks. Likewise, I sincerely appreciate the help of Dr. Carol Ballinger for help addressing all the tasks necessary to gain and maintain my status as a fellow on the Rheumatology T32 grant.

Finally, I have been lucky beyond measure to have had the support of two parents who never gave a second thought to putting my needs above their own. Your unconditional love is the reason I am where I am. Last, I would like to thank Mr. Laurence Laufer, Ms. Wesli Turner, Ms. Anastasia Hartzes-Hirschy, Mr. Tarrant McPherson and Drs. Josh Cohen, Alice Weaver, and Alex Dussaq for sustaining me during this long endeavor.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table *Page*

INTEGRATIVE APPROACHES TO UNDERSTANDING THE ROLE OF PATHOGENIC VARIATION IN RHEUMATIC DISEASE

DENSE GENOTYPING IDENTIFIES LOCI FOR RHEUMATOID ARTHRITIS RISK AND DAMAGE IN AFRICAN AMERICANS

GENETIC INFLUENCES ON SUSCEPTIBILITY TO RHEUMATOID ARTHRTIS IN THREE GLOBAL POPULATIONS

LIST OF FIGURES

INTRODUCTION

DENSE GENOTYPING IDENTIFIES LOCI FOR RHEUMATOID ARTHRITIS RISK AND DAMAGE IN AFRICAN AMERICANS

GENETIC INFLUENCES ON SUSCEPTIBILITY TO RHEUMATOID ARTHRTIS IN THREE GLOBAL POPULATIONS

LIST OF ABBREVIATIONS

A1	– Allele 1; the effect allele.
A2	– Allele 2; the alternate allele.
AA	– African-American
ACPA	– Anti-citrullinated peptide/protein antibodies
AFR	– 1000 Genomes super population code for African
ALL	– Acute lymphocytic leukemia
BP	– Base pair location as per genome build
CI	– Confidence Interval
CHB	– 1000 Genomes population code for Han Chinese in Bejing, China
JPT	– 1000 Genomes population code for Japanese in Tokyo, Japan
CLEAR	– Consortium for Longitudinal Evaluation of Early Arthritis Registiry
EAF	– Expected allele frequency
EAS	– East Asian
EUR	– 1000 genomes super population code for European
HIV	– Human immunodeficiency virus
HLA	– Human leukocyte antigen
LD	– Linkage disequilibrium
MAF	– Minor allele frequency
MEGA	– Multi Ethnic Genotyping Array
MHC	– Major histocompatibility complex

MMR	– Measles, mumps, rubella
MS	– Multiple sclerosis
MTX	– Methotrexate
OR	– Odds ratio
ORL	– The lower bound of the confidence interval on an odds ratio
ORU	– The upper bound of the confidence interval on an odds ratio
PICS	– Probabilistically identified causative SNP
RA	– Rheumatoid arthritis
RF	– Rheumatoid Factor
SNP	– Single nucleotide polymorphism
T1D	– Type 1 diabetes
T2D	– Type 2 diabetes
TEMA	– Trans-ethnic meta-analysis
TEFM	– Trans-ethnic fine-mapping
VARA	– Veterans Affairs Rheumatoid Arthritis registry
YRI	– 1000 Genomes population code for Yoruban

INTRODUCTION

Background

Overview

The studies presented in this dissertation are an endeavor to bring understanding of the genetic basis of RA in African-American populations to the same standard as that of European and Asian populations. Along the way, we make several observations about the genetics of RA in all populations that draw on prior studies as well as on the field of population genetics. To do this, we had three specific goals. First, we aimed to discover novel associations with RA that have not been found before in other ethnicities. Second, we endeavored to validate in African-Americans known associations identified in genetic studies of RA in Asians and Europeans with RA. Last, we employed trans-ethnic fine-mapping algorithms to isolate candidate causal variants in the loci we identified. Pursuant to the first aim, we find 3 novel associations with RA in the *CSMD3*, *GPC5*, and *RBFOX1* loci that appear to be unique to individuals of African ancestry. Second, we replicate 28 genetic risk loci discovered in other populations, and present evidence that 4 such loci are unlikely to replicate. Last, we identify several new candidate pathogenic variants, including several that may have relevance for precision medicine.

Incidence and Prevalence of Rheumatoid Arthritis

Rheumatoid Arthritis (RA) is a complex autoimmune condition affecting about 1.3 million U.S. adults and 0.5-1% of the population worldwide [1]. Like many autoimmune conditions, the incidence of RA appears to be rising in developed nations such as the United States [2]. RA is extremely variable in its clinical course, both in terms of its clinical features and the severity of its presentation. Indeed, its hallmark clinical feature, a symmetrical polyarthritis, can range from a mild, indolent swelling to fulminant disease quickly leading to joint destruction, deformity, and impairment with activities of daily living [3]. Because these articular manifestations as well as extra-articular manifestations such as cardiovascular disease can be both prolonged and severe, there is significant morbidity, mortality, and economic cost resulting from RA [4].

Diagnostic Algorithms for Rheumatoid Arthritis

As a result of its myriad manifestations, clinicians have created systems and diagnostic algorithms not only for presence and absence of RA [5], but for severity of joint damage [6], disease activity [7], and for other features of the disease. In the current study, the ACR/EULAR 2010 criteria were used for diagnosis with RA, but the modified Total Sharp Score (mTSS) was used to assign patients a score from 0-448 indicating the severity of their joint disease, and the Disease Activity Score (DAS) 28 was used to measure the clinical activity of disease (see also Methods in Section II). However, numerous other classification algorithms exist and their application to create research cohorts for use in genetic studies has not been uniform. Thus, although these criteria were applied consistently in the current study, the meta-analyses performed may include subjects recruited under different inclusion and exclusion criteria. For discussion how this influences the current study, please see Limitations of the current studies in Section V.

Environmental risk factors for RA

Numerous environmental and genetic risk factors have been established for RA, and distinct combinations of these risk factors appear to interact not merely to increase risk of RA, but to modulate risk of specific RA subtypes (e.g. seronegative and seropositive RA, see below) and subphenotypes (i.e. disease subtraits [8]). Smoking, which is the best-established and strongest environmental risk factor for RA, provides an instructive example. A study of 370,000 women from the Women's Health Cohort Study found that women who smoked at least 25 cigarettes a day for more than 20 years found a modest relative risk (RR) of 1.4 compared to never smokers [9]. A Danish study of 515 patients separated study subjects into anti-citrullinated peptide antibody (ACPA) positive and ACPA-negative groups, and found that heavy smoking (>25 pack years) did not elevate risk of ACPA-negative RA, while the odds ratio for ACPA-positive RA was 17.8. Additionally, they observed an odds ratio of 53 for individuals homozygous for the shared-epitope and who had ACPA antibodies. These observations strongly suggest that a gene-by-environment interaction effect is present. However, it has remained difficult to ascribe a mechanism for this association. Cigarette smoke is thought to contain >4,000 toxic substances, which in aggregate affect dozens of immune phenotypes, including (but not limited to) IL-6 signaling, TNF-α signaling, production and processing of reactive oxygen species, and T-cell response to dendritic cells [10]. We highlight this example because it is indicative of several over-arching trends in RA research. First, the knowledge of specific associations and interaction effects has not translated into mechanistic knowledge in part due to the complexity of the phenomena involved. Second, etiologic risk factors that may look weak when studied in isolation may in fact confer substantial risk in the right context. Third, interaction effects can predispose patients to specific disease subtypes and indeed to particular manifestations of RA, but not others.

Sex-specific factors also exert a strong effect on RA risk. Women are 2-3 times more likely to develop RA than men [11]. Nulliparity is positively correlated with RA susceptibility, while third trimester pregnancy and breastfeeding are associated with remission of the disease. Though the reasons for this are not entirely clear, there is evidence both from animal models and human studies that estrogen and testosterone may underlie this differential susceptibility [12]. That these sex hormones affect B and T cell populations has long been known, but more recent studies of C57BL/6, C3H/lpr, and B/W mice have shown that they affect the most suspect adaptive immune cell subsets during active disease [13].

Occupational exposure to a variety of substances including silica and asbestos also appear to increase the likelihood of developing RA. In addition, alcohol intake, obesity, stress and physical disorders, birthweight, lower socioeconomic status (SES) positively correlate with RA [14]. As with the example of smoking highlighted above, these risk factors also exert effects through complicated pathways and probably through multiple immune intermediaries. While a complete description of how each of these contributes is beyond the scope of this dissertation, understanding that these environmental risk factors interact with heritable risk factors within the context of a person's genetic background is crucial to both the design of genetic studies of RA and also to their interpretation.

Selected Etiological Features of RA

Cellular interactions between immune and non-immune cells

RA etiology is a vast network of complex interactions between numerous innate and adaptive immune cell types as well as non-immune cells. Here, we focus on three processes relevant to the current studies.

1. We illustrate intercellular interactions that produce the articular manifestations of RA in order to provide a basis for understanding studies of radiographic severity of RA.
2. We describe the development of pathogenic autoantibodies within the RA synovium (but not osteoarthritic or healthy synovial tissue), which provides crucial biological understanding that underpins the distinction between seropositive and seronegative RA.
3. We summarize alterations to T-cell activation, including both T cell stimulation and co-stimulation as organizes about two dozen of the strongest genetic risk factors for RA into a well-defined biological event.

While other biological events and processes are important to RA, these will provide important context for the studies described herein.

Cellular interactions leading to joint disease in RA

In RA, the homeostasis of both bone and cartilage extracellular matrix (ECM) is tipped towards net catabolism. Clinical and animal studies indicate that this occurs via reduced anabolism, e.g. through nitric oxide-dependent inhibition of proteoglycan

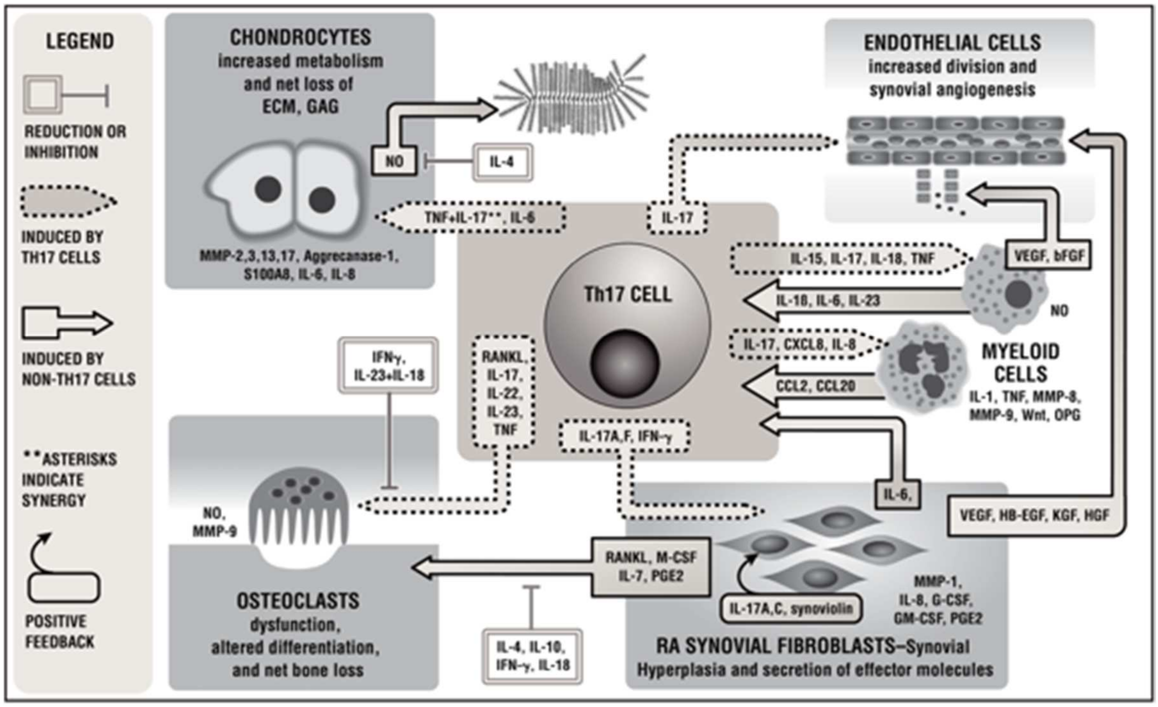

LEGEND
REDUCTION OR INHIBITION
INDUCED BY TH17 CELLS
INDUCED BY NON-TH17 CELLS
**ASTERISKS INDICATE SYNERGY
POSITIVE FEEDBACK
CHONDROCYTES
increased metabolism and net loss of ECM, GAG
NO
IL-4
TNF+IL-17**, IL-6
MMP-2,3,13,17, Aggrecanase-1, S100A8, IL-6, IL-8
ENDOTHELIAL CELLS
increased division and synovial angiogenesis
IL-17
Th17 CELL
IL-15, IL-17, IL-18, TNF
VEGF, bFGF
IL-18, IL-6, IL-23
NO
IL-17, CXCL8, IL-8
MYELOID CELLS
IL-1, TNF, MMP-8, MMP-9, Wnt, OPG
CCL2, CCL20
IFNγ, IL-23+IL-18
RANKL, IL-17, IL-22, IL-23, TNF
IL-17A,F, IFN-γ
IL-6,
VEGF, HB-EGF, KGF, HGF
NO, MMP-9
OSTEOCLASTS
dysfunction, altered differentiation, and net bone loss
RANKL, M-CSF IL-7, PGE2
IL-4, IL-10, IFN-γ, IL-18
IL-17A,C, synoviolin
MMP-1, IL-8, G-CSF, GM-CSF, PGE2
RA SYNOVIAL FIBROBLASTS–Synovial Hyperplasia and secretion of effector molecules

Figure 1 - Direct and indirect interactions between cells and chondrocytes, osteocytes, endothelial cells, myeloid cells, and fibroblasts in RA. Interactions between these cell types produce many of the hallmarks of RA pathology, including loss of cartilage and bone, synovial hyperplasia, and synovial angiogenesis. The upper myeloid cell represents a macrophage, the lower myeloid cell represents a neutrophil. Arrows may represent cytokine- based influences or influences dependent on other processes, such as cell contact. A double box with a flat- headed arrow signifies an inhibitory relationship. A single dotted box indicates that the factors within the box are produced by Th17 cells, while a single solid box indicates the factors are produced by any other cell. Two asterisks next to a pair of factors indicates they produce a synergistic effect with one of the indicated items. Osteoclasts are not listed next to other myeloid cells for reasons of visual representation. This figure illustrates relationships drawing on data from both human and on human and animal studies.

synthesis, as well as through increased catabolism, e.g. through TNF-dependent upregulation of matrix metalloproteinases (MMPs). IL-17A appears to potentiate ECM degradation quite potently, perhaps through additional pro-inflammatory mediators such as IL-6, IL-8, and G-CSF. Over time, these catabolic changes lead to joint erosion, subluxation, and ultimately destruction that are measured on clinical indices of RA radiographic severity.

Many pro-inflammatory cytokines such as IL-17 help induce angiogenesis. In the rheumatoid joint, these same processes result in angiogenesis within the rheumatoid synovium itself. This phenomenon is termed synovial hyperplasia and is measured by indices of disease activity and severity like those mentioned above. Stimulated by pro-inflammatory cytokines such as IL-17, fibroblast like synoviocytes (FLS) produce VEGF. Division of endothelial cells and infiltrating T and B cells arriving through this neovascularization contribute to the joint space narrowing observed on RA radiographs. These processes are represented visually in Figure 1. For the purposes of understanding the current studies, it is as important to catalog the individual cytokines mediating signal transduction as it is to observe the recurring tropes that arise repeatedly in RA pathology: feedforward loops, synergistic (non-additive) effects, and cellular cross-talk.

In our studies, one key result centers on the molecule glypican-5, which is encoded by the gene *GPC5*. Glypicans are components of heparin sulfate proteoglycans, which bind to the outer surface of the plasma membrane of cells, anchored by GPI. There are 6 glypican molecules in mammals, and several of these are clinically important. The main function of these proteins appears to be positive and negative regulation of key signaling cascades: Wnt, Hedgehog pathways, FGF and BMP pathways, etc. The best characterized of these is *GPC3*, an important oncogene the knockout of which leads to Simpson-Golabi-Behmel syndrome (OMIM #312870); an overgrowth syndrome resulting from loss of inhibition of the hedgehog signaling pathway. Both *GPC3* and *GPC5* bind important molecules highly expressed in T cells, which is discussed further below.

Pathogenic autoantibody production

The complexity of intercellular interactions leading to RA pathogenesis is also evident when considering pathogenic autoantibody production. There is a battery of RA risk genes important in follicular helper T cells (Tfh) that do not appear to replicate in either East Asian or African populations including the *IL2-IL21* locus as well as *CXCR5*. We identified *CXCR5* as nominally associated with RA severity in our studies (see Section III). However, *CXCR5* did not replicate as a risk locus for RA susceptibility in either East Asian or African-American populations (see section IV). Briefly, $CD4^+$ T cells express the chemokine receptor CXCR5, which enables migration from the T cell areas of secondary lymphoid organs into the B cell follicles following a CXCL13 gradient. There, these follicular helper T cells (Tfh) interact with B cells, providing them with activation and survival signals, such as CD40L and Interleukin 21 (IL-21), which are critical for the formation and maintenance of the germinal center (GC). B cells interacting with Tfh cells in the highly selective environment of the GC undergo extensive cycles of clonal expansion and somatic hypermutation, and ultimately differentiate into memory B and long-lived plasma cells. Human studies and preclinical data demonstrate that self-reactive Tfh cells expand and contribute to the pathogenesis of not only RA but several antibody antibody-mediated autoimmune diseases, most likely by favoring the development of auto-reactive plasma cells. T and B cell infiltration, T cell receptor (TCR) oligoclonality, and B-cell somatic hypermutation in the joints of RA patients indicate that antigen-driven events occur in the normally thin, delicate RA synovium. Supporting this, pathogenic autoantibodies in RA undergo affinity maturation as well, and these effects largely depend on help from Tfh cells. Recent studies have localized Tfh cells to RA synovial tissue, while such cells are absent from osteoarthritic and healthy tissues [15-19].

We detect strong, trans-ethnic support for the association of PAD2 and PAD4 in RA in all global populations. These proteins encode protein arginine deiminases that have distinct specificities for cellular substrates, which likely influences autoantigen selection and ultimately pathogenic autoantibody production in the context of RA. Both genetic and experimental data suggest the association of *PADI2* with RA is independent of that of *PADI4* [20, 21], a known risk allele for RA and a key enzyme in RA due to its role in citrullination and the generation of the ACPA response [22]. The contribution of PAD enzymes to the pathogenesis of RA is covered in detail in the discussion in Section IV.

Alteration to dynamics of T-cell activation

Engagement of the T cell receptor (TCR) by an MHC molecule bound to antigen kicks off a cascade of events that will determine the fate of a naïve T cell and its role in the body. Briefly, if this “primary” signal is followed by a co-stimulatory “secondary” signal, then the cell is said to become ‘activated’ and differentiate into an effector T cell of one type or another depending on the cytokine milieu. If the primary signal is not accompanied by a secondary signal, however, the T cell may become anergic – that is, enter a long-lasting hyporesponsive state. In reality, the situation is not black and white but is rather influenced by numerous factors that influence signal strength and type. For instance, number of MHC molecules engaged and the strength of the interactions affect the primary signal, just as outcome of co-stimulation depends on the number and type co-stimulatory or co-inhibitory molecules are activated and to what degree.

Once physiologic T-cell activation is understood as an output dependent on the ranges of several variables such as ligand concentration, spatial and temporal constraints, cell surface receptor affinity and expression, etc., it becomes easy to see how minor variation in the structure or expression of these molecules could confer risk of either of immunodeficiency, autoimmune disease, or both. Indeed, risk factors for RA and other autoimmune conditions cluster tightly in pathways relating to T-cell activation [23]. For example, ~13% of the 101 genes convincingly associated with RA [24] belong to the GO TCR Pathway (*TEC, CD2, LY9, PTPN22, PRPRC, CTLA4, RASGRP1, PTPN11, CD28, FCRL3, CD5, PRKCQ,* and *SH2B3*; $p = 1.37 \times 10^{-9}$) and another ~10% belong to the GO TCR Downstream Signaling Pathway (*PRKCQ, RL, IRF4, PTPN11, TNFRSF9, SH2B3, CD83, GATA3, CD28,* and *TRAF6*; $p = 3.07 \times 10^{-8}$). The GO BCR Signalling and Co-stimulatory signal during T-cell activation pathways are similarly enriched (http://cpdb.molgen.mpg.de/; date of access 4/19/2018). The importance of modulation of signal levels also suggests that one or more of these molecules might be a fruitful therapeutic target, and indeed the drug Abatacept (CTLA-Ig) is a mainstay of RA therapy.

RA as a genetic disease

Historical understanding of RA Heredity

Heritability is defined as the proportion of phenotypic variation accounted for by genotypic variation. The first observation that RA is in some degree heritable dates back at least 200 years to William Heberden, who asked if the disease “is [RA] not in some degree hereditary” as early as 1806 [25]. Thus, evidence that RA is at least partially genetically motivated predates the work of Johann Mendel. Since the observation of the familial clustering of RA evidently predated knowledge of even simple, autosomal dominant traits, it is understandable that a disease that clusters into families but defies specific rules that seem to govern simpler traits would generate substantial interest and confusion. Early attempts focused on studies of familial clustering, for instance that of Kroner, who published a study on a family having 4 generations of women with RA in 1928 [26].

Following this, in the mid and late 20th century, a variety of techniques were employed. However, because an account of such methods, including those based on kinship coefficients and identity by descent (IBD) as well as sibling recurrence risk ratios, twin studies, and parent-child trio studies can be found below (in Section II) we will defer discussion of these techniques. Also deferred are discussions of missing heritability, which are found in the same section. Here, it is most important to mention that estimates of the heritability of RA are usually around 50-60%. Estimates as low as 12% have been positied, with the difference attributed to shared environmental effects rather than genetics. Estimates of heritability of radiographic severity of RA are of a similar magnitude; for instance, a recent study of the inhabitants of Iceland arrived at an estimates of the heritability of joint destruction rate at 45% (using kinship coefficients), and 58% (based on IBD data) [27].

The first genetic association with RA was discovered in the 1970s [28]. This association was disambiguated in 1987 by Gregersen *et al.*, who clarified the allelic associations found in class II MHC loci that had remained opaque to understanding [29]. The key conceptual step was to grasp that gene conversion events could result in serologically distinct class II (specifically HLA-DR) alleles that nevertheless share short stretches of sequence – or epitopes. As such it became clear that conventional serological analysis would only incompletely correlate with disease, which opened the way for more accurate classifications based on DNA sequence and epitope conformation. These results led researchers ultimately to the recognition that the amino acid sequences QRRAA, RRRAA, and QKRAA in *HLA-DRB1* positions 70-74 define these so-called ‘shared-epitope’ alleles, but these results did not completely explain the association of the MHC region to RA. Further progress in pinpointing the genetic association of the HLA region to RA came in 2012, when Raychaudhuri *et al.* used conditional analysis and haplotype analysis coordinately to identify 5 amino acids in 3 HLA proteins that explain the vast majority of the association of that region to susceptibility to seropositive RA [30]. The striking finding was that all of these amino acids lay within the binding grooves of these proteins (*HLA-DRB1*, *HLA-B*, and *HLA-DPB1*), and indeed all of them mapped to positions between amino acid 9 and 13 within the groove. This finding suggests a clear biological rationale despite relying heavily on statistical and bioinformatics techniques. Taken together, the variants in these 3 HLA proteins account for roughly one-third of the heritable basis of seropositive RA, but a comparatively negligible portion of seronegative RA see also the Results of Section III, as well as prior studies [30].

Non-HLA genetic risk factors for RA

After identification of risk alleles in the HLA region, candidate-gene approaches led to successful identification of a handful of non-HLA risk genes, specifically, *PTPN22*, *PADI4*, *FCRL3*, *CD244*, and *CTLA4* [31]. These initial successes and the advent of cheapening genome-wide assays led to the advent of systematic exploration of the heritable basis of RA beginning around 2005. These technologies required some key analytical changes and innovations compared to family-based methods. Specifically, GWAS and GWA meta-analysis (GWAMA) either suppose individuals to be unrelated, or account for relatedness using mixed modelling approaches that account for the genotypic correlation structure between groups (approaches have been devised to account for relatedness at both an individual and a population level [32]). Typically, the former approach makes use of principal components analysis to control for population stratification, while the latter handles population structure by estimating the phenotypic covariance that is due to genetic similarity. This has been accomplished in a variety of ways e.g. through kernel regression frameworks [32].

Initial studies led to the identification of a growing number of RA risk loci. Following this, these studies began to be organized into meta-analyses, then trans-ethnic meta-analyses over approximately the next decade, culminating in a study of ~100,000 Europeans and Asians with RA in 2014 [18]. Because this latter study subsumes many of the prior results, we may summarize the literature by stating the key findings of Okada *et al.* 2014. Briefly, ~100 non-MHC risk loci combined explain only ~5.5% of the heritable basis of RA in Europeans and ~4.7% in Asians. Okada *et al.* constructed a trans-ethnic risk model based on the data from both ethnicities. This model explained roughly 80% of the known heritability in either population it was applied to, leading the investigators to conclude the heritable basis of RA is largely shared. We will return to this observation throughout Section IV and Section V. Nevertheless, it remains that the majority of the heritable basis of RA (and many other complex conditions) remains unexplained, and what little is explained is mostly attributable to a handful of variants in the HLA.

Current understanding of the heritable basis of RA

Genome-wide scans of susceptibility to RA and other diseases revealed a substantial amount about autoimmune risk architecture beyond the list of loci generated. Multiple studies leveraging several lines of evidence have suggested that autoimmune disease is an amalgam of a small number of coding variants (~10%) alongside a larger number of non-coding variants (~90%). In fact, several studies have suggested that 60% of autoimmune risk variants map to enhancer regions alone [33].

Statistical methods used to study the genetics of complex disease

This study makes use of several statistical methodologies. First, like many GWA studies of dichotomous traits, this study employs logistic regression, considering principal components and other key risk factors as covariates. However, unlike most other studies, we also employed zero-inflated negative binomial regression to model counts of swollen, tender and damaged joints in our study of the radiographic severity of RA.

Second we employ both a fixed effects meta-analysis (using METASOFT) of our African American datasets, and a random effects meta-analysis of all three global populations [34, 35]. Analyzing the data jointly increased our statistical power to detect both population-specific and trans-ethnic associations. Moreover, this analysis enabled us to generate M-values, which are akin to posterior probabilities that a variant of interest has a true association in a given study. Thus, these M-values gave us a consistent, quantitative way to adjudicate which of the variants are truly associated and which are not, and they are easily interpretable since, like other probabilities, they range between 0 and 1.

Using association summary statistics from our trans-ethnic meta-analysis, we conducted trans-ethnic fine-mapping. We selected CAVIARBF [36] and PAINTOR [37, 38] to complete this as these leverage the strength of our diverse data sets.

Contemporary goals in RA genetic research and the "post-GWAS" era

Efforts to organize even larger datasets into meta-analyses are ongoing. These strategies will no doubt be successful in identifying novel risk loci for RA, just as recent large meta-analyses for other traits have [39]. However, in our view the creation of ever larger datasets is primarily valuable insofar as it enables other thoughtful and creative applications beyond the use of single variant association testing to discover additional disease associations.

Broadly speaking, the post-GWAS era revolves around the use accumulated GWA data to empower the goals of precision medicine. Association summary statistics from a finalized, quality-controlled, well-powered GWAS are commonly the input in such an approach; a typical example might be a bioinformatic algorithm that integrates DNA microarray data (at the level of either genotypes or the association summary statistics) with other –omic datasets to predict disease genes or "causal" genetic variants. The trans-ethnic fine-mapping approaches employed in studies conducted Section III and Section IV are examples of this kind. Insofar as the goal of precision medicine is to enable individual-level, data-driven clinical decision-making, the goal of post-GWA research into complex disease can be understood as using aggregated –omic datasets to develop mechanistic understanding of genetic variants that influence disease.

Although many post-GWAS approaches use functional annotations, or ATAC-seq data, or single cell RNA-seq data, others use differences between ethnic populations to triangulate risk variants from non-risk. Most DNA microarray chips were designed by drawing on genetic variants found in European populations, and most GWAS have been conducted in European and Asian populations. But since differences between populations are a valuable source of information, addressing health disparities in genotyping of ethnic minorities in the United States and in Europe offers advantages not only in terms of distributive justice, but in its value to the scientific community and its clinical utility. In section IV we add ~2,500 RA patients of African American descent to ~100,000 Europeans and Asians with RA. While this addition results in the identification of only one new locus using standard frequentist association testing in a RA GWAS, we nearly double the number of candidate risk variants identified with high posterior probability. There are many other fascinating examples of ongoing post-GWAS projects that are relevant to RA genetics. I discuss several of these more fully in Section II, and I revisit this in Section V when discussing future applications of the current studies.

Motivation for the present studies

With the foregoing as context we mention several related motivations for the present research over and above the aims described in the Abstract and Overview sections of the dissertation, above. The first is to understand similarities and differences between RA risk in African populations in order that they can reap the rewards of precision medicine. Failure to do so constitutes a modern-day health disparity. Second, we integrate our findings with those from Europeans and Asians in order to improve results of post-GWAS algorithms for people of all populations. To accomplish this, we first integrate genotyping data from African Americans using both meta and mega-analysis. We then organize this into a joint trans-ethnic meta-analysis (TEMA). Finally, we use the results of this TEMA to prioritize candidate risk variants using trans-ethnic fine-mapping (TEFM).

INTEGRATIVE APPROACHES TO UNDERSTANDING THE ROLE OF PATHOGENIC VARIATION IN RHEUMATIC DISEASE

by

LAUFER VA, CHEN JY, LANGEFELD CD, AND BRIDGES SL JR.

In submission for *Rheumatic Disease Clinics of North America*

Format adapted for dissertation

Abstract

The use of high-throughput omics may help to understand the contribution of genetic variants to the pathogenesis of rheumatic diseases. We discuss the concept of missing heritability: that known genetic variants do not explain the heritability of rheumatoid arthritis and related rheumatologic conditions. In addition to an overview of how integrative data analysis can lead to novel insights into mechanisms of rheumatic diseases, we describe statistical approaches to prioritizing genetic variants for future functional analyses. We illustrate how analyses of large datasets provide hope for improved approaches to the diagnosis, treatment, and prevention of rheumatic diseases.

Key points

- Large genetic studies of rheumatic diseases have implicated many risk loci.
- Within risk loci, the identity and function of the pathogenic variants that underlie rheumatic diseases remain largely unknown, but methods in development will address these gaps in knowledge.
- Integrative analysis of omics datasets will yield new insights into the molecules, cells, tissues, and pathways that initiate and perpetuate rheumatic diseases.
- Functional characterization of prioritized genetic variants will pave the way for better diagnosis, treatment, and prevention of rheumatic diseases.

Introduction

The study of rheumatic diseases draws on many genome-scale technologies. Box 1 defines relevant terms that will be used in this discussion. Genome-wide association studies (GWAS) and other genetic studies have identified and replicated numerous loci associated with rheumatic diseases. Although these findings have led to increased awareness of particular pathogenetic pathways, there are multiple impediments to the translation of these results to the clinic. First, and as expected, the variants identified thus far do not account for the entirety of the heritable basis of any given rheumatic disease. Second, genetic variants in close physical proximity tend to be inherited together (linkage disequilibrium, or LD, see Box 1). As a result, a rheumatic disease risk locus usually contains multiple associated variants, from which the actual pathogenic variants are difficult to separate. This is most pronounced in the major histocompatibility complex region, where there are hundreds of associated variants, many of which are in strong LD. However, new techniques that leverage trans-ethnic and annotation data will help narrow the search for single-nucleotide polymorphisms (SNPs) that are directly pathogenic. Finally, determining the mechanisms of action of pathogenic variants is challenging, due to interaction effects, cell type–specific gene expression, the local tissue milieu, the temporal course of gene expression, and complicating environmental factors.

There is hope, however. Although rheumatic diseases are complex and have considerable differences in etiology, clinical presentation, and treatment, there is overlap in the pathogenic mechanisms involved. For example, pathobiology involving the adaptive immune system (e.g., autoantibodies) is similar among rheumatoid arthritis (RA), systemic lupus erythematosus (SLE), inflammatory myositis, and Sjögren syndrome. In these conditions, failure of adaptive immune cells (B and T lymphocytes) to maintain self-tolerance opens the way to several aspects of autoimmune pathogenesis, such as autoantibody production. These commonalities stem in part from genetic variants that affect multiple rheumatic diseases and similar conditions; for instance, dysregulation of autoantibody production characterizes patients with a risk variant in *PTPN22*, and this variant is associated with many rheumatic conditions, including RA, SLE, type 1 diabetes, and others [1]. Identifying such shared risk factors may provide insights into causes of rheumatic diseases. Furthermore, ongoing technological and bioinformatic advancements have enabled increasingly accurate and sensitive characterization of cells, tissues, organisms, and diseases through analyses of the genome, transcriptome, epigenome, proteome, and metabolome (see Box 1 for definitions). This review discusses how integration of data can help characterize and prioritize genetic variants for laboratory-based studies of their functional and biological consequences that will lead to better understanding of the mechanisms of human rheumatic diseases.

Table 1: Glossary of Key Terms

Term	Definition
5' untranslated region (5'-UTR)	The region directly upstream of the initiation codon and translation start site. In mRNA the sequence of this region strongly influences translation, and likewise the corresponding regions of template DNA contain many elements that can produce a marked effect on transcription.
ATAC-Seq	(Assay for Transposase-Accessible Chromatin with high throughput sequencing). This technique is used to study chromatin accessibility (accessible or protected), which is related to transcription factor binding and gene expression.
Copy number variation (CNV)	A form of genetic variation resulting in a change in the number of copies of a gene or genomic element. Deletion and insertion of DNA by a variety of mechanisms can produce genetic variants affecting as little as a few kilobases (kb) or as much as an entire chromosomes. CNVs have been difficult to assay using common technologies, affect a substantial portion of the genome, and influence a variety of diseases including rheumatic diseases.

CpG site

A DNA sequence consisting of a 5’ guanine nucleotide joined to a cytosine residue by a phosphate group. Cytosines in CpG sites can be methylated to form 5-methylcytosine, which can change its expression.

DNA methylation

Modification of DNA by attachment of a methyl group to DNA nucleotides. One common site of methylation is CpG sites (see definition above).

Epigenetics

The study of genetic effects produced by mechanisms that do not alter the primary sequence of DNA. For instance, methylation of DNA (above) or of histones producing differences in gene regulation are examples of epigenetic effects. Epigenetic modifications may result in changes to gene expression and regulation.

Extrinsic filtering

Data filtering based on information outside of the dataset, such as the inclusion of genomic annotations from the NIH Roadmap Epigenomics Mapping Consortium (ROADMAP) or the Encyclopedia of DNA elements (ENCODE).

Genome-wide association study (GWAS)

Examination of a genome-wide set of genetic variants (typically SNPs) to uncover associations between genotypic variation and a phenotype or trait. Similarly, epigenetic variation such as DNA methylation can be investigated in epigenome wide association studies.

Haplotype

A set of SNPs on the same DNA strand that are inherited together due to linkage to one another (below).

Heritability

The proportion of phenotypic variation that can be accounted for based on genotypic variation.

Imputation

Statistical inference of unobserved data, such as predicting the most likely allele of a particular SNP due to known LD/haplotype structure. Imputation methods are most well-established for genotyping data.

Intrinsic filtering

Filtering of data based on information calculated from the dataset itself, such as filtering genetic variants based on linkage to another variant strongly associated in that dataset.

Linkage disequilibrium (LD)

The non-random association of two or more genetic variants. Genetic recombination during meiosis allows for independent assortment of alleles and genetic variants. Genomic proximity, as well as forces like selection, population structure, and genetic drift, can maintain the association of two variants over a considerable period of time.

Long non-coding RNAs (lncRNA)

lncRNA are molecules of RNA greater than 200 nucleotides in length that do not code for protein products. These RNAs interact with several levels of gene specific transcription, splicing, translation, post-translational modification, and gene regulation. RNA-Seq studies have mostly targeted a genomic locus and having high depth can identify associated these lncRNAs for further analysis.

Mendelian randomization

An epidemiologic method in which genetic variation in genes of known function is used to examine whether a modifiable exposure has a causal effect relationship to disease in non-experimental studies. This method can be used to test for causal effects among two phenotypes (often an intermediate phenotype and a disease outcome) without conducting a randomized controlled trial.

Metaorganism

A community of organisms including the host and others that is indicated by the metagenome. The metagenome comprises the all genetic material associated with a human being including host DNA, microbial DNA, the virome, etc.

Multiple enhancer variant hypothesis

The hypothesis based on the observation that multiple variants in linkage may act cooperatively to regulate the expression of a target gene, and in diseases such as RA, SLE, and MS.

Non-additive genetic effects

Effects for which the contribution of alleles influencing a trait are not independent of one another, or not independent of the environment.

Metabolomics

The study of metabolites (small molecules left behind as part of specific cellular processes) within cells, fluids, or tissues or organisms. Collectively, these small molecules are referred to as the metabolome.

Pathogenic variant

A variant that contributes to the pathogenesis of a specified disease state. Such variants may also contribute to or protect against other phenotypes. Pathogenic variants need be neither necessary nor sufficient to produce a disease state due to incomplete penetrance.

Phased haplotype

With short read sequencing, it is uncertain whether variants are inherited from the maternal or paternal copy of a given chromosome. Algorithms have been devised to deduce phased haplotypes, or the most likely assignment of variants in a region to one or the other parental copy of a chromosome, enabling inference of haplotypes.

Phenome

The phenome refers to the set of all phenotypic states for a given biological unit of interest, such as an organism or population.

Polygenic traits

Traits influenced by genetic variation in several or many genes or genetic loci. Recent studies of rheumatic diseases suggest that thousands of genetic variants of small effect may modify disease risk.

Proteomics

Analysis of the full complement of proteins produced by a given biological entity of interest, such as a cell, tissue, or organism, including those modified through splicing or post-translational modification.

Quantitative trait locus

A genetic variant that is associated with a quantitative difference in the measurement of a phenotype or trait. For instance, an expression quantitative trait locus is a genetic variant correlated with expression level of either local genes (<5Mb; a cis-eQTL) or faraway genes (>5Mb; a trans-eQTL). The presence of a SNP that correlates with the methylation state of one or more genomic elements, such as nearby CpG sites is referred to as a methylQTL or meQTL.

RNA-Seq

A next generation sequencing technology that allows quantitative profiling of the transcriptome (identifying the presence and amount of messenger RNA in a sample of cells, tissues, etc.).

Single nucleotide polymorphism (SNP)

A DNA sequence variation affecting only one nucleotide, typically present in at least 1% of a given population. For instance, in the hypothetical sequence AGT(C)TA, the substitution of cytosine by thymine resulting in a sequence of AGT(T)TA would define a SNP.

Structural variation (SV)

Large-scale DNA sequence variants. Copy number variants (above) producing deletion or duplication of a genomic segment are structural variants, as are genomic rearrangements not resulting in a gain or loss of genetic material such as an inversion or translocation.

Transcriptomics	Study the set of all RNA transcripts produced by the genome, usually studied in particular tissues or organs (e.g. blood), or cell types (e.g. CD4+ T lymphocytes).

Heritability of rheumatic diseases

Historical evidence for the heritability of rheumatic diseases comes from studies of familial clustering, sibling recurrence risk ratios, twin studies, and parent-child trio studies [2]. More recently, a large number of advanced methodologies based on genome-wide assays for estimating heritability have been devised [3, 4, 5]. Heritability estimates for rheumatic diseases are often approximately 0.5, [2] but this is highly variable. GWAS (see Box 1) conducted to date have identified hundreds of risk loci for autoimmune diseases and thousands of associations with disease and traits [1, 2].

Although in aggregate these studies explain a meaningful proportion of disease risk, much of the heritable basis of rheumatic disease remains unexplained. There are many possible explanations for this problem, which is referred to as "missing heritability" [6] (Box 2). In some cases, it is possible that the estimate of heritability is inflated. Recent studies of the heritability of RA reported only 12% of phenotypic variance in the susceptibility to RA due to additive genetic effects, [7] whereas typical estimates from previous studies ranged between 50% and 60% [8]. This study instead found a 50% contribution from shared environmental effects and 38% from nonshared environmental effects [7].

Possible reasons for missing heritability in large-scale genomic studies

Alternatively, methodological factors may be implicated. For example, a recent GWAS of RA identified thousands of variants, which individually do not meet the threshold of association, but collectively constitute a substantial fraction (~20%) of disease risk. [9] Cohort design, study design, and disease definition also may contribute to this problem. For instance, trans-ethnic meta-analyses of a disease performed by many groups in collaboration often include different disease subtypes or include patients using different diagnostic criteria, leading to failure to capture available heritability through inclusion of heterogeneous subtypes of patients.

Other studies indicate that some missing heritability may reside in genetic variants not readily identifiable with current technologies. For instance, the genetic association of *FCGR3B* with SLE may be caused by structural variation (SV) (see Box 1), [10] although to our knowledge there are no conclusive data that a pathogenic SV produces a specific rheumatic disease association indexed in the National Human Genome Research Institute GWAS Catalog. [1] Advances in technology, such as the ability to infer "phased haplotypes" (see Box 1) from genomic DNA using long-read next-generation sequencing (NGS) [11, 12] technology platforms, could enable better identification of contributions of SVs and haplotypes to missing heritability than current NGS platforms.

Many recent functional studies have shown that rare variants (genetic variants having an allele frequency of <5%) contribute to a spectrum of phenotypes, including rheumatic diseases. [13, 14] Several studies have sought to quantify the contribution of rare variants to missing heritability of complex diseases. For instance, 1 large exome-sequencing study examined risk loci from 6 autoimmune diseases and found that rare variants contributed less than 3% of the heritability explained by common variants at known risk loci. [15] However, more recent studies have found that most autoimmune risk variants lie in noncoding elements (~90%), greatly limiting the value of an exome-based approach for autoimmune applications. Indeed, until very large, high-depth, whole-genome sequencing studies are widely available, quantifying the contribution of rare variants in rheumatic disease will remain problematic. Further description of likely sources of missing heritability is provided in Box 2.

Table 2: Factors accounting for Missing Heritability in Rheumatic diseases

Factor contributing to Missing Heritability	Explanation
Polygenicity and non-additive genetic effects	Non-additive genetic effects (see Box 1 for definition) are not well measured by traditional methods of estimating heritability, and therefore might represent sources of missing heritability. For instance, haplotypes of common SNPs could explain a fraction of the missing heritability, which could be related to epistasis (the interaction of genes which changes their effect) or better tagging of pathogenetic variants. Recent studies of complex diseases suggest that thousands of variants may each contribute a small fraction of disease risk. However, the contributions of marginally associated variants are often not included in the heritability accounted for in a given study. Thus, estimates of heritability based only on highly significant SNPs would not include such effects, resulting in inability to account for disease heritability. This appears to be a major source of missing heritability.

Rare variants and structural variants	Many kinds of genetic variants (e.g. short insertions and deletions), structural variants (e.g. copy number variants), and rare variants are not well-assayed by current genotyping arrays. There are several reports of structural variant association with rheumatic diseases, but these associations have generally been difficult to reproduce. Current NGS technology can capture information on these variants, but with much lower accuracy than on common variants, making their contribution difficult to assess. Despite this limited accuracy, several large sequencing studies have been performed with the goal of identifying rare variants implicated in autoimmune disease, and often conclude they do not account for a substantial fraction of missing heritability. It will be necessary to obtain very large studies of long-read sequencing data in order to accurately assess the contribution of rare and structural variants.
Inflated heritability estimates	If heritability is overestimated, then the amount of missing heritability will also be high. It is possible that many estimates in the literature misattribute environmental effects on disease risk as genetic liability. Though debate continues, many experts do not expect inflated heritability estimates to be a major contributor to the problem of missing heritability.
Epigenetic effects	Because some forms of epigenetic variation are inherited, phenotypic variance incorrectly attributed to genetic rather than epigenetic mechanisms could produce artificially high heritability estimates. Such effects may account for a moderate or large portion of missing heritability.
Biotechnology effects	The use of different platforms and technologies, superimposed on other effects, may lead to errors in heritability estimates. With sound analytical practices, such effects on estimation of missing heritability should be minor.

High-throughput omics approaches that can be integrated with genetic data to understand rheumatic diseases

RNA-seq

Advances in biotechnology have led to high-throughput studies of gene expression (the transcriptome) that can lend insight into the pathogenesis of rheumatic diseases. RNA-Seq (see Box 1) is a method of interrogating the transcriptome that uses NGS to identify and quantify RNA transcripts, and offers several advantages over array-based technologies [16]. Notably, this includes the ability to identify allelic imbalance, to quantify gene expression in a transcript-specific manner, and to capture unexpected alternative splicing, truncation, and post-transcriptional modification events [17]. RNA-Seq has been used to perform biomarker discovery in peripheral blood monocytes in RA, and to study differential expression in synovial fibroblasts [18] in RA and monocytes in SLE [19]. Focused analyses of a single locus using RNA-Seq can provide a detailed picture of mRNA and noncoding RNA [20]. A recent study of the *TRAF1-C5* locus revealed a long noncoding RNA (lncRNA) (see Box 1) that influences C5 levels in RA [21]. In SLE, single-gene profiling of *IRF5* was performed to assess the well-known population-specific diversity and genetic associations in the locus. Notably, this study identified 14 new differentially spliced *IRF5* transcript variants and found that one of the risk haplotypes for SLE is among the most abundant transcripts produced in the disease [22]. RNA-Seq has also been used to study microRNA in the salivary glands of patients with SS [23] and to investigate gene regulation in RA as a part of an integrative bioinformatics approach [24].

Typically, RNA-Seq measures a bulk sample of cells of a given type. However, individual cells isolated from samples of whole blood are frequently in different states producing different amounts of transcript [25]. Single-cell RNA-Seq may detect differences in transcript splicing or transcript isoform expression between cells that are lost on aggregation even among seemingly phenotypically similar cells isolated from the same tissue [25, 26]. This rapidly maturing technology has been used to analyze expanded CD4$^+$ T cells in the peripheral blood and synovium of patients with RA [27]. This study revealed that skewing of phenotypes of expanded CD4+ T-cell clones is likely due to nonspecific expansion of naïve and memory T-cell subsets. RNA-Seq is also well-suited to studying regulatory variants found in rheumatic diseases [28].

Expression quantitative trait loci

Integration of genetic/genomic data with expression data has provided important insights. SNPs (see Box 1) that influence gene expression are called expression quantitative trait loci (eQTL); these may affect the expression of nearby genes (cis-eQTLs) or distant genes (trans-eQTLs) (see QTL in Box 1). eQTLs are enriched among suspected pathogenic variants in autoimmune risk loci. A recent study used eQTLs to quantify the contribution of gene expression to heritability and found that, on average, 21% of disease heritability was attributable to the cis-genetic component of gene expression levels for many complex phenotypes, including rheumatic diseases [29]. In general, cis-eQTLs are more commonly associated with complex disease and tend to have a greater impact on gene expression compared with trans-eQTLs [30].
Although there is much evidence that eQTLs are important in rheumatic diseases, relatively few variants have well-characterized pathogenic effects. rs140490, a cis-eQTL associated with SLE, is one such example [2, 31]. rs140490 is just upstream of the 5′ untranslated region (see Box 1) of *UBE2L3*, and is associated with increased expression and translation of *UBE2L3*, probably through diminished degradation of the nuclear factor (NF)-κB inhibitor-α (IκBα) [32]. The resulting activation of NF-κB leads to increased B-cell survival in both healthy controls and patients with SLE. Patients with

SLE with the risk genotype have elevated *UBE2L3* within proliferating and activated B cells, as well as increased counts of antibody-producing plasmablasts. This example demonstrates the complexity of characterizing pathogenic variants in SLE: the variant is located in a regulatory region affecting only some disease-relevant cell types (B cells and monocytes) and functions in a temporally variable and an activation-state–dependent manner.

Recent studies have integrated genetic and eQTL data on a massive scale. An approach called summary data–based Mendelian randomization (see Box 1) used genome-wide genetic data from more than 338,000 people and eQTL data from more than 5,000 people to link genes in 126 risk loci to 5 different phenotypes, including RA. [33] This finding highlights the ability of this technique to identify novel disease associations. Methods of imputation (see Box 1) of gene expression based on reference panels are being devised. Such advancement will allow identification of expression-trait associations of small effect [34] while avoiding the high cost of obtaining eQTL data.

Epigenetics/Epigenomics

Several major classes of epigenetic regulation are relevant to human disease, [35] and are assayed using a growing number of technologies. [36] Analogous to eQTLs, [37] methylQTLs (meQTLs; see Quantitative trait locus in Box 1) are CpG sites the methylation state of which correlates with a genetic variant. Such methylated DNA impedes transcriptional proteins and increases affinity for proteins that alter DNA accessibility. Thus, cellular outcomes, such as transcription and cell fate, are altered in ways that can impact rheumatic disease. For example, in a study of SLE among twins, DNA methylation differences are associated with twin discordance, [38] and a study of 24 patients with SLE and controls showed evidence of differential DNA methylation in CD4 + T cells. [39] Many meQTLs are found in tight LD with risk variants for rheumatic diseases, [40] although the possibility that both independently affect transcription cannot be excluded. Other epigenetic assays, such as ATAC-Seq (see Box 1), have been used to investigate the correlation of nucleosome modifications with functional elements genome-wide, for instance in naïve B cells in SLE, [41] whereas DNAse-seq pairs DNAse hypersensitivity methodology with deep sequencing to identify portions of the genome accessible to transcriptional machinery and transcription factors. One methodology integrating epigenomic and other assays, RASQUAL (Robust Allele Specific QUAntification and quality controL), has been applied to autoimmune risk loci and is described in further detail later in this article.

Epigenetic factors likely contribute to the heritability of rheumatic diseases, [42] and may alter heritability estimates derived from genetic data. Because epigenetic modifications can be inherited across generations, but are not assayed by genotyping chips or by whole genome sequencing [43], it is difficult to exclude the possibility that epigenetic alterations could account for a proportion of disease risk normally attributed to genetics. This assertion gains further support from studies suggesting that such alterations affect human disease phenotypes [44]. In SLE, there is an association between DNA methylation patterns and twin discordance [38]. Several mechanisms, including imprinting, incomplete erasure of DNA methylation, and persistence of histone markings [42, 43, 45] could produce such an effect. Nevertheless, at this time, estimates of epigenetic missing heredity are not widespread for complex diseases.

Proteomics and Metabolomics

Proteomic studies typically profile a tissue or fluid important to the disease process, such as synovial fluid in RA, often with the goal of finding biomarkers to enable early diagnosis or identification of common pathways. [46] Other common applications in RA include monitoring disease activity, disease severity, and treatment efficacy. [47] Importantly, levels of a particular protein (for example in serum, plasma, or synovial fluid) cannot be inferred from gene expression data [48], due in part to the presence of uncoupled posttranslational protein regulatory mechanisms [49], and in part to the aggregation of proteins from various nearby tissues and bodily fluids. Advances in mass spectrometry [3] have enabled proteomic assays to evolve beyond simple catalogs of protein abundance by capturing the rates of protein synthesis, degradation, and turnover. Incorporation of subcellular localization and tissue abundance of the proteins being studied [34] adds a further dimensionality to these increasingly rich datasets. Therefore, proteomic technology is becoming more useful in determining biological effects (e.g., increased protein levels) of upstream events (e.g., gene expression levels influenced by genetic variants and epigenetic factors) in rheumatic diseases.

Informatics-based approaches to characterizing protein interaction networks have also been useful to understand rheumatic diseases. Current evidence suggests genetic risk variants for rheumatic diseases are organized in pathways, physically interact with one another, [50] and are enriched for protein-protein interaction network modules. [51] Recognition of this pattern in complex diseases has already led to identification of drug targets, [52] and has been used to prioritize gene relevancy within risk loci. [53, 54] Such interaction network modeling techniques have been used to perform functional characterization of pathogenic variants among genomic, transcriptomic, and proteomic data. [55]

Techniques used in large-scale metabolomics studies include proton nuclear magnetic resonance and mass spectrometry. These have been used to profile small molecules in rheumatic diseases and have been reviewed in detail recently. [56] Integration of the metabolome with GWAS data has enabled detection of genetic variation that affects metabolite levels (referred to as metabolome QTLs) in several organisms [57, 58, 59, 60], but continued improvements to modeling and methodology are needed before application to human rheumatic diseases becomes widespread.

Cell-Specific and Tissue-Specific Gene Expression: Influence on Integrated Multi-Omic Analysis of Rheumatic Diseases

One crucial consideration in the interpretation of transcriptomic, epigenomic, and proteomic analyses is the cell-based or tissue-based specificity of expression of variants that may exert pathogenic effects. Consortia such as Encyclopedia of DNA elements (ENCODE) provide a catalog of functional elements across the human genome.[61] More recently, the National Institutes of Health Roadmap Epigenomics Mapping Consortium provided a publically available catalog of methylation, histone modification, chromatin accessibility, and other data. [62] Drawing heavily on such datasets, a recent study fine mapped pathogenic autoimmune disease variants. There was markedly different enrichment in acetylation of cis-regulatory elements of 33 different cell types across 39 autoimmune diseases and related traits. [30] The investigators of another variant prioritization methodology provide more than 8000 genome-wide annotations to aid investigators in study of risk loci. [63] Thus, although it can be daunting to isolate the effect of a genetic variant to the appropriate tissue, the increasingly comprehensive annotation of the human genome can aid investigators in selecting the appropriate tissues and focusing hypotheses.

One recent study combined an impressive array of publically available data and tools for analysis of microRNA, transcription factor binding sites, epigenetic data, data from ENCODE, and chromatin immunoprecipitation data. By analyzing these aggregated data, the investigators were able to frame and test the hypothesis that a variant in the autoimmune risk locus *ETS1* increases pSTAT1 binding and decreases *ETS1* expression in Asian individuals, but not other populations, with SLE. [64] Thus, tissue-specific effects are often also subject to additional complicating factors, such as activation state dependency, trans-ethnic differences, and temporal variability, necessitating careful design of follow-up functional studies. The emergence of large datasets to provide reference points that can be used to interpret data from cells and tissues from diseases will be critical to these future studies.

Approaches to data integration

Integration of high-throughput data for analysis of rheumatic diseases is a complicated topic reviewed in detail elsewhere. [65] We include a brief summary of 2 types of techniques (multistage analysis and metadimensional analysis) and 2 forms of data filtering (intrinsic and extrinsic) to provide helpful context. Intrinsic data filtering uses information from the dataset itself, such as filtering genetic variants based on linkage to other variants of interest in that dataset. Extrinsic data filtering is based on information outside of the dataset, such as the inclusion of genomic annotations from separate studies such as ENCODE. Currently, both intrinsic and extrinsic data filtering are essential for efficient characterization of complex disease genetics.

Multistage Analysis

Multistage analysis sequentially examines relationships between each dataset and the other datasets, and also between each dataset and the trait. For instance, the correlation of a genetic variant with gene expression level is performed in 1 analytical step, then correlation of such a variant or gene expression level with a disease state like RA is performed in a subsequent step. Such designs are common in rheumatic disease genetic research, such as the analysis of eQTLs, which includes analysis of genetic variants (e.g., SNPs) and gene expression levels.

Multistage analyses have much greater power to detect the effect of a single SNP on gene expression than several weak independent effects that together lead to important changes in gene expression [65, 66, 67]. Although multistage analyses are useful, they also have limitations. In rheumatic disease research, the combined effects of multiple variants on gene regulation may be critical (the multiple enhancer variant hypothesis), therefore making them difficult to discover in multistage analysis. [68]

Metadimensional Analysis

For complex multivariate data sets from different platforms, metadimensional analysis may perform well [57]. This technique takes advantage of simultaneous combination of multiple data types into a single search space to construct a final model.

Metadimensional analysis can use 3 types of integration strategies; that is, to combine raw or processed data sets by directly "concatenation" of them before modeling and analysis, to perform data mapping or "transformation" first before modeling and analysis, and to model the data independently before merging all models toward the final analysis. Metadimensional analysis approaches can draw on a variety of techniques, such as regression trees [69], Bayesian networks [70], and evolutionary computation [71]. These algorithms may be tuned to search for known conventional relationships, or could be relaxed to search for new or unexpected complex relationships, but are often computationally intensive.

Selected Insights Gleaned from Integrative Analyses

In the following section, we describe several examples in which integrative analyses of large datasets have been used to provide novel insights into the pathogenesis of rheumatic diseases. For example, a recent study noted genetic and epigenetic interactions affect the expression of the gene *LBH* and potentially risk for diseases such as RA, SLE, and celiac disease [72]. Making use of intrinsic and extrinsic data filtering, this study integrated GWAS, gene expression, and DNA methylation data in RA with publically available data from the ENCODE project. Using reporter constructs methylated in vitro and transfected into synovial fibroblasts, the investigators showed that the RA-associated SNP in *LBH* decreased *LBH* transcription. This study illustrates how confluence of data from multiple genomic assays, specifically a GWAS risk variant for RA, a differentially methylated locus, and open DNA regulatory elements, can aid in characterization of RA pathobiology by showing how a functional SNP and a differentially methylated enhancer regulate aggressiveness of RA fibroblast-like synoviocyte(s).

Another recent study presented a method to integrate transcriptomic data and epigenetic data in a highly novel way. The RASQUAL method [73] was used to combine expression and chromatin conformation data (from ATAC-Seq) and led to significant findings within RA risk loci. Strikingly, an SNP identified by previous GWAS of RA, rs909685 in *SYNGR1*, may act to alter gene expression by altering chromatin structure and accessibility. The finding that a genetic variant affects the 3-dimensional conformation of DNA and histone folding illustrates the utility of integrative analysis by offering an example of how these methods naturally provide a springboard for future studies compared with association analysis alone.

Integrated proteomics approaches are becoming more common as well. The COMBINE (Controlling chronic inflammatory diseases with combined efforts) study integrated DNA, RNA, flow cytometry, and proteomic data to predict clinical response to tumor necrosis factor (TNF) inhibitors (TNFi) in RA [74]. Specifically, results from commercial protein biomarker panels, DNA microarrays, and RNA-Seq were filtered based on publically available datasets on TNFi responsiveness. Measurements of these biomarkers, genetic variants, and expression levels were then fit into a linear regression model with treatment response 3 months after initiation of TNFi as the primary outcome. This approach replicated 11 biomarkers for anti-TNF treatment in RA and successfully combined multiple levels of omics data into a predictive model with a sensitivity of 73% and a specificity of 78%. Studies such as these are early indicators of the potential of integrative approaches in precision medicine.

Prioritization of genomic variants/pathways for functional analysis

One of the problems arising from the analyses of large numbers of patients with rheumatic diseases using high-throughput methodologies is the generation of a very large number of candidate risk loci to be examined. The expense and logistics of analyzing a large number of loci is daunting, and there is a need for systematic, rational, and biology-based approaches to identify variants with the highest likelihood of clinically relevant effects. Numerous methodologically distinct approaches to variant prioritization have been developed and several have been applied to rheumatic diseases. Prioritization approaches using intrinsic or extrinsic data filtering, or both (see Box 1) are routinely used, and leverage enrichment of variants bearing certain functional annotations [63, 75, 76], trans-ethnic differences in genetic variation [63, 77], and association strength of genetic variants [30, 63, 76, 77, 78]. Other tools, such as OMIM Explorer, integrate high-dimensional clinical phenotyping data with genotype information, and offer powerful frameworks for variant prioritization as well [79]. Due to the number, complexity, and size of the datasets used to filter variants, we anticipate these increasingly sophisticated methods will be critical to guiding optimal variant prioritization based on empirical classifiers. Here, we discuss a few variant prioritization tools that have been calibrated on, or applied to, rheumatic diseases. Many other variant prioritization tools are available and under development but are not discussed in this review [75, 79, 80, 81].

The Probabilistically Identified Causal Single-Nucleotide Polymorphism Algorithm

A study of dense genotyping of a large number of patients with different autoimmune diseases and controls was used to develop an algorithm called

Probabilistically Identified Causal SNPs (PICS) [30]. This algorithm estimates the likelihood that each variant is pathogenic, using the strength of association and LD values of variants in a locus. Application of this algorithm to genetic data from 21 autoimmune diseases, including Sjögren syndrome, RA, SLE, and seronegative spondyloarthritis (SNSA), resulted in identification of approximately 9000 candidate causal variants called PICS. Nearly 90% of these “autoimmune” PICS mapped outside of coding regions, and 60% to immune-cell enhancers. Strikingly, these PICS tended to be near to, but outside of, canonical binding sites of regulators of immune differentiation, in less well-characterized regions of the enhancer. Therefore, in addition to providing a prioritization tool, and identifying a large list of candidate SNPs, this study is significant for its suggestion that current gene regulatory models may be incomplete.

The Probabilistic Annotation INTegratOR Algorithm

The Probabilistic Annotation INTegratOR (PAINTOR) algorithm is an open-source fine-mapping program that integrates association summary statistics (Z-scores) from GWAS, LD scores, and functional annotation information. It was developed to model the likelihood of causality of 1 or more SNPs in a risk locus. One of the strengths of this algorithm is that it can leverage trans-ethnic studies to better prioritize variants. PAINTOR2 was recently used to perform a meta-analysis of a large genetic dataset from RA [53]. The algorithm assigned a very high posterior probability of causality to rs2476601, a missense variant in *PTPN22*. Given that the effects of rs2476601 are relatively well studied and that it underlies risk of many autoimmune and rheumatic conditions, this finding might be regarded as a positive control. Intriguingly, it also assigned a very high posterior probability to variants in 4 other RA risk loci, *ANKRD55*, *TNFRSF14*, *UBASH3A*, and *TYK2*, the functional roles of which are not well-established. Identification of variants such as these increases cost efficacy by reducing the number of likely candidates (the credible set), thereby increasing the likelihood of studying a pathogenic variant. PAINTOR 3.0, the most recently released version of the software, extends the PAINTOR framework to multiple traits across transethnic studies, and is capable of modeling 1 or more causal variants per locus (http://bogdan.bioinformatics.ucla.edu/2016/11/03/paintor-3-0/, accessed December 26, 2016). PAINTOR uses functional annotation as input to enable better prioritization of candidate variants, or output enrichment of candidate causal variants within functional classes [63].

A distinct Bayesian approach to accomplish the latter goal was recently described [82]. It uses association statistics computed across the genome to identify classes of genomic elements that are enriched with (or depleted of) loci influencing a trait. Thus, this approach incorporates internal filtering to make inferences about the relative importance of annotation data. Reweighting each GWAS by using information from functional genomics increased the number of loci with high-confidence associations by approximately 5%.

The Molecular Interaction Network-Based Ranking Algorithm

There is a need for developing systems biology approaches to integrate comprehensive genetic information and provide new insight on complex disease biology. We took such an approach [54] to study type 2 diabetes (T2D); however, the method is readily applicable to the study of other rheumatic diseases, such as RA. The method works by bringing in protein-protein interaction data to construct a disease-specific molecular interaction network, which consists of disease-specific genetic risk genes and all their direct interacting gene partners. Then, network centrality measures using "network topological features," such as hubs and clustering coefficients are used to rank genes from the network or network modules. These genes can be further ranked based on additional GWAS association hazard ratio data and related pathway enrichment and gene set enrichments results. We found that *PI3KR1*, *ESR1*, and *ENPP1* were the interconnected T2D disease network "hub" genes most strongly associated to T2D genetic risks [54]. Contrary to expectations, the well-characterized gene *TCF7L2* was not among the highest-ranked genes in the T2D gene list. However, many highly relevant pathways were reaffirmed from the integrated data sets, including pathways involved in insulin signaling, T2D, mature-onset diabetes, adipocytokine signaling pathways, and cancer-related pathways. Similar pathway and network analysis approaches based on this framework [83] are critical for improving interpretations of genetic variations and genetic risk factors. These approaches may facilitate attribution of complex disease genetic risk to the summative genetic effects of many genes involved in a broad range of signaling pathways and functional networks.

Methods for exploration of relationships between clinical phenotypes

Given the overlap between autoimmune diseases, techniques that can be used to study relationships between phenotypes are of intense interest. The principles of Mendelian randomization (see Box 1) can be applied in a variety of ways to make inferences about environmental determinants of disease, among other applications [84]. An innovative algorithm was used to examine 43 GWAS on 42 human traits to identify pairs of traits sharing association of multiple (common; minor allele frequency >5%) genetic variants [85]. This analysis found that variants that increase risk of coronary artery disease (CAD) tend to decrease risk of RA, whereas variants affecting RA appear to have little effect on CAD risk. This can be interpreted as evidence of a causal link between CAD and RA, but this result could not be confirmed in a larger study despite other successes of the algorithm [85]. The results obtained in this detailed study generally agree with data from randomized controlled trials and Mendelian randomization studies [85]. Importantly, inferred causal relationships obtained from Mendelian randomization frameworks such as these may be used even if the studies paired share no common subjects. Therefore, given the paucity of high-quality comprehensive phenomic datasets [85, 86, 87], these methodologies are exceedingly valuable because they can provide information that would otherwise be available only through costly trials. Equally important are the implications to the paradigm of personalized medicine: if pleiotropic effects are widespread in the phenome, then even a targeted intervention aimed at a single pathogenic variant is likely to affect other phenotypes. Alternatively, if legal and organizational obstacles can be overcome, high-dimensional phenomic data may ultimately become available through the electronic medical record, curated by clinical centers, enabling exploration of the extent of pleiotropy and genome-phenome interactions. Early steps toward this goal have been made in the form of phenome-wide association studies [88].

Functional validation of pathogenic regulatory variation in complex disease

We have argued that the integration of multiple omics technologies can be used to frame specific hypotheses and design experiments to test them. However, even when such analysis is done well, functional validation of findings from -omic assays remains a crucial limiting step to advancement of our understanding. This may be particularly difficult if the distribution and concentration of multiple genetic variants in noncoding elements, such as enhancers, is critical to autoimmune disease risk, as is currently expected [68]. Nevertheless, there are several promising technologies that are currently used for functional validation of regulatory variants, and that are potentially scalable. The creation of multiple distinct Cas9 mutant enzymes facilitates different functions, such as gene silencing, gene activation, or site-specific DNA recognition and cleavage and enables study of complex disease variants [89]. Certain Cas9 mutants can introduce precise mutations or knock-ins such as those found in immune enhancer regions. A recent review covers developments in CRISPR/Cas9 relevant for rheumatologists [90]. Combining reporter assays with DNA synthesis [91], DNAse-seq [92], and barcoding [93] has substantially increased throughput of these assays, even allowing massively parallel interrogation of regulatory variants in human cells [94, 95]. RNAi-based screens, which could recapitulate loss-of-function analyses produced by pathogenic risk variants in regulatory regions in vivo [96], also could be used to study rheumatic diseases. Continued refinement of these technologies and others may eventually prove commensurate to the challenge presented by integrated omics data; namely, understanding the context and biologic roles of thousands of pathogenic risk variants acting in concert to produce rheumatic disease.

Summary

In summary, we begin by providing an overview of heritability of rheumatic diseases and describing potential explanations for why much of genetic risk remains unknown. We then highlight how the use of high-throughput omics approaches, such as RNA-Seq, expression (and other forms) of QTLs, epigenetics, and proteomics can help understand the genetic basis for the pathogenesis of rheumatic diseases. Having outlined several genomic technologies, we then describe approaches to integrating multiple forms of data, including multistage and metadimensional analytical designs (using extrinsic and intrinsic data filtering) and provide specific examples of novel insights into the mechanisms of rheumatic diseases that these analyses provide. We also describe statistical approaches to prioritizing genomic variants for functional analysis. We provide examples showing that these integrative analytical approaches are valuable because they are better at providing context necessary for researchers to frame targeted hypotheses. Overall, we believe coordinated study of human biology alongside programs to analyze large datasets in detail raise hope for better approaches to the diagnosis, treatment, and prevention of complex conditions such as the rheumatic diseases.

References

1. Welter D, MacArthur J, Morales J, et al. *The NHGRI GWAS catalog, a curated resource of SNP-trait associations*. Nucleic Acids Res 2014;42:D1001–6.

2. Gutierrez-Arcelus M, Rich SS, Raychaudhuri S. *Autoimmune diseases—connecting risk alleles with molecular traits of the immune system*. Nat Rev Genet 2016; 17:160–74.

3. Visscher PM, Macgregor S, Benyamin B, et al. *Genome partitioning of genetic variation for height from 11,214 sibling pairs*. Am J Hum Genet 2007;81:1104–10.

4. Vinkhuyzen AA, Wray NR, Yang J, et al. *Estimation and partition of heritability in human populations using whole-genome analysis methods*. Ann Rev Genet 2013; 47:75–95.

5. Visscher PM, Goddard ME. *A general unified framework to assess the sampling variance of heritability estimates using pedigree or marker-based relationships.* Genetics 2015;199:223–32.

6. Manolio TA, Collins FS, Cox NJ, et al. *Finding the missing heritability of complex diseases*. Nature 2009;461:747–53.

7. Svendsen AJ, Kyvik KO, Houen G, et al. *On the origin of rheumatoid arthritis: the impact of environment and genes–a population based twin study*. PLoS One 2013;8(2):e57304.

8. Kurko J, Besenyei T, Laki J, et al. *Genetics of rheumatoid arthritis - a comprehen- sive review*. Clin Rev Allergy Immunol 2013;45(2):170–9.

9. Stahl EA, Wegmann D, Trynka G, et al. *Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis*. Nat Genet 2012;44:483–9.

10. Ptacek T, Li X, Kelley JM, et al. *Copy number variants in genetic susceptibility and severity of systemic lupus erythematosus*. Cytogenet Genome Res 2008;123: 142–7.

11. Chaisson MJ, Huddleston J, Dennis MY, et al. *Resolving the complexity of the human genome using single-molecule sequencing.* Nature 2015;517:608–11.

12. Zheng GX, Lau BT, Schnall-Levin M, et al. *Haplotyping germline and cancer genomes with high-throughput linked-read sequencing.* Nat Biotechnol 2016;34: 303–11.

13. Rieux-Laucat F, Casanova JL. Immunology. *Autoimmunity by haploinsufficiency.* Science 2014;345:1560–1.

14. Rice GI, del Toro Duany Y, Jenkinson EM, et al. *Gain-of-function mutations in IFIH1 cause a spectrum of human disease phenotypes associated with upregulated type I interferon signaling*. Nat Genet 2014;46:503–9.

15. Hunt KA, Mistry V, Bockett NA, et al. *Negligible impact of rare autoimmune-locus coding-region variants on missing heritability*. Nature 2013;498:232–5.

16. Giannopoulou EG, Elemento O, Ivashkiv LB. *Use of RNA sequencing to evaluate rheumatic disease patients*. Arthritis Res Ther 2015;17:167.

17. Maher CA, Kumar-Sinha C, Cao X, et al. *Transcriptome sequencing to detect gene fusions in cancer.* Nature 2009;458:97–101.

18. Heruth DP, Gibson M, Grigoryev DN, et al. *RNA-seq analysis of synovial fibroblasts brings new insights into rheumatoid arthritis*. Cell Biosci 2012;2:43.

19. Shi L, Zhang Z, Yu AM, et al. *The SLE transcriptome exhibits evidence of chronic endotoxin exposure and has widespread dysregulation of non-coding and coding RNAs*. PLoS One 2014;9:e93846.

20. Clark MB, Mercer TR, Bussotti G, et al. *Quantitative gene profiling of long non-coding RNAs with targeted RNA sequencing.* Nat Methods 2015;12:339–42.

21. Messemaker TC, Frank-Bertoncelj M, Marques RB, et al. *A novel long non-coding RNA in the rheumatoid arthritis risk locus TRAF1-C5 influences C5 mRNA levels.* Genes Immun 2016;17:85–92.

22. Stone RC, Du P, Feng D, et al. *RNA-Seq for enrichment and analysis of IRF5 transcript expression in SLE.* PLoS One 2013;8:e54487.

23. Tandon M, Gallo A, Jang SI, et al. *Deep sequencing of short RNAs reveals novel microRNAs in minor salivary glands of patients with Sjogren's syndrome*. Oral Dis 2012;18:127–31.

24. Song YJ, Li G, He JH, et al. *Bioinformatics-based identification of microRNA-regulated and rheumatoid arthritis-associated genes*. PloS One 2015;10: e0137551.

25. Vieira Braga FA, Teichmann SA, Chen X. *Genetics and immunity in the era of single-cell genomics.* Hum Mol Genet 2016;25(R2):R141–8.

26. Stubbington MJ, Lonnberg T, Proserpio V, et al. *T cell fate and clonality inference from single-cell transcriptomes*. Nat Methods 2016;13:329–32.

27. Ishigaki K, Shoda H, Kochi Y, et al. *Quantitative and qualitative characterization of expanded CD41 T cell clones in rheumatoid arthritis patients*. Sci Rep 2015; 5:12937.

28. Lappalainen T, Sammeth M, Friedlander MR, et al. *Transcriptome and genome sequencing uncovers functional variation in humans*. Nature 2013;501:506–11.

29. Luke J, O'Connor AG, Liu X, et al. *Estimating the proportion of disease heritability mediated by gene expression levels.* New York: Cold Spring Harbor Laboratory; 2017. p. 118018.

30. Farh KK, Marson A, Zhu J, et al. *Genetic and epigenetic fine mapping of causal autoimmune disease variants.* Nature 2015;518:337–43.

31. Harley JB, Alarcon-Riquelme ME, Criswell LA, et al. *Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in ITGAM, PXK, KIAA1542 and other loci.* Nat Genet 2008;40:204–10.

32. Lewis MJ, Vyse S, Shields AM, et al. *UBE2L3 polymorphism amplifies NF-kappaB activation and promotes plasma cell development, linking linear ubiquitination to multiple autoimmune diseases*. Am J Hum Genet 2015;96:221–34.

33. Zhu Z, Zhang F, Hu H, et al. *Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets.* Nat Genet 2016;48:481–7.

34. Gusev A, Ko A, Shi H, et al. *Integrative approaches for large-scale transcriptome-wide association studies.* Nat Genet 2016;48:245–52.

35. Brookes E, Shi Y. *Diverse epigenetic mechanisms of human disease.* Annu Rev Genet 2014;48:237–68.

36. Greenleaf WJ. *Assaying the epigenome in limited numbers of cells.* Methods 2015;72:51–6.

37. Gibbs JR, van der Brug MP, Hernandez DG, et al. *Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain*. PLoS Genet 2010;6:e1000952.

38. Javierre BM, Fernandez AF, Richter J, et al. *Changes in the pattern of DNA methylation associate with twin discordance in systemic lupus erythematosus.* Genome Res 2010;20:170–9.

39. Jeffries MA, Dozmorov M, Tang Y, et al. *Genome-wide DNA methylation patterns in CD41 T cells from patients with systemic lupus erythematosus.* Epigenetics 2011;6:593–601.

40. Lemire M, Zaidi SH, Ban M, et al. *Long-range epigenetic regulation is conferred by genetic variation located at thousands of independent loci*. Nat Commun 2015;6:6326.

41. Scharer CD, Blalock EL, Barwick BG, et al. *ATAC-seq on biobanked specimens defines a unique chromatin accessibility structure in naive SLE B cells.* Scientific Rep 2016;6:27030.

42. Zhou Y, Simpson S Jr, Holloway AF, et al. *The potential role of epigenetic modifications in the heritability of multiple sclerosis.* Mult Scler 2014;20:135–40.

43. Trerotola M, Relli V, Simeone P, et al. *Epigenetic inheritance and the missing heritability.* Hum genomics 2015;9:17.

44. Yehuda R, Daskalakis NP, Bierer LM, et al. *Holocaust exposure induced intergenerational effects on FKBP5 methylation.* Biol Psychiatry 2016;80:372–80.

45. Connolly S, Heron EA. *Review of statistical methodologies for the detection of parent-of-origin effects in family trio genome-wide association data with binary disease traits. Brief Bioinformatics* 2015;16:429–48.

46. Bhattacharjee M, Balakrishnan L, Renuse S, et al. *Synovial fluid proteome in rheumatoid arthritis*. Clin Proteomics 2016;13:12.

47. Park YJ, Chung MK, Hwang D, et al. *Proteomics in rheumatoid arthritis research.* Immune Netw 2015;15:177–85.

48. Hillenmeyer ME, Fung E, Wildenhain J, et al. The chemical genomic portrait of yeast: uncovering a phenotype for all genes. Science 2008;320:362–5.

49. Breker M, Schuldiner M. *The emergence of proteome-wide technologies: systematic analysis of proteins comes of age*. Nat Rev Mol Cel Biol 2014;15:453–64.

50. Rossin EJ, Lage K, Raychaudhuri S, et al. *Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology*. PLoS Genet 2011;7:e1001273.

51. Marson A, Housley WJ, Hafler DA. *Genetic basis of autoimmunity*. J Clin Invest 2015;125:2234–41.

52. Chen JY, Pinkerton SL, Shen C, et al. *An integrated computational proteomics method to extract protein targets for Fanconi anemia studies*. 21st annual ACM symposium on applied computing. Dijon, France, April 23–27, 2006. 173–9.

53. Okada Y, Wu D, Trynka G, et al. *Genetics of rheumatoid arthritis contributes to biology and drug discovery.* Nature 2014;506:376–81.

54. Hale PJ, Lopez-Yunez AM, Chen JY. *Genome-wide meta-analysis of genetic susceptible genes for type 2 diabetes.* BMC Syst Biol 2012;6(Suppl 3):S16.

55. Wu X, Chen JY. *Molecular Interaction Networks: Topological and Functional Characterizations. In: Alterovitz G, Benson R, Ramoni M, editors. Automation in Proteomics and Genomics: An Engineering Case-Based Approach*. Chichester (UK): John Wiley & Sons Ltd.

56. Guma M, Tiziani S, Firestein GS. *Metabolomics in rheumatic diseases: desperately seeking biomarkers*. Nat Rev Rheumatol 2016;12:269–81.

57. Joseph B, Corwin JA, Li B, et al. *Cytoplasmic genetic variation and extensive cytonuclear interactions influence natural variation in the metabolome*. ELife 2013;2:e00776.

58. Reed LK, Lee K, Zhang Z, et al. *Systems genomics of metabolic phenotypes in wild-type Drosophila melanogaster.* Genetics 2014;197:781–93.

59. Nicholson G, Rantalainen M, Li JV, et al. *A genome-wide metabolic QTL analysis in Europeans implicates two loci shaped by recent positive selection.* PLoS Genet 2011;7:e1002270.

60. Shin SY, Fauman EB, Petersen AK, et al. *An atlas of genetic influences on human blood metabolites.* Nat Genet 2014;46:543–50.

61. *An integrated encyclopedia of DNA elements in the human genome.* Nature 2012;489:57–74.

62. Bernstein BE, Stamatoyannopoulos JA, Costello JF, et al. *The NIH Roadmap Epigenomics Mapping Consortium.* Nat Biotechnol 2010;28:1045–8.

63. Kichaev G, Pasaniuc B. *Leveraging functional-annotation data in trans-ethnic fine-mapping studies.* Am J Hum Genet 2015;97:260–71.

64. Lu X, Zoller EE, Weirauch MT, et al. *Lupus risk variant increases pSTAT1 binding and decreases ETS1 expression*. Am J Hum Genet 2015;96:731–9.

65. Holzinger ER, Ritchie MD. *Integrating heterogeneous high-throughput data for meta-dimensional pharmacogenomics and disease-related studies*. Pharmacogenomics 2012;13:213–22.

66. Culverhouse R, Suarez BK, Lin J, et al. *A perspective on epistasis: limits of models displaying no main effect.* Am J Hum Genet 2002;70:461–71.

67. Ritchie MD, Holzinger ER, Li R, et al. *Methods of integrating data to uncover genotype-phenotype interactions.* Nat Rev Genet 2015;16:85–97.

68. Corradin O, Saiakhova A, Akhtar-Zaidi B, et al. *Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits.* Genome Res 2014;24:1–13.

69. Schwarz DF, Konig IR, Ziegler A. *On safari to random jungle: a fast implementa- tion of random forests for high-dimensional data.* Bioinformatics 2010;26:1752–8.

70. Jiang X, Barmada MM, Visweswaran S. *Identifying genetic interactions in genome-wide data using Bayesian networks.* Genet Epidemiol 2010;34:575–81.

71. Turner SD, Dudek SM, Ritchie MD. *ATHENA: a knowledge-based hybrid backpropagation-grammatical evolution neural network algorithm for discovering epistasis among quantitative trait loci.* BioData Min 2010;3:5.

72. Hammaker D, Whitaker JW, Maeshima K, et al. *LBH gene transcription regulation by the interplay of an enhancer risk allele and DNA methylation in rheumatoid arthritis.* Arthritis Rheumatol 2016;68:2637–45.

73. KumasakaN, KnightsAJ, GaffneyDJ. *Fine-mapping cellular QTLs with RASQUAL and ATAC-seq.* Nat Genet 2016;48:206–13.

74. Folkersen L, Brynedal B, Diaz-Gallo LM, et al. *Integration of known DNA, RNA and protein biomarkers provides prediction of anti-TNF response in rheumatoid arthritis: results from the COMBINE study*. Mol Med 2016;22:322–8.

75. Hou L, Zhao H. *A review of post-GWAS prioritization approaches.* Front Genet 2013;4:280.

76. Kichaev G, Yang WY, Lindstrom S, et al. *Integrating functional data to prioritize causal variants in statistical fine-mapping studies.* PLoS Genet 2014;10:e1004722.

77. Zaitlen N, Pasaniuc B, Gur T, et al. *Leveraging genetic variability across populations for the identification of causal variants.* Am J Hum Genet 2010;86:23–33.

78. Chen W, Larrabee BR, Ovsyannikova IG, et al. *Fine mapping causal variants with an approximate Bayesian method using marginal test statistics*. Genetics 2015; 200:719–36.

79. James RA, Campbell IM, Chen ES, et al. *A visual and curatorial approach to clinical variant prioritization and disease gene discovery in genome-wide diagnostics.* Genome Med 2016;8:13.

80. Salatino S, Ramraj V. *BrowseVCF: a web-based application and workflow to quickly prioritize disease-causative variants in VCF files.* Brief Bioinformatics 2016.

81. Glanzmann B, Herbst H, Kinnear CJ, et al. *A new tool for prioritization of sequence variants from whole exome sequencing data.* Source code Biol Med 2016;11:10.

82. Pickrell JK. *Joint analysis of functional genomic data and genome-wide association studies of 18 human traits.* Am J Hum Genet 2014;94:559–73.

83. Wu X, Hasan MA, Chen JY. *Pathway and network analysis in proteomics.* J Theor Biol 2014;362:44–52.

84. Smith GD, Ebrahim S. *Mendelian randomization: prospects, potentials, and limitations*. Int J Epidemiol 2004;33:30–42.

85. Pickrell JK, Berisa T, Liu JZ, et al. *Detection and interpretation of shared genetic influences on 42 human traits.* Nat Genet 2016;48:709–17.

86. Gratten J, Visscher PM. *Genetic pleiotropy in complex traits and diseases: implications for genomic medicine*. Genome Med 2016;8:78.

87. Visscher PM, Yang J. *A plethora of pleiotropy across complex traits*. Nat Genet 2016;48:707–8.

88. Denny JC, Ritchie MD, Basford MA, et al. *PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations.* Bioinformatics 2010;26:1205–10.

89. Hilton IB, D'Ippolito AM, Vockley CM, et al. *Epigenome editing by aCRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers.* Nat Biotechnol 2015;33:510–7.

90. Gibson GJ, Yang M. *What rheumatologists need to know about CRISPR/Cas9.* Nat Rev Rheumatol 2017;13(4):205–16.

91. Patwardhan RP, Hiatt JB, Witten DM, et al. *Massively parallel functional dissection of mammalian enhancers in vivo.* Nat Biotechnol 2012;30:265–70.

92. Murtha M, Tokcaer-Keskin Z, Tang Z, et al. *FIREWACh: high-throughput functional detection of transcriptional regulatory modules in mammalian cells.* Nat Methods 2014;11:559–65.

93. Arnold CD, Gerlach D, Spies D, et al. *Quantitative genome-wide enhancer activity maps for five Drosophila species show functional enhancer conservation and turnover during cis-regulatory evolution.* Nat Genet 2014;46:685–92.

94. Vockley CM, Guo C, Majoros WH, et al. *Massively parallel quantification of the regulatory effects of noncoding genetic variation in a human cohort.* Genome Res 2015;25:1206–14.

95. Vanhille L, Griffon A, Maqbool MA, et al. *High-throughput and quantitative assessment of enhancer activity in mammals by CapStarr-seq.* Nat Commun 2015;6:6905.

96. Crotty S, Pipkin ME. *In vivo RNAi screens: concepts and applications.* Trends Immunol 2015;36:315–22.

DENSE GENOTYPING IDENTIFIES LOCI FOR RHEUMATOID ARTHRITIS RISK AND DAMAGE IN AFRICAN AMERICANS

by

DANILA MI*, LAUFER VA*, REYNOLDS RJ, YAN Q, LIU N, GREGERSEN PK, KERN M, LANGEFELD CD, ARNETT DK, BRIDGES SL JR.

Molecular Medicine

Format adapted [and errata corrected] for dissertation

Abstract

More than 100 risk loci for rheumatoid arthritis (RA) have been identified in individuals of European and Asian descent, but the genetic basis for RA in African Americans is less well understood. We genotyped 610 African Americans with autoantibody-positive RA and 933 African American controls on the Immunochip (iChip) array. Using multivariable regression, we evaluated the association between iChip markers and the risk of RA and radiographic severity. The single nucleotide polymorphism (SNP) rs1964995 (odds ratio = 1.97, p = 1.28 × 10–15) near HLA-DRB1 was the most strongly associated risk SNP for RA susceptibility; SNPs in *AFF3*, *TNFSF11* and *TNFSF18* loci were suggestively associated (10^{-4}< p < 3.1 × 10^{-6}). Trans-ethnic fine mapping of *AFF3* identified a 90% credible set containing previously studied variants, including rs9653442, rs7608424 and rs6712515, as well as the novel candidate variant rs11681966; several of these likely influence AFF3 gene expression level. Variants in *TNFRSF9*, *CTLA4*, *IL2RA*, *C5/TRAF1* and *ETS1* – but no variants within the major histocompatibility complex – were associated with RA radiographic severity. Conditional regression and pairwise linkage disequilibrium (LD) analyses suggest that additional pathogenic variants may be found in *ETS1* and *IL2RA* beyond those found in other ethnicities. In summary, we used the dense genotyping of the iChip array and the unique LD structure of African Americans to validate known risk loci for RA susceptibility and radiographic severity, and to better characterize the associations of *AFF3*, *ETS1* and *IL2RA*.

Introduction

Rheumatoid arthritis (RA) is a systemic autoimmune disorder characterized by synovial joint inflammation, with disease phenotype ranging from mild joint involvement to severe joint destruction and permanent disability (1). The factors responsible for RA heterogeneity are poorly understood, but both genetic and environmental factors contribute to its pathogenesis and clinical expression. Most RA patients have serum autoantibodies, such as rheumatoid factor (RF) or anticyclic citrullinated peptide antibody (ACPA), which can be present before the onset of clinically relevant disease (2) and are associated with radiographic severity (3).

To replicate and fine map risk loci identified in genome-wide association studies (GWAS) of autoimmune and inflammatory disorders such as RA, the Immunochip Consortium designed the Immunochip (iChip), a custom Illumina Infinium high-density array that has been used to study RA in patients of several racial and ethnic backgrounds (4–8). Using the iChip and many other arrays, 100 RA risk loci of genome-wide significance ($p < 5 \times 10^{-8}$) have been identified in individuals of European and Asian ancestry, including *HLA-DRB1*, *PADI4*, *PTPN22* and *CTLA4* (9,10). However, there is a paucity of genetic association data on RA in African Americans. Many of the genetic influences on RA are similar among those of European ancestry and African Americans (11). However, there are important differences; for instance, polymorphisms in *CCR6*, *TAGAP* and *TNFAIP3* have discordant odds ratios (ORs) compared with those reported in European RA patients (11). Furthermore, the *PTPN22* risk allele containing rs2476601, which has the highest effect size on RA susceptibility of any locus outside the major histocompatibility complex (MHC) in European populations, is essentially absent from the Yoruban population and is present in Asian and African American populations in low frequency (12).

HLA-DRB1 alleles encoding the shared epitope (SE) (13) have the strongest association with RA in Europeans and Asians. In addition to their role in susceptibility of RA, *HLA-DRB1* SE alleles are associated with erosive disease (14) and mortality in Europeans (15,16). Our group has shown that 43% of African Americans with RA have at least one *HLA-DRB1* SE allele compared with ~60–70% of Europeans with RA (17). At the level of *HLA-DRB1* amino acid residues and their association with RA susceptibility, there are both similarities and major differences between Europeans and African Americans. The valine residue at position 11, as found in Europeans, is most strongly associated with RA in African Americans (18). However, an aspartic acid residue at position 11, indicative of the classical allele *09:01, confers a two-fold increased risk of RA in African Americans and is also associated with RA in Koreans (19), but not in individuals of European ancestry. After conditioning on residue substitutions at position 11, amino acid positions 71 and 74 are not significantly associated with RA in African Americans, as they are in Europeans (18). Subphenotypes (ie, disease subtraits) (20) are frequently more heritable than the complex disease traits of which they are a part (21). In addition, genetic association studies on specific subphenotypes tend to focus on less heterogeneous patients than studies on overall disease susceptibility. Radiographic severity is a characteristic RA subphenotype, with an estimated heritability between 45% and 58% (22). Genetic influences on radiographic severity have been examined in several ethnic groups (23–31), and ~30 risk loci have been identified in European and Asian populations, including *CXCR5*, *AFF3*, *C5-TRAF1*, *IL2RA*, *IL6*, *IL10* and *FCRL3*. Typically, these studies have included smaller numbers of participants than studies on susceptibility of RA, and the definition of radiographic severity has not been uniform, which may limit statistical power and complicate replication. Finally, few studies have addressed radiographic severity of RA in African Americans.

The objective of our study is to investigate the associations of known autoimmune disease risk loci with RA and its radiographic severity in African Americans. In view of the heterogeneity of disease associations among ethnic groups, shorter haplotype blocks and differences in allele frequency in African Americans, we hypothesize that fine mapping will identify differences in the genetic architecture of RA in African Americans compared with other ethnicities. Strengths of this study include analysis of the largest group of African Americans with RA currently available in the world, with accompanying high-quality radiographic outcomes data (34). A major goal of research into the genetics of complex diseases is to identify pathogenic variants that produce disease associations. In view of this, we draw on available data from association testing of >100,000 Asians and Europeans and use cutting-edge algorithms to prioritize genetic risk variants in African Americans with RA in the *AFF3* locus. This study represents an important addition to the literature on the genetics of RA, which has primarily involved participants of Asian and European descent.

Materials and Methods

Study Population

The CLEAR (Consortium for the Longitudinal Evaluation of African Americans with Early Rheumatoid Arthritis) study enrolled African Americans with RA of <2 years' disease duration (CLEAR I), African Americans with RA irrespective of disease duration (CLEAR II) and African American healthy controls, as previously described (17). Participants were enrolled at five academic sites: University of Alabama at Birmingham (coordinating center); Grady Hospital/Emory University, Atlanta, Georgia; University of North Carolina, Chapel Hill; Washington University, St. Louis, Missouri; and Medical University of South Carolina, Charleston. CLEAR controls were African Americans without rheumatic diseases who were matched (as a group) by age, sex and geographic location to CLEAR RA patients. The Institutional Review Boards of the participating institutions approved human subject research protocols. Biologic specimens and patient information, including sociodemographic characteristics, medical history, medications and disease activity measures, were collected (17). The majority of CLEAR participants were ACPA-positive, as previously reported (32). Of the 837 African American healthy controls included for analysis in the current study, 404 were from CLEAR and 433 were from the Birmingham, Alabama, area (33).

Radiographs of hands/wrists and feet were obtained at the CLEAR enrollment visit for participants with RA and assigned a modified total Sharp score (mTSS) (range 0–448) using the modified Sharp/van der Heijde method (34). Scoring was performed using state-of-the-art methods under the auspices of Désirée van der Heijde, a world expert in quantitative assessment of radiographs in rheumatic diseases (34). Furthermore, mTSS scores for participants from CLEAR I and CLEAR II have been validated extensively (35,36).

Sample Genotyping

Genotyping was carried out using the iChip array at the Feinstein Institute for Medical Research in Manhasset, New York. Genotype clustering was performed using the GenTrain2 clustering algorithm. Genotype calling was performed with the genotyping module of the GenomeStudio data analysis software package.

Quality Control

Rigorous quality control procedures were employed, including checks for gender inconsistency, relatedness (duplicates and first- or second-degree relatives) and ethnic outliers. The sample call rate threshold was 95%. The marker call rate was >98.5% across all SNPs, after removing low-quality SNPs and rare SNPs, those with minor allele frequency (MAF) <5% and SNPs out of Hardy-Weinberg equilibrium (using control samples only, using p value >1 × 10–5).

Association Testing of iChip Markers with RA Susceptibility

Of 610 RA cases, 593 (97%) were autoantibody-positive (defined as positive for either RF or anti-CCP antibody tests) and were included in the analysis of RA susceptibility. Multivariable logistic regression was used to evaluate the association between iChip markers and autoantibody-positive RA. Sex and European admixture proportion (calculated using Eigenstrat v6.0) (17) were included as covariates. Two-sided p values are reported, except as noted for trans-ethnic fine mapping of the AFF3 locus (see Trans-ethnic fine mapping of the *AFF3* locus section). To adjust for variability due to *HLA-DRB1* in the extended MHC locus (Chr6:26,000,000–34,000,000), we fit a model accounting for the variability of all four-digit *HLA-DRB1* SE alleles. LocusZoom plots were used to display the fine mapping results graphically (37).

Association Testing of iChip Markers with RA Severity

The modified total radiographic scores (mTSS) were over-dispersed in the CLEAR cohort, with a high proportion of individuals having no erosions or joint space narrowing (mTSS = 0): 156 of 230 CLEAR I participants (67.8%) and 150 of 365 CLEAR II participants (41.1%) (see Supplementary Figure S1). We assessed several count regression models and found that the zero-inflated negative binomial model had the best fit for the data, likely due to the high proportion of participants without damage (mTSS = 0). Thus, we used this method to evaluate the association of genetic markers with radiographic severity (under an additive genetic model). Association testing was carried out using the PSCL package in R (38) after adjusting for body mass index, sex, smoking status, percent European admixture (see [17] for details) and disease duration (in months) as covariates. Due to the inclusion criteria, disease duration was much shorter in CLEAR I (early RA) (median 1.01 years; interquartile range 0.57–1.52 years) than in CLEAR II (any disease duration) (median 9.25 years; interquartile range 3.42–17.75 years). Using a square root transformation for disease duration improved the model fit and reduced genomic inflation ($\lambda_{GC} = 1.10$) compared with a model using untransformed disease duration. After removal of the SNPs in the extended MHC and other associated loci, the λ_{GC} value was further reduced to 1.04 (See quantile-quantile plot in Supplementary Figure S2).

Trans-Ethnic Fine-Mapping of the AFF3 Locus

We conducted trans-ethnic fine mapping of the *AFF3* locus, combining our RA susceptibility data with those from a previously published large trans-ethnic meta-analysis (10). To accomplish this, we: (1) aligned reference and alternate alleles from all Asian, European and CLEAR populations to match those from the 1000 Genomes project (39); (2) generated LD matrices either from our genotyping data (African Americans) or from the 1000 Genomes project (Asian and European populations; data from Okada *et al.*) (10); (3) annotated SNPs from all three ethnicities using 8,138 genomic annotations (for example, DNAse hypersensitivity, enhancer markings and so on) provided with the PAINTOR3 algorithm; and (4) trimmed these to the top five uncorrelated annotations (correlation coefficient <0.10), excluding the annotations with lower Bayes factors.

We then confirmed the algorithm was working properly by examining the results it produced in RA loci in which the causal variant was known. For instance, we generated a posterior probability of 1.0 for rs2476601 in *PTPN22* in Europeans with RA. Following this, we calculated the posterior probability that each variant in the *AFF3* locus was pathogenic using PAINTOR3 (39), which assigns a probability ranging from 0 (very unlikely) to 1 (highly likely). We ran the algorithm using genetic data from all three populations, and defined a “90% credible set” for candidate pathogenic variants as previously reported (40) (see Table 3). Although PAINTOR3 is capable of modeling more than one causal variant per locus, in this study we conducted trans-ethnic fine mapping under the assumption of one causal variant.

Calculation of Number of Effective Markers for the iChip Array

Because the iChip contains many variants concentrated in specific loci and in LD with one another, the number of independent tests is much smaller than the actual number of variants genotyped. Estimates of the number of effective markers for custom genotyping arrays such as the iChip vary widely, between 2,800 and 60,000 LD-independent markers (41–43). To find independent SNPs, we used Plink (44), as previously utilized for iChip data (45), and found 16,154 LD-independent SNPs. We thus defined an iChip-wide statistical significance threshold as 0.05 divided by 16,154 LD-independent SNPs, or p = 3.1 × 10–6, similar to previous reports. We report any variants having p < 1 × 10–4 as showing suggestive statistical associations (for both susceptibility and severity).

Results

Following quality control procedures, 100,268 SNPs with MAF >0.05 were available for analysis in 610 RA cases and 837 healthy controls (as stated in Materials and Methods, 593 [97%] were autoantibody-positive and were included in subsequent analyses). The demographic characteristics of African Americans from the CLEAR registry included in this study are presented in Table 1. Characteristics of Birmingham controls did not differ significantly from the CLEAR registry with respect to sex, European admixture proportion or other variables (17).

Table 1 - Demographic, clinical, genetic and radiographic characteristics of African American participants with RA and healthy controls from the CLEAR registry.

Demographic Characteristic	Cases	Controls	Cases	Controls
Baseline characteristics	CLEAR I N = 233	CLEAR II N = 360	CLEAR I N = 139	CLEAR II N = 265
Age in years, mean (SE)	50.0 (13.0)	56.0 (11.8)	48.1 (12.4)	57.3 (8.7)
Sex (female), %	82.7	85	75.5	72
Disease duration in months, mean (SE)	12.9 (7.1)	114.0 (119.2)	–	–
Body mass index, mean (SE)	31.4 (7.8)	–	31.7 (7.6)	–
Global European admixture estimate, mean (SE)	0.17 (0.09)	0.16 (0.10)	0.16 (0.09)	0.17 (0.10)
Number of tender joints, median (IQR 25–75)	4.0 (1.0–12.0)	4.0 (1.0–9.5)	–	–
Number of swollen joints, median (IQR 25–75)	3.0 (1.0–7.0)	4.0 (1.0–10.0)	–	–
Medications				
Biologics ever used (%)	4.4	–	19.5	–
Other DMARDs (%)	81.4	–	87.1	–
Methotrexate, current use (%)	63.9	60.3	–	–
Radiographic score, mean (SE)				
Joint-erosion score	1.6 (4.3)	10.7 (18.5)		
Joint-narrowing score	2.1 (5.7)	18.1 (27.7)		
Total score	3.7 (9.4)	28.8 (44.1)		

SE: standard error; IQR: interquartile range; DMARD: disease-modifying anti-rheumatic drug.

We evaluated the association between iChip markers and RA using logistic regression and adjusting for the proportion of overall European admixture. We observed seven non-HLA loci suggestively associated with RA (defined as $p < 10^{-4}$); the lead SNP (ie, the most strongly associated) at each locus is shown in Table 2. As expected, the markers with the strongest association with RA were found in the MHC region (Figure 1A). We identified rs1964995 in *HLA-DRB1* as the variant with strongest association with RA (OR = 1.97, $p = 1.28 \times 10^{-15}$).

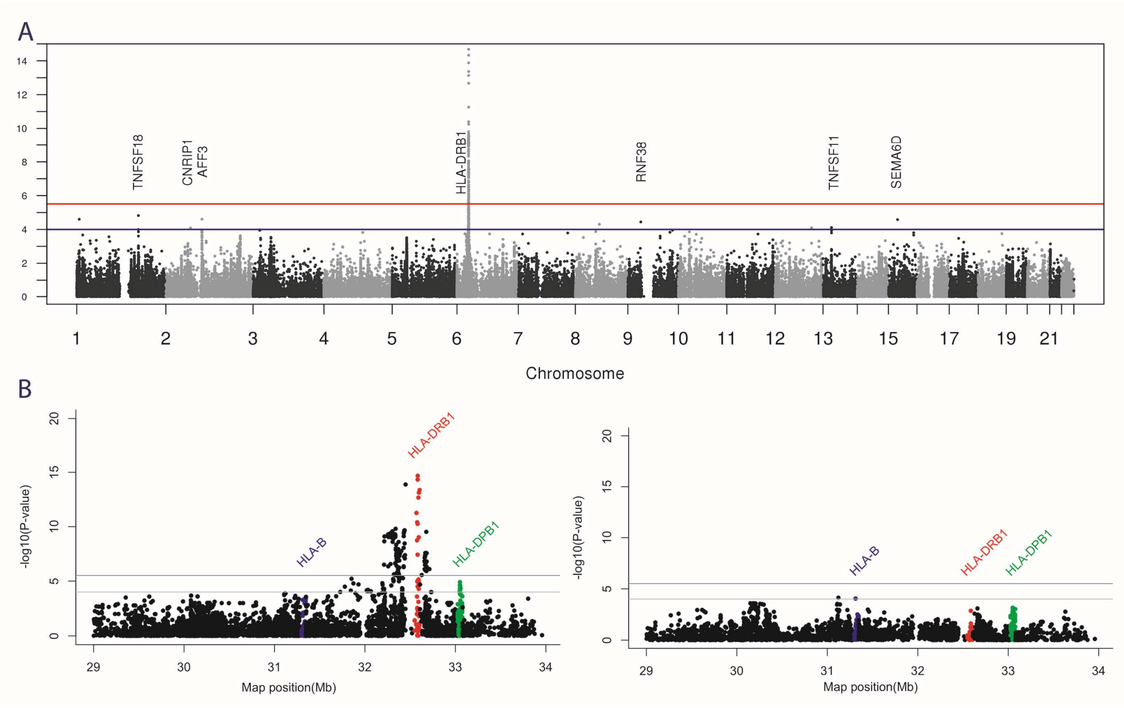


Figure 1. (A) Manhattan plot of the association of iChip variants with autoantibody-positive RA in African Americans. The x-axis indicates chromosome and position, the y-axis indicates association strength –log(p). The blue line illustrates suggestive statistical association (p = 1x10^{-4}). The red line illustrates iChip-wide level of statistical significance (p = 3.1x10^{-6}). (B) Conditional analysis in the extended MHC region (chr6:26,000,000- 34,000,000). Axes have the same meanings as in (A). Red, green and blue dots are SNPs in the regions immediately surrounding *HLA-DRB1*, *HLA-DPB1* and *HLA-B*, respectively. The left panel shows the association summary statistics before conditioning on *HLA-DRB1* alleles. The strongest association maps to *HLA-DRB1*. The right panel shows the locus after conditioning on *HLA-DRB1* 4-digit alleles.

We performed a conditional analysis of the variation contained within the extended MHC region as previously described by Raychaudhuri *et al.* (46). As shown in Figure 1B, conditioning on the *HLA-DRB1* alleles substantially attenuated the strength of association of other variants within the extended MHC region. rs3134792 near *HLA-B* displayed an OR of 2.01 (95% confidence interval [CI] = 1.42–2.88; p = 9.92 × 10–5) for the association with RA susceptibility after conditioning. This effect size and direction of effect are consistent with those reported for amino acid position 9 of *HLA-B* in Europeans (OR = 2.12, CI = 1.89–2.38). No variants in *HLA-DPB1* were associated with RA after controlling for the *HLA-DRB1* alleles. However, the direction of effect and ORs measured in this locus were similar to those found in studies of European populations. Specifically, although above the threshold for statistical significance, rs9277357 had an OR of 1.34 (95% CI = 1.14–1.59; p = 5.39 × 10–4), which is consistent with that previously reported for amino acid position 9 in *HLA-DPB1* (OR = 1.40, CI = 1.31–1.50).

Table 2 - Variants outside the HLA region associated with autoantibody-positive RA

rsID	Chr	Position	A1[a]	OR	95% CI	*P* value	Nearest genes
rs61828386	1	172863647	G	0.69	0.58–0.82	1.79×10^{-5}	*TNFSF18*, *FASLG*
rs67164098	2	68556131	A	1.67	1.31–2.13	4.65×10^{-5}	*CNRIP1*
rs11681966	2	100759457	C	1.5	1.23–1.78	4.04×10^{-5}	*AFF3*[b]
rs10758368	9	36310778	A	0.7	0.58–0.83	5.48×10^{-5}	*RNF38*
rs9533119	13	43049426	A	1.36	1.17–1.59	6.32×10^{-5}	*TNFSF11*
rs2934178	15	48218221	C	0.7	0.59–0.82	2.96×10^{-5}	*SEMA6D*

Variants having $p < 10^{-4}$ in African-Americans with RA. Chr: chromosome. [a]Indicates the test allele and minor allele for this the study. [b]Indicates a validated risk locus for RA.

We performed more detailed analysis on *AFF3* (see Materials and Methods and Figure 2), a validated RA risk locus among Europeans and Asians (10,47,48). We found that rs11681966 was suggestively associated with RA in African Americans (OR = 1.5, 95% CI 1.23–1.78, $p = 4.04 \times 10^{-5}$). The lead *AFF3* SNP (rs9653442) associated with RA in European ancestry ($p = 3.6 \times 10^{-12}$) (10) was not strongly associated with RA susceptibility (p = 0.015) in African Americans. Similarly, rs10209110, the index variant in AFF3in another study of RA in Europeans (4), was not associated in our dataset (p = 0.84). Therefore, due to differing association strengths and LD patterns in the locus, we conducted trans-ethnic fine mapping of this locus using data from African Americans, Asians and European RA patients and controls using PAINTOR3 (39) (see Methods).

Most association studies on *AFF3* have examined roughly the region from chr2:100,800,000 to 100,850,000, which contains index variants identified by multiple prior GWAS. However, our index variant (rs11681966, at chr2:100,759,457) is outside this region, located near the 5’ end of AFF3 (>1 kb). Thus, we defined the risk locus as a broader region (from chr2:100,709,000 to 100,875,000). We then conducted trans-ethnic fine mapping using PAINTOR 3. Figure 2 shows the 90% credible set for variants in the

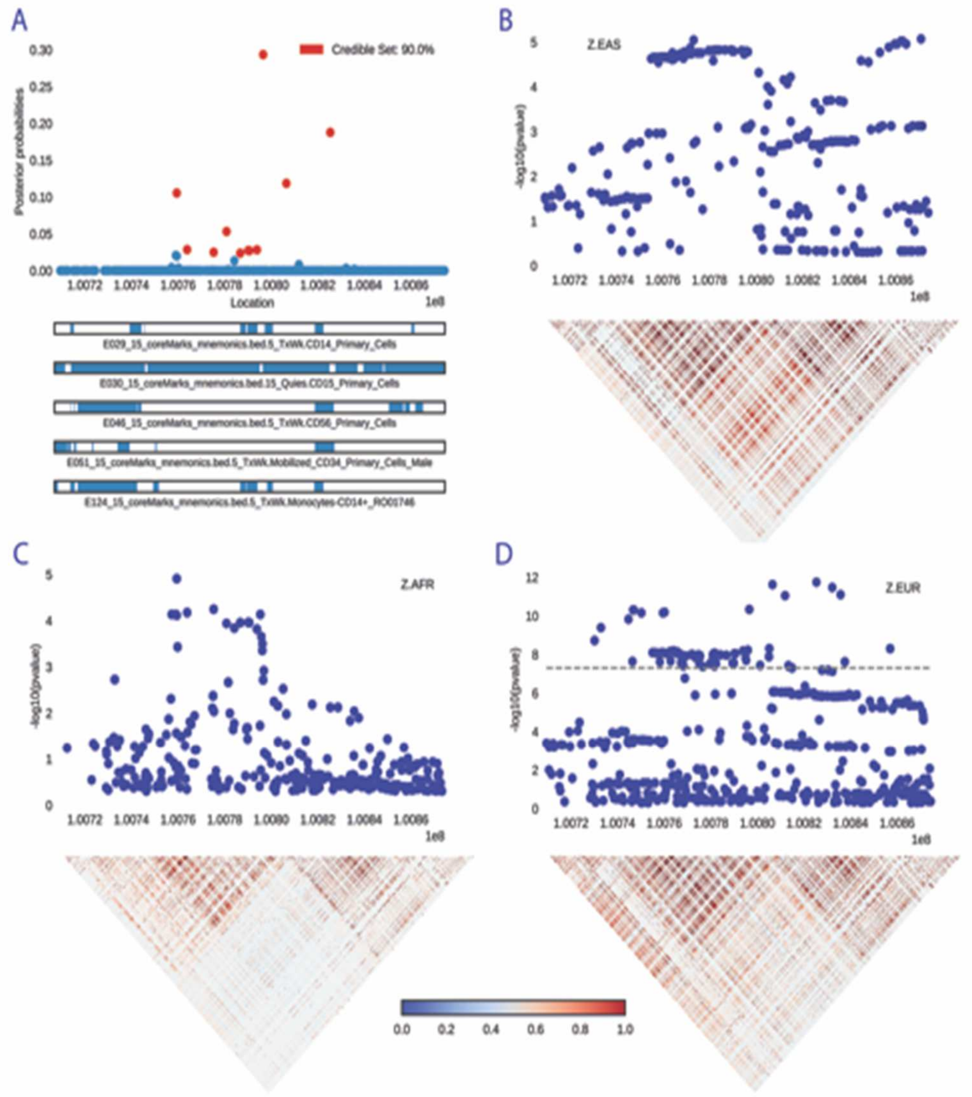


Figure 2 - Results from trans-ethnic fine mapping of the AFF3 locus. (A) The 90% credible set for candidate pathogenic variants (top) and selected annotations used to prioritize variants, with blue coloration indicating variants having a given annotation (bottom). (B–D) Zoom plot of association summary statistics versus genomic position for Asians, African Americans, and Europeans with RA. Each includes a heatmap for variants in the locus, colored according to a linkage disequilibrium LD heatmap generated from that population. The color bar (bottom middle) indicates the degree of linkage disequilibrium for variants in each LD heatmap.

AFF3 locus, enriched genomic annotations used to help construct the credible set and zoom plots in European, Asian and African American populations with LD heatmaps for each (Figures 2B–D, respectively). Doing so revealed that rs11681966 and a linked variant, rs13003982, were also in the 90% credible set defined by PAINTOR3 (see Figure 2A and Table 3). Consistent with previous reports of autoimmune disease in other ethnicities, our trans-ethnic fine mapping analysis of the AFF3 locus in combined African American, European and Asian RA identified rs9653442, rs6712515 and rs7608424 as likely candidates to be pathogenic variants (see Figure 2 and Table 3). Several of these SNPs are listed as index variants in the National Human Genome Research Institute (NHGRI) GWAS catalog (49) and have been noted in prior studies (50).

Table 3 -90 % Credible Set of candidate causal variants in the *AFF3* locus.

rsID	Chr	Position	Effect allele	Alt. allele	Z AA[a]	Z EAS[b]	Z EUR[c]	Posterior probability
rs13003982	chr2	100759078	T	C	–3.80	–4.09	–5.65	0.021
rs11681966[d]	chr2	100759457	A	C	–4.22	–4.09	–5.65	0.105
rs12712067	chr2	100763900	T	G	–3.82	–4.08	–5.69	0.029
rs4851257	chr2	100775297	T	C	–3.86	–4.14	–5.59	0.025
rs4851258	chr2	100780830	T	C	–3.68	–4.17	–5.60	0.053
rs4851261	chr2	100786717	A	G	–3.69	–4.17	–5.60	0.024
rs10185059	chr2	100790172	T	C	–3.70	–4.16	–5.63	0.027
rs10185510	chr2	100790581	T	C	–3.69	–4.16	–5.63	0.027
rs12712071	chr2	100793876	A	G	–3.61	–4.17	–5.69	0.028
rs7608424[d]	chr2	100796543	T	G	–2.89	–4.15	–6.48	0.293
rs6712515[d]	chr2	100806514	T	C	–2.31	–3.66	–6.91	0.119
rs9653442[d]	chr2	100825367	T	C	–2.43	–3.50	–6.95	0.188

Chr: chromosome. a - Z-score from our study of African Americans. b - Z-score for East Asians from the trans-ethnic meta-analysis of Okada *et al.* c - Z-score for Europeans from Okada *et al.* d - Denotes that the variant was the index variant reported in this study, or in another genome-wide association study.

Table 4 - Association between iChip Markers and Radiography Severity of RA

rsID	Chr	Position	A1[a]	Stat	Effect size	95% CI	*P* value	Gene locus
rs228702	1	7945520	G	IRR	0.58	0.44–0.76	6.33×10^{-5}	*TNFRSF9*[b]
rs13014054	2	33678924	A	IRR	0.53	0.39–0.71	2.92×10^{-5}	*RASGRP3*
rs73055463	2	204712807	C	OR	1.99	1.38–2.86	4.72×10^{-5}	*CTLA4*[b]
rs7034499	9	123687231	C	OR	2.27	1.49–3.46	9.60×10^{-5}	*TRAF1-C5*[b]
rs7077067	10	6132692	A	IRR	1.48	1.22–1.78	5.16×10^{-5}	*IL2RA*[b]
rs7101785	11	696437	G	IRR	1.52	1.24–1.86	5.17×10^{-5}	*TMEM80*
rs7127742	11	118521637	G	IRR	0.4	0.25–0.62	5.66×10^{-5}	*PHLDB1/CXCR5*
rs4362159	11	128305571	A	IRR	0.48	0.34–0.69	6.26×10^{-5}	*ETS1*[b]
rs506746	13	101981771	A	IRR	0.53	0.41–0.68	4.19×10^{-7}	*NALC/ITGBL1*
rs7193451	16	11050356	G	IRR	1.65	1.30–2.09	7.09×10^{-5}	*CLEC16A*

Variants genotyped on the iChip associated with RA radiographic severity in African Americans. Zero and count refer to the model coefficients for the portions of the zero inflated negative binomial model. Chr: chromosome; IRR: incident rate ratio. a - The allele tested in this study. b - Indicates a validated risk locus for RA.

After quality control procedures, 100,169 SNPs with MAF >0.05 were available for analysis in 548 autoantibody-positive RA patients who had radiographic scores. A Manhattan plot illustrating the genetic variants associated with severity is shown in Figure 3. In contrast to studies in individuals of European ancestry, we did not find a statistically significant association between SNPs tagging the HLA region and radiographic severity (Table 4). We detected several suggestive associations, including variants in or near *AFF3* (Supplementary Figure S3A), TNFRSF9 (Supplementary Figure S3B), *CTLA4* (Supplementary Figure S3C), *IL2RA* (Supplementary Figure S3D),

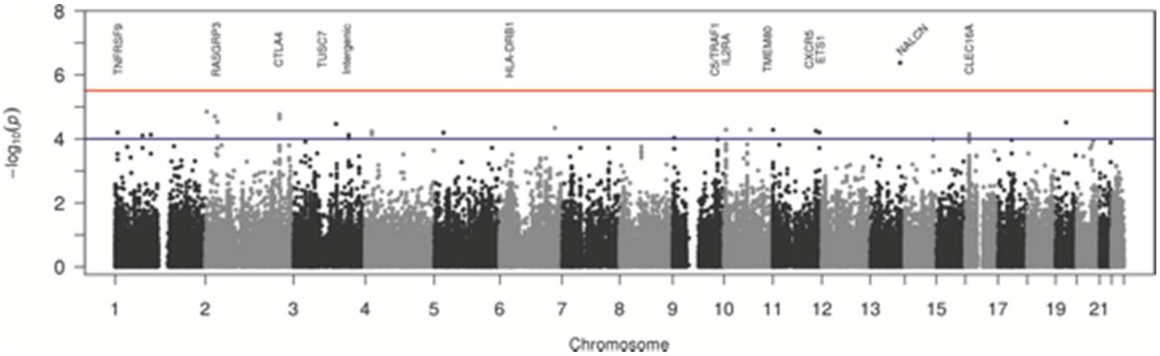


Figure 3 - Manhattan plot of the association of iChip variants with radiographic severity in African Americans with autoantibody-positive RA. The x-axis indicates chromosome and position, the y-axis indicates association strength –log(p). The blue line illustrates suggestive statistical association (p = 1*10–4). The red line illustrates the iChip-wide level of statistical significance (p = 3.1*10-6).

C5/TRAF1 (Supplementary Figure S3E) *and NALCN/ITGBL1*(Supplementary Figure S3F). rs506746 (near *NALCN/ITGBL1*) was the most strongly associated variant with radiographic severity (p = 4.33 × 10–7), but we could not evaluate support from LD for this association due to low marker density for this region on the iChip array, so no further analysis was performed.

We chose to examine two loci (*IL2RA* and *ETS1*) in more detail. Multiple *IL2RA* variants have been associated with autoimmune conditions (juvenile idiopathic arthritis, type 1 diabetes, systemic lupus erythematosus [SLE], multiple sclerosis, Graves' disease and so on) (49). Similar to previous reports, we observed a suggestive association in the *IL2RA* locus (rs7077067) with radiographic severity (p = 5.16 × 10–5) (51). rs7077067 was the lead SNP in this study, which differs from that in Europeans, rs2104286 (51). In our study, rs2104286 had MAF = 0.05 and was only weakly associated with RA radiographic severity (p = 0.024). There is a paucity of trans-ethnic association summary statistics for RA radiographic severity. Thus, we relied on conditional analysis, pairwise

LD estimates and prior studies (not trans-ethnic fine mapping) to further understand these loci. Adjusting for the effect of rs2104286 did not eliminate the association of rs7077067 with severity ($p = 8.15 \times 10^{-5}$). Previous studies in other ancestries have noted substantial LD between variants in *IL2RA* and the surrounding region, including *RBM17* (51,52). In this locus, we found shorter haplotype blocks and lower LD between genetic variants, so we sought to localize an association signal in this locus. We found that the most strongly associated SNPs in our dataset are in the first intron of *RBM17*, specifically in a ~5kb section of the genome displaying the H3K27Ac histone marks and DNAse hypersensitivity (Figure 4).

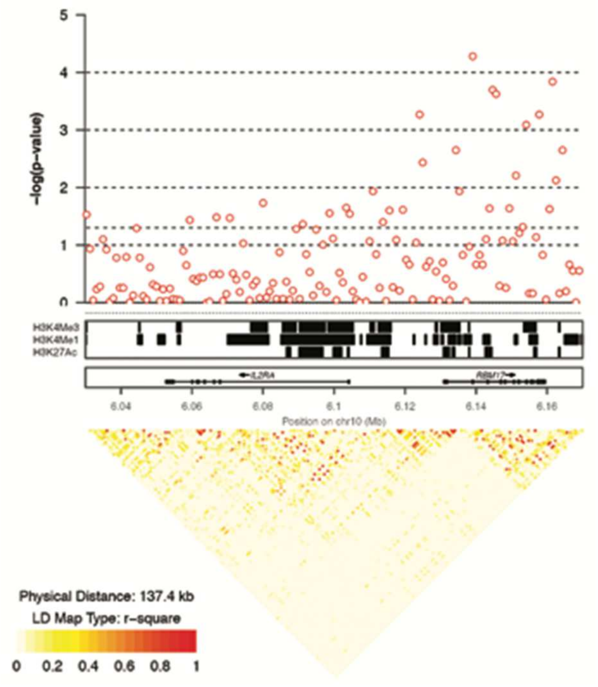


Figure 4 - (Top) Zoom plot indicating the strength of association in the region of chromosome 10 surrounding the *IL2RA* locus. Circles represent single nucleotide polymorphisms (SNPs); the y-axis measures negative log association p value and the x-axis represents genomic position. (Middle) Diagram indicating histone markings and gene diagrams corresponding to the genomic region in the zoom plot. (Bottom) LD heatmap for the region.

With regard to *ETS1*, we found that rs4362159 was associated with radiographic severity ($p = 6.26 \times 10^{-5}$). We also identified a variant linked to rs4362159, rs7108537, which was more weakly associated with RA radiographic severity ($p = 2.8 \times 10^{-4}$), but exists in the transcription factor binding site–rich region. Similarly, a previous study of SLE identified rs6590330 as an SLE risk variant that alters binding of pSTAT1 and affects *ETS1* expression in persons of Asian ancestry only. As expected, this SNP was not associated with RA radiographic severity ($p = 0.11$) in our dataset, nor was the lead SNP in our study in LD with rs6590330 ($r^2 = 0.03$), or with other previously described variants reported in the NHGRI GWAS catalog (49), for example, rs1128334 ($p = 0.27$; $r^2 = 0.01$) (53).

Discussion

Our analyses led to several important findings regarding RA in African Americans. First, SNPs tagging *HLA-DRB1* were significantly associated with RA susceptibility, but not radiographic severity. Second, *AFF3*, *TNFSF11* and *TNFSF18* (all previously validated loci for RA susceptibility) were associated suggestively with RA susceptibility ($1.0 \times 10^{-4} < p < 3.1 \times 10^{-6}$).

Third, *TNFRSF9*, *CTLA4*, *IL2RA*, *C5/TRAF1* and *CXCR5* were associated suggestively with radiographic severity. Finally, leveraging the differential LD pattern between Europeans and African Americans, we defined suggestive novel lead SNPs for the associations of *AFF3* with susceptibility and *IL2RA* with severity.

As expected from previous studies and our prior work (18), we found that the strongest association with RA susceptibility lies in the MHC region near *HLA-DRB1*. When the SNPs in the extended MHC were conditioned on the classical *HLA-DRB1* alleles, the association signal elsewhere in the MHC region is lost. This finding illustrates that the genome-wide significant SNPs in our study are tagging *HLA-DRB1* classical alleles. This was first noted in Europeans (46), but in that study residual significant association signal remained near *HLA-B* and *HLA-DPB1* after the conditioning analysis. We observed residual signals near *HLA-B* and *HLA-DPB1* having the same effect size and direction of effect, but they were not significantly associated with RA. Considering the consistency of effect size, this likely reflects statistical power, but could reflect biological differences as well.

Despite our study being well powered to detect an effect of similar magnitude, we found no association between MHC region SNPs and radiographic severity comparable to other reports. Viatte *et al.* reported an association between haplotypes defined by amino acid residues at positions 11, 71 and 74 of *HLA-DRB1* and radiographic damage (15). There are several possible explanations for this discrepancy. First, the differences could result from cohort inclusion criteria. Viatte *et al.* included autoantibody-positive RA, autoantibody-negative RA and inflammatory polyarthritis (not meeting ACR criteria for classification of RA), while our study focused exclusively on autoantibody-positive RA. Second, because they did not stratify based on autoantibody positivity, it is possible that their findings reflect the known association between radiographic severity and autoantibody positivity. Finally, biological differences between ethnicities cannot be ruled out.

AFF3 encodes LAF4, a transcriptional activator with suspected roles in lymphoid tissue development and oncogenesis (54). The locus has been associated with RA susceptibility in Europeans (4,10) as well as SLE and juvenile idiopathic arthritis (10,55,56). rs9653442 in particular has been the subject of several investigations as an autoimmune risk variant, and it was the index variant for RA risk in a trans-ethnic meta-analysis (10). In this study, it was found in the 90% credible set for pathogenic variants. However, the *AFF3* locus has been identified as containing multiple independent effects for common complex diseases (50). Consistent with this, our results further suggest several promising candidate pathogenic variants in addition to rs9653442. rs6712515 is an index variant reported in the NHGRI GWAS catalog (49), but has previously been associated with cognitive phenotypes rather than autoimmunity. These two variants as well as rs7608424 are known to be expression quantitative trait loci for *AFF3* expression (57). The index variant in our study, rs11681966, is found only ~400 bases from the transcription start site of *AFF3* in a conserved region capable of binding numerous transcription factors. Another variant in tight linkage with rs11681966, rs13003982, is located only ~40 bp from the transcription start site of *AFF3*. Thus, our data not only suggest candidate pathogenic variants, but an initial finding for functional studies of the contribution of RA genetic variants to *AFF3* to test.

We also detected several suggestive associations with RA radiographic severity. *CTLA4* is associated with RA in several populations (10,58), and the importance of *CD28/CTLA4* co-stimulation in RA is highlighted by the efficacy of CTLA4Ig (abatacept) (59). We detected a suggestive association of *TNFRSF9* with RA radiographic severity. *TNFRSF9* (*CD137*) is a member of the TNF receptor family known for its role in T cell co-stimulation. In RA, a soluble form of *CD137* is released by activated lymphocytes and is present in the serum (60). In collagen-induced arthritic mice, treatment with an anti-CD137 antibody protects against disease progression, possibly by amplifying antigen-specific CD11c + /CD8 + T lymphocytes and suppressing the pathogenic CD4 + T lymphocyte subset (61). While rs506746 (chr13:101981771, near NALCN and ITGBL1) showed the strongest association with radiographic severity, this finding should be interpreted cautiously because of the low coverage of this region on the iChip array.

We subjected two loci to additional analyses based on context provided by prior studies. We observed a suggestive association between RA severity and rs7077067, a variant near IL2RA. This locus has previously been linked to RA susceptibility (10), radiographic severity (31,51) and decreased likelihood of disease remission (62). Interleukin 2 receptor α (*IL2RA* or *CD25*) gene, together with *IL2RB* and *IL2RG*, encodes the high-affinity *IL2* receptor. In the absence of *IL2RA*, there is abnormal proliferation and migration of T cells, resulting in widespread inflammation. This may be due to reduced T cell apoptosis in the thymus, resulting in autoreactive T cell survival (63). In addition, rs2104286 in the *IL2RA* locus has been shown to reduce T cell activation in healthy individuals (64) and is associated with radiographic severity of RA in Europeans (49,52). Specifically, the minor allele of rs2104286 was associated with decreased progression of joint destruction and lower levels of soluble IL-2Rα. rs7077067 was not in LD (r2≤ 0.02) with any of the 10 index variants previously reported, including rs2104286. The relatively weak association of rs2104286 in our study and the low LD suggest that additional pathogenic variants may be found upstream of *IL2RA*. It is possible that different risk haplotypes predominate in African Americans with RA. Alternatively, low LD between variants in the *IL2RA* locus may preclude tagging the same pathogenic variant.

We also found an association of radiographic severity of RA with *ETS1*, a highly conserved transcription factor whose expression in B cells, T cells and natural killer cells strongly affects immune cell function. Ets1 knockout mice display aberrant T cell differentiation, altered cytokine expression and increased differentiation into memory and effector T cells (65). Downregulation of Ets1 increases formation of plasma cells in part by upregulating Pax-5 and inhibiting Blimp1 activity (66). In humans, lupus risk alleles are associated with lower *ETS1* mRNA expression, and the genetic basis of these findings differs in an ethnic-specific fashion (67). Specifically, increased binding of pSTAT1 to oligonucleotides containing the rs6590330 risk allele correlates with decreased *ETS1* expression in Asian SLE patients, but not in other populations, including African Americans (67). Consistent with this finding, we found that rs7108537, but not rs6590330, was associated with radiographic severity. There are transcription factor binding sites in the immediate vicinity of rs7108537, and the genotype of this SNP appears to affect *ETS1* expression in persons of Yoruban ancestry, but not in other populations (68). Therefore, our study of radiographic severity provides additional evidence that population- specific variants may contribute to risk of autoimmunity by decreasing *ETS1* expression.

Our association testing results should be interpreted cautiously, as the sample size may result in inflated effect size estimates. In addition, our study was not well powered to detect association of common SNPs (MAF 0.15-0.50) with effect sizes <1.3. Nevertheless, we used the largest registry of African Americans with RA for whom clinical, radiographic and genetic data are available. We were unable to attempt to replicate our findings, because no other cohorts of African Americans with RA are available.

Finally, it should be noted that the markers selected during the design of the custom iChip array were derived from the 1000 Genomes project from European individuals, which might be suboptimal for analysis of disease-associated variants in African Americans, and thus additional novel risk variants RA may yet exist in this ethnic group. Limitations of the fine mapping study include exclusion of some genotyped variants due to absence from the reference dataset used for LD and inability to confirm uniform alignment of some variants to reference genomes based on LD and Z-score information. This may lead to the exclusion of potentially interesting variants. For instance, rs11676922, an RA index SNP previously studied in a meta-analysis of RA in Han Chinese and Europeans (69), was not examined in this study. This SNP is in near-perfect LD with rs9653442 as well as rs6712515. As such, investigators who wish to carry out functional studies on variants in *AFF3* should note that including this variant might alter the posterior probabilities attributed to other variants.

Conclusion

In contrast to other reports, we find that SNPs in the MHC region do not appear to be associated with radiographic severity of RA in African Americans. Our study also demonstrates the utility of ethnic-specific analysis of genetic data. We confirm the association of *AFF3* with RA susceptibility and *IL2RA* and *ETS1* with radiographic severity, and our analysis of these loci suggests several candidate variants for functional validation. Our analysis of the *AFF3* locus suggests that rs11681966, rs9653442, rs7608424 and rs6712515 are high-priority targets for functional studies, but may exert effects in population-specific contexts. Our data add to evidence that *ETS1* autoimmune risk is mediated by ethnic-specific variants decreasing expression. In the *IL2RA* locus, our data suggest that trans-ethnic fine mapping studies could be valuable for RA susceptibility and radiographic severity due to different LD patterns. Overall, our study suggests that trans-ethnic genetic analysis is likely to be an important step in bringing precision medicine to complex autoimmune diseases, including RA.

References

1. McInnes IB, Schett G. (2011) *The pathogenesis of rheumatoid arthritis*. N. Engl. J. Med. 365: 2205–19.

2. Nielen MM, et al. (2004) *Specific autoantibodies precede the symptoms of rheumatoid arthritis: a study of serial measurements in blood donors*. Arthritis Rheum. 50:380–6.

3. Mewar D, et al. (2006) *Independent associations of anti-cyclic citrullinated peptide antibodies and rheumatoid factor with radiographic severity of rheumatoid arthritis.* Arthritis Res. Ther. 8:R128.

4. Eyre S, et al. (2012) *High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis*. Nature Genet. 44:1336–40.

5. Danila MI, et al. (2015) *The role of genetic variants in CRP in radiographic severity in African Americans with early and established rheumatoid arthritis.* Genes Immun. 16:446–51.

6. Govind N, et al. (2014) *Immunochip identifies novel, and replicates known, genetic risk loci for rheumatoid arthritis in black South Africans*. Mol. Med. 20:341–9.

7. Yang SK, et al. (2015) *Immunochip analysis identification of 6 additional susceptibility loci for Crohn's disease in Koreans.* Inflamm. Bowel Dis. 21:1–7.

8. Isobe N, et al. (2015) *An ImmunoChip study of multiple sclerosis risk in African Americans.* Brain. 138(Pt 6):1518–30.

9. Ramos PS, Shedlock AM, Langefeld CD. (2015) *Genetics of autoimmune diseases: insights from population genetics.* J. Hum. Genet. 60:657–64.

10. Okada Y, et al. (2014) *Genetics of rheumatoid arthritis contributes to biology and drug discovery.* Nature. 506:376–81.

11. Hughes LB, et al. (2010) *Most common single nucleotide polymorphisms associated with rheumatoid arthritis in persons of European ancestry confer risk of rheumatoid arthritis in African Americans*. Arthritis Rheum. 62:3547–53.

12. Elshazli R, Settin A. (2015) *Association of PTPN22 rs2476601 and STAT4 rs7574865 polymorphisms with rheumatoid arthritis: a meta-analysis update.* Immunobiology. 220:1012–24.

13. Gregersen PK, Silver J, Winchester RJ. (1987) *The shared epitope hypothesis. An approach to understanding the molecular genetics of susceptibility to rheumatoid arthritis*. Arthritis Rheum. 30:1205–13.

14. Gorman JD, et al. (2004) *Impact of shared epitope genotype and ethnicity on erosive disease: a metaanalysis of 3,240 rheumatoid arthritis patients*. Arthritis Rheum. 50:400–12.

15. Viatte S, et al. (2015) *Association of HLA-DRB1 haplotypes with rheumatoid arthritis severity, mortality, and treatment response*. JAMA. 313:1645–56.

16. Mattey DL, et al. (2007) *Association of DRB1 shared epitope genotypes with early mortality in rheumatoid arthritis: results of eighteen years of follow-up from the early rheumatoid arthritis study*. Arthritis Rheum. 56:1408–16.

17. Hughes LB, et al. (2008) *The HLA-DRB1 shared epitope is associated with susceptibility to rheumatoid arthritis in African Americans through European genetic admixture.* Arthritis Rheum. 58:349–58.

18. Reynolds RJ, et al. (2014) *HLA-DRB1–Associated rheumatoid arthritis risk at multiple levels in African Americans: hierarchical classification systems, amino acid positions, and residues*. Arthritis Rheumatol. 66:3274–82.

19. Lee HS, et al. (2004) *Increased susceptibility to rheumatoid arthritis in Koreans heterozygous for HLA-DRB1*0405 and *0901.* Arthritis Rheum. 50:3468–75.

20. Terao C, Raychaudhuri S, Gregersen PK. (2016) *Recent advances in defining the Genetic basis of rheumatoid arthritis.* Annu. Rev. Genomics Hum. Genet. 17:273–301.

21. Weidinger S, Baurecht H, Naumann A, Novak N. (2010) *Genome-wide association studies on IgE regulation: are genetics of IgE also genetics of atopic disease?* Curr. Opin. Allergy Clin. Immunol. 10:408–17.

22. Knevel R, et al. (2012) *Genetic predisposition of the severity of joint destruction in rheumatoid arthritis: a population-based study.* Ann. Rheum. Dis. 71:707–9.

23. Maehlen MT, et al. (2011) *FCRL3–169C/C genotype is associated with anti-citrullinated protein antibody-positive rheumatoid arthritis and with radiographic progression.* J. Rheumatol. 38:2329–35.

24. Marinou I, et al. (2007) *Association of interleukin-6 and interleukin-10 genotypes with radiographic damage in rheumatoid arthritis is dependent on autoantibody status.* Arthritis Rheum. 56:2549–56.

25. Cantagrel A, et al. (1999) *Interleukin-1beta, interleukin-1 receptor antagonist, interleukin-4, and interleukin-10 gene polymorphisms: relationship to occurrence and severity of rheumatoid arthritis*. Arthritis Rheum. 42:1093–100.

26. Knevel R, et al. (2012) *Genetic variants in IL15 associate with progression of joint destruction in rheumatoid arthritis: a multicohort study.* Ann. Rheum. Dis. 71:1651–7.

27. Teare MD, et al. (2013) *Allele-dose association of the C5orf30 rs26232 variant with joint damage in rheumatoid arthritis*. Arthritis Rheum. 65:2555–61.

28. Pawlik A, Wrzesniewska J, Florczak M, Gawronska-Szklarz B, Herczynska M. (2005) *The –590 IL-4 promoter polymorphism in patients with rheumatoid arthritis*. Rheumatol. Int. 26:48–51.

29. Ceccarelli F, et al. (2011*) Transforming growth factor beta 869C/T and interleukin 6 –174G/C polymorphisms relate to the severity and progression of bone-erosive damage detected by ultrasound in rheumatoid arthritis*. Arthritis Res. Ther. 13:R111.

30. Song GG, Bae SC, Kim JH, Lee YH. (2014) *Associations between TRAF1-C5 gene polymorphisms and rheumatoid arthritis: a meta-analysis*. Immunol. Invest. 43:97–112.

31. Ruyssen-Witrand A, et al. (2014) *Association of IL-2RA and IL-2RB genes with erosive status in early rheumatoid arthritis patients (ESPOIR and RMP cohorts).* Joint Bone Spine. 81:228–34.

32. Mikuls TR, et al. (2006) *Anti-cyclic citrullinated peptide antibody and rheumatoid factor isotypes in African Americans with early rheumatoid arthritis.* Arthritis Rheum. 54:3057–9.

33. Freedman BI, et al. (2014) *End-stage renal disease in African Americans with lupus nephritis is associated with APOL1.* Arthritis Rheumatol. 66:390–6.

34. Bridges SL Jr, et al. (2010) *Radiographic severity of rheumatoid arthritis in African Americans: results from a multicenter observational study*. Arthritis Care Res. 62:624–31.

35. Mikuls TR, et al. (2008) C*igarette smoking, disease severity and autoantibody expression in African Americans with recent-onset rheumatoid arthritis.* Ann. Rheum. Dis. 67:1529–34.

36. Tang Q, et al. (2015) *Expression of interferon- gamma receptor genes in peripheral blood mononuclear cells is associated with rheumatoid arthritis and its radiographic severity in African Americans*. Arthritis Rheumatol. 67:1165–70.

37. Pruim RJ, et al. (2010) *LocusZoom: regional visualization of genome-wide association scan results*. Bioinformatics. 26:2336–7.

38. Achim Zeileis CK, Jackman S. (2008) *Regression models for count data* in R. J. Stat. Softw. 27:1–25.

39. Kichaev G, et al. (2017) *Improved methods for multi-trait fine mapping of pleiotropic risk loci.* Bioinformatics. 33:248–55.

40. Edwards W, Lindman H, Savage LJ. (1963) *Bayesian statistical inference for psychological research.* Psychol. Rev. 70:193–242.

41. Gao X, Starmer J, Martin ER. (2008) *A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms.* Genet. Epidemiol. 32:361–9.

42. Li MX, Yeung JM, Cherny SS, Sham PC. (2012) *Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets.* Hum. Genet. 131:747–56.

43. Gao X, Becker LC, Becker DM, Starmer JD, Province MA. (2010) *Avoiding the high Bonferroni penalty in genome-wide association studies.* Genet. Epidemiol. 34:100–5.

44. Purcell S, et al. (2007) *PLINK: a tool set for whole-genome association and population-based linkage analyses.* Am. J. Hum. Genet. 81:559–75.

45. Trynka G, et al. (2011) *Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease.* Nature Genet. 43:1193–201.

46. Raychaudhuri S, et al. (2012) *Five amino acids in three HLA proteins explain most of the associa- tion between MHC and seropositive rheumatoid arthritis.* Nature Genet. 44:291–96.

47. Barton A, et al. (2009) *Identification of AF4/ FMR2 family, member 3 (AFF3) as a novel rheumatoid arthritis susceptibility locus and confirmation of two further* pan-autoimmune susceptibility genes. Hum. Mol. Gen. 18:2518–22.

48. Stahl EA, et al. (2010) *Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci*. Nature Genet. 42:508–14.

49. Welter D, et al. (2014) *The NHGRI GWAS catalog, a curated resource of SNP-trait associations.* Nucleic Acids Res. 42(Database issue):D1001–06.

50. Ke X. (2012) *Presence of multiple independent effects in risk loci of common complex human diseases*. Am. J. Hum. Genet. 91:185–92.

51. Knevel R, et al. (2013) *Association of variants in IL2RA with progression of joint destruction in rheumatoid arthritis*. Arthritis Rheum. 65:1684–93.

52. Lowe CE, et al. (2007) *Large-scale genetic fine-mapping and genotype-phenotype associations implicate polymorphism in the IL2RA region in type 1 diabetes.* Nature Genet. 39:1074–82.

53. Yang W, et al. (2010) *Genome-wide association study in Asian populations identi es variants in ETS1 and WDFY4 associated with systemic lupus erythematosus.* PLoS Genet. 6:e1000841.

54. Hiwatari M, et al. (2003) *Fusion of an AF4-related gene, LAF4, to MLL in childhood acute lymphoblastic leukemia with t(2;11)(q11;q23).* Oncogene. 22:2851–5.

55. Cen H, et al. (2012) *Association of AFF1 rs340630 and AFF3 rs10865035 polymorphisms with systemic lupus erythematosus in a Chinese population.* Immunogenetics. 64:935–8.

56. Hinks A, et al. (2010) *Association of the AFF3 gene and IL2/IL21 gene region with juvenile idiopathic arthritis.* Genes Immun. 11:194–8.

57. Jansen R, et al. (2017) *Conditional eQTL analysis reveals allelic heterogeneity of gene expression.* Hum. Mol. Genet. 26:1444–51.

58. Lei C, et al. (2005) *Association of the CTLA-4 gene with rheumatoid arthritis in Chinese Han population.* Eur. J. Hum. Genet. 13:823–8.

59. Kormendy D, et al. (2013) *Impact of the CTLA-4/ CD28 axis on the processes of joint inflammation in rheumatoid arthritis*. Arthritis Rheum. 65:81–7.

60. Michel J, Langstein J, Hofstadter F, Schwarz H. (1998) *A soluble form of CD137 (ILA/4–1BB), a member of the TNF receptor family, is released by activated lymphocytes and is detectable in sera of patients with rheumatoid arthritis*. Eur. J. Immunol. 28:290–5.

61. Seo SK, et al. (2004) *4-1BB–mediated immunotherapy of rheumatoid arthritis.* Nature Med. 10:1088–94.

62. van Steenbergen HW, et al. (2015) *IL2RA is associated with persistence of rheumatoid arthritis.* Arthritis Res. Ther. 17:244.

63. Roifman CM. (2000) *Human IL-2 receptor alpha chain deficiency*. Pediatr. Res. 48:6–11.

64. Dendrou CA, et al. (2009) *Cell-specific protein phenotypes for the autoimmune locus IL2RA using a genotype-selectable human bioresource*. Nature Genet. 41:1011–15.

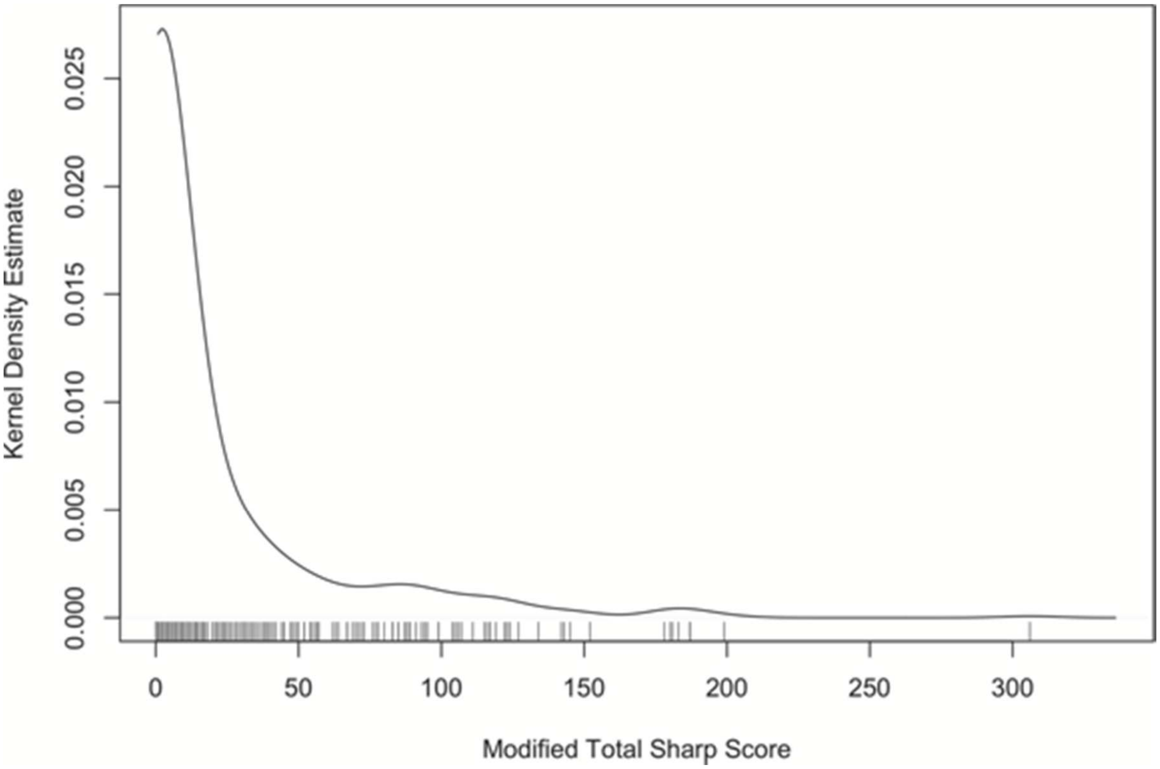
65. Garrett-Sinha LA. (2013) *Review of Ets1 structure, function, and roles in immunity.* Cell. Mol. Life Sci. 70:3375–90.

66. John SA, Clements JL, Russell LM, Garrett-Sinha LA. (2008) *Ets-1 regulates plasma cell differentiation by interfering with the activity of the transcription factor Blimp-1*. J. Biol. Chem. 283:951–62.s

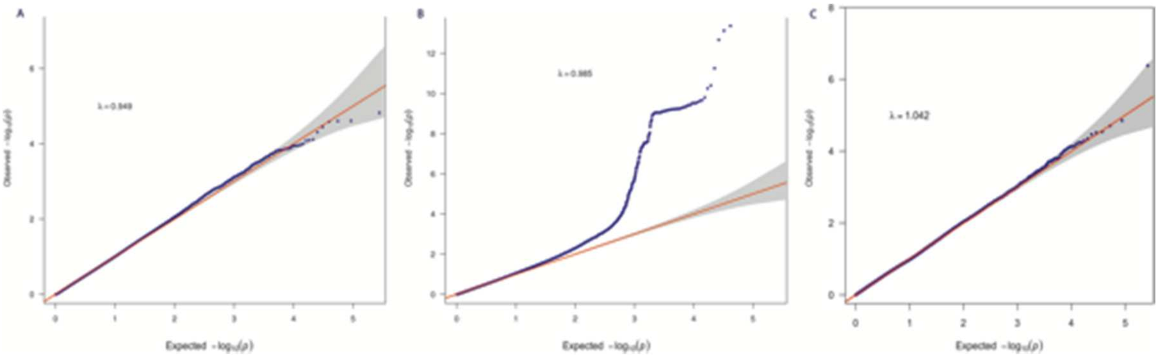
67. Lu X, et al. (2015) *Lupus risk variant increases pSTAT1 binding and decreases ETS1 expression*. Am. J. Hum. Genet. 96(5):731–9.

68. Stranger BE, et al. (2012) *Patterns of cis regulatory variation in diverse human populations.* PLoS Genet. 8:e1002639.

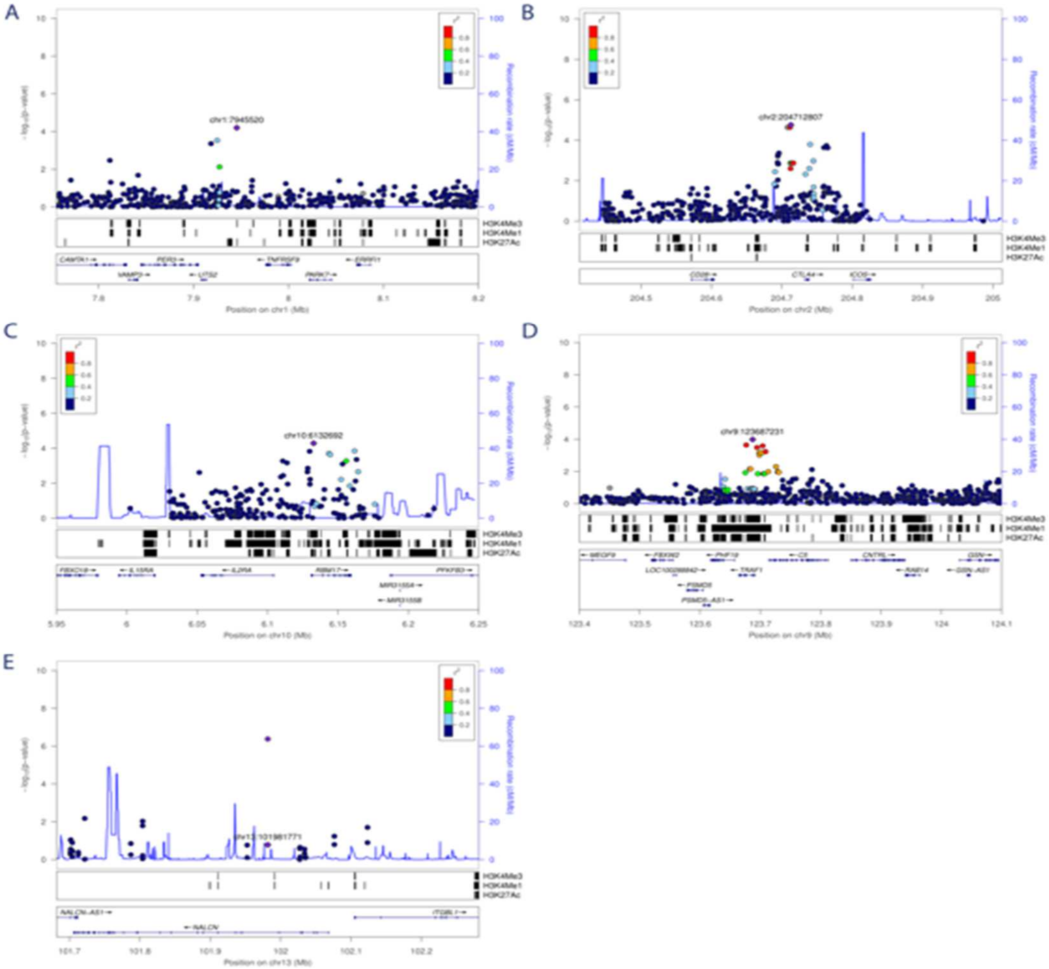
69. Jiang L, et al. (2014) *Novel risk loci for rheumatoid arthritis in Han Chinese and congruence with risk variants in Europeans*. Arthritis Rheumatol. 66:1121–32.



Supplementary Figure S1. Distribution of modified total radiographic scores for our study population using kernel density estimates. Estimates were calculated according to the density R package. Vertical marks are individual subject's overall total modified total Sharp scores on which the density estimates are based.



Supplementary Figure S2. Quantile-quantile plots of p values for the iChip study. P values are from logistic regression (A and B), and from azero-inflated negative binomial model for radiographic severity of RA (C). (A) is with the HLA region excluded (chr6:26,000,000-34,000,000), (B) is with all markers. (C) includes markers from the zero and count portions of the model zero-inflated negative binomial model.



Supplementary Figure S3. Locus zoom plots of suggestively associated genomic loci. For each zoom plot, the x-axis indicates genomic position and the y-axis indicates association strength –log(p). The color bar indicates the strength of LD with the index SNP in purple and is based on LD patterns from persons of African ancestry from the March 2012 release of the 1000 Genomes project. (A) Locus zoom plot showing the association of AFF3 to RA radiographic severity, according to the count portion of the zero-inflated negative binomial model. (B) Locus zoom plot showing the association of TNFRSF9 to RA radiographic severity, according to the count portion of the zero-inflated negative binomial model. (C) The association of CTLA4 to RA radiographic severity according to the zero portion of the zero-inflated negative binomial model. (D) Locus zoom plot showing the association of IL2RA to RA radiographic severity according to the count portion of the zero-inflated negative binomial model. (E) Locus zoom plot showing the association of TRAF1-C5 found to be associated with RA radiographic severity according to the zero portion of the zero-inflated negative binomial model. (F) Locus zoom plot showing the association of NALCN/ITGBL1 to RA radiographic severity. The marker density in this region is low, making it difficult to assess how much support this SNP has from surrounding SNPs in LD.

GENETIC INFLUENCES ON SUSCEPTIBILITY TO RHEUMATOID ARTHRTIS IN THREE GLOBAL POPULATIONS

by

LAUFER VA, TIWARI HK, REYNOLDS RJ, DANILA MI, WANG J, EDBERG JC,
KIMBERLY RP, KOTTYAN LC, HARLEY JB, MIKULS TR, GREGERSEN PK,
ABSHER DM, LANGEFELD CD, ARNETT DK, BRIDGES SL JR.

In submission for *Rheumatic Disease Clinics of North America*

Format adapted for dissertation

Abstract

Large meta-analyses of RA susceptibility in European and Asian populations have been used to identify >100 RA risk loci and inform drug discovery. Despite this, systematic genetic studies of RA in African populations are lacking. We address this disparity with the largest study of RA genetics in African-Americans to date, a two-phase joint analysis of 916 RA patients and 1392 controls, then aggregate our data with >100,000 European and Asian RA patients and controls. We provide evidence of shared effect in 28 risk loci reported in Europeans and Asians. Nevertheless, each population harbors a small number of risk loci specific to it. In African-Americans we identify *GPC5*, *RBFOX1* and *CSMD3* ($p_{AA}<5 \times 10^{-9}$; M_{EAS} and $M_{EUR}<0.2$). Among loci that do not replicate, we observe an enrichment of uncommon and rare variants with large effect sizes – findings that shed light on conflicting reports from the past. Specifically, only 2 of the 16 largest effect index variants in Europeans appear to confer similar risk in both Asian and African populations. Finally, we use CAVIARBF and PAINTOR3 to fine-map >90 shared and population-specific RA risk loci. Addition of African-American genotypes enabled identification of 8 candidate pathogenic variants ($p_{POST}>0.8$), not identified in a previous study, bringing the total above 20. Of these, we highlight rs2233434 in *NFKBIE* and rs3087243 in *CTLA4*, two variants that may influence RA susceptibility as well as RA therapeutic response. Thus, our results illustrate the use of multi-ethnic cohorts to provide information relevant to precision medicine.

Introduction

Rheumatoid arthritis (RA) affects 0.5-1% of populations worldwide [1] and has both environmental and genetic influences [2]. The genetic basis of RA has been explored extensively among persons of European (EUR) and of East Asian (EAS) ancestry [3, 4], but is much less well studied in populations of African ancestry (AFR). Our group [5] and others [3] have previously shown significant overlap between risk loci for RA in African-Americans, but genome-wide association studies (GWAS) of differences in the genetic influences on RA in African-Americans are lacking.

There is growing evidence of important differences in RA heritability among global populations. *HLA-DRB1* is the strongest genetic risk factor for RA in all racial/ethnic groups analyzed [6, 7], in particular among patients with RA-associated serum autoantibodies (rheumatoid factor [RF] and anti-citrullinated peptide/protein antibodies [ACPA]) [8]. Nevertheless, both the effect sizes and the HLA alleles [9] and amino acid residues [10] implicated differ between ethnicities [11]. For instance, in European ancestry RA, a valine residue at amino acid position 11 of *HLA-DRB1* confers risk [10]. In African-Americans, an aspartic acid at position 11 is strongly associated with RA, which is not observed in European ancestry RA [12]. Furthermore, residues at positions 71 and 74 appear to account for more RA risk in Europeans than in African-Americans. Thus, genetic variants in a common RA risk locus may differ among ethnicities, and support the need for more in-depth analysis of RA in African-Americans.

Recent meta- and mega-analyses suggest multiple differences in risk loci outside the HLA region as well [13-15]. Okada et al. conducted a large trans-ethnic mega-analysis of RA in populations of European and Asian ancestry and identified many new RA risk loci not found in either population [3]. Of note, in this analysis only 2 of the 11 largest effect non-HLA RA risk loci found in Europeans, (*TNFAIP3* and *NFKBIE*) were demonstrated to have a concordant effect in Asians; the others either were tested but did not replicate or were not tested due to low minor allele frequency (MAF) [3]. Since both the effect size and the allele frequency of a variant directly impact the heritable proportion of a disease it accounts for, in either event these lead to differences in genetic architecture across populations. One well-known example is *PTPN22*, which harbors the strongest non-HLA risk variant in Europeans (OR_{EUR} = 1.81 MAF = 0.10) but the risk allele is very rare in African and Asian populations, limiting its impact [5, 16]. Other studies provide additional examples of non-MHC risk loci identified in European or Asian RA populations that differ in African-Americans or black Africans [5, 17, 18]. Thus, a major goal of the present study is to quantify similarities and differences among RA susceptibility loci in African-Americans compared to other global populations.

Another major gap in the study of RA genetic risk is to identify variants driving the biology of the risk locus (so-called causal or pathogenic variants) from those merely associated with the disease but not linked to disease biology. Addressing this issue will help expedite mechanistic studies of likely causal variants in RA, which could lead to more precise diagnosis and patient stratification for prognosis and optimal targeted therapies for individual patients [19]. Many approaches to identify pathogenic variants are used, such as the CAVIARBF [20] and PAINTOR algorithms [21, 22], which use differential linkage disequilibrium (LD) and association patterns across multiple populations to assign each disease-associated SNP a posterior probability of being pathogenic. It has been shown that inclusion of African genotypes improves studies to fine-map loci and to identify pathogenic variants [14, 15, 23-25]. Owing to the paucity of available data in African-Americans, fine-mapping approaches have largely been limited to European and Asian ancestry individuals [21]. Thus, a second major goal of our study is to to address this critical "post-GWAS" dilemma by integrating genetic studies of RA susceptibility from three global ancestries to narrow the lists of associated variants (credible sets) to those most likely to be pathogenic.

Our study represents the most comprehensive assessment of genetic risk for RA in African-Americans yet reported. Our analyses were performed in three stages (Figure 1). In Phase I, we conducted an RA GWAS on African-Americans using Omni 1M and 1S arrays and jointly analyzed these results with genotyping data from the Omni 5M array on an additional set of African-American RA and controls. In Phase 2, we merged data our African-American RA data with European and Asian data from Okada *et al.* [3] to conduct a large trans-ethnic meta-analysis (TEMA). In Phase 3, we used the results from the TEMA to conduct trans-ethnic fine-mapping (TEFM) of RA risk loci and identify candidate pathogenic variants.

Our analyses provide several new insights into the genetic basis of RA. We identify a small number of associations in African-Americans that appear to be population-specific. Second, through analysis of the effect sizes of all index variants identified to date, we were able to distinguish true lack of effect from lack of statistical power to detect an effect. We report the number of concordant, discordant, and ambiguous variants in each population. Next, we found several likely pathogenic variants with a posterior probability ($p_{POST} > 0.8$) not previously thought to be pathogenic. These

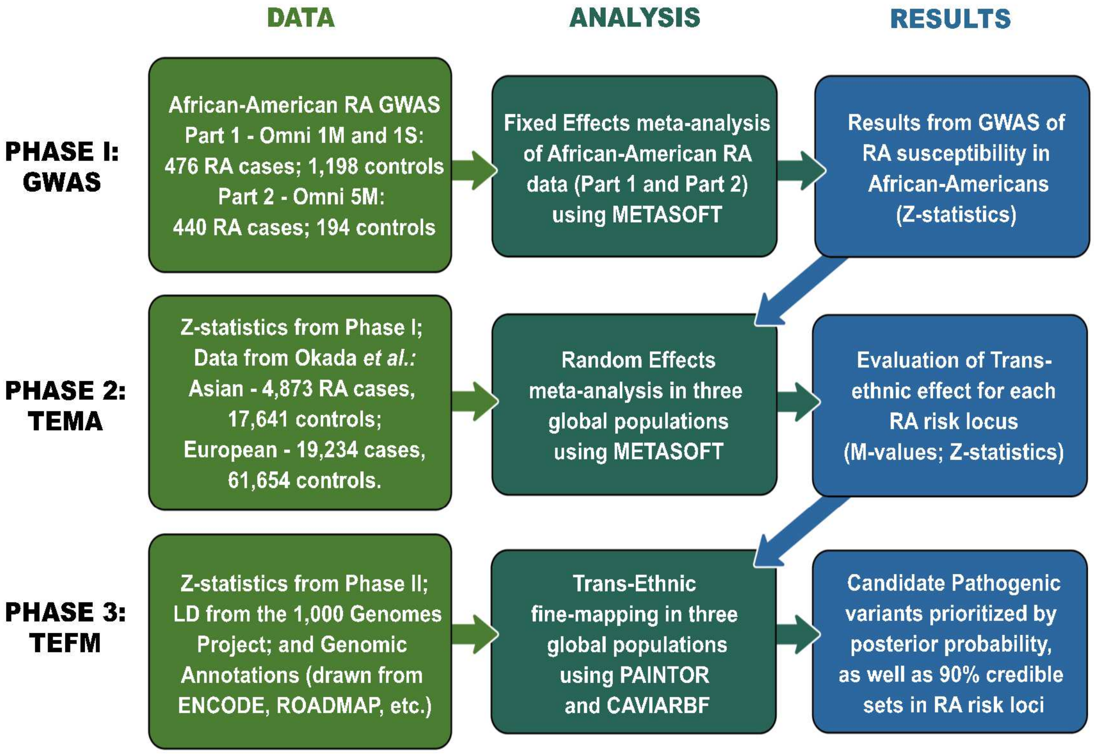

DATA
ANALYSIS
RESULTS
PHASE I: GWAS
African-American RA GWAS Part 1 - Omni 1M and 1S: 476 RA cases; 1,198 controls Part 2 - Omni 5M: 440 RA cases; 194 controls
Fixed Effects meta-analysis of African-American RA data (Part 1 and Part 2) using METASOFT
Results from GWAS of RA susceptibility in African-Americans (Z-statistics)
PHASE 2: TEMA
Z-statistics from Phase I; Data from Okada *et al.:* Asian - 4,873 RA cases, 17,641 controls; European - 19,234 cases, 61,654 controls.
Random Effects meta-analysis in three global populations using METASOFT
Evaluation of Trans-ethnic effect for each RA risk locus (M-values; Z-statistics)
PHASE 3: TEFM
Z-statistics from Phase II; LD from the 1,000 Genomes Project; and Genomic Annotations (drawn from ENCODE, ROADMAP, etc.)
Trans-Ethnic fine-mapping in three global populations using PAINTOR and CAVIARBF
Candidate Pathogenic variants prioritized by posterior probability, as well as 90% credible sets in RA risk loci

Figure 1. Consort diagram showing the components and flow of the study. Phase I is a two-part GWAS of RA in African-Americans, which was jointly analyzed in a fixed-effects meta-analysis. Phase II is a trans-ethnic meta-analysis, in which we combine our data with association summary statistics graciously provided by Okada et al. [3]. We then conduct a random effects meta-analysis using METASOFT. In addition to association p-values (RE2 p-values), we generate M-values, which represent posterior probabilities that an effect exists in a target population [27, 28]. Phase III takes the results of our trans-ethnic meta-analysis (TEMA) and enters them as input to two different trans-ethnic fine-mapping (TEFM) algorithms, CAVIARBF and PAINTOR3. We then construct credible sets for RA risk loci and report variants having posterior probability > 0.8 of being the pathogenic variant in the locus.

findings have improved our understanding of the biology of RA-associated genetic variants and may ultimately lead to better ways to diagnose and treat patients with RA.

Results

Genome-Wide Association Study in African-Americans (Phase I)

Our study took part in several distinct phases that are visualized in a consort diagram in Figure 1. We initially genotyped 683 subjects on Omni 1M and 1S arrays. This included 476 African-American RA patients (421 from CLEAR and 55 VARA; see Table S1A); and 207 African-American controls (38 from CLEAR and 169 from the local Birmingham area; see Table S1B). For descriptions of cohorts, see Methods section and Table S1. We also used previously existing Omni 1M and 1S genotyping data from 991 additional out-of-study African-American controls from SLEGEN [23]. We also performed genotyping of an independent set of 634 African-American subjects on the Omni 5M array (440 RA patients and 194 controls; Table S1). Summary statistics (beta and standard error) from all three sets of arrays (1M, 1S, 5M) were calculated jointly[26] in a fixed-effects meta-analysis of RA in African-Americans (916 RA and 1,392 controls) using METASOFT [27, 28]. Consistent with results of genetic studies of RA in other populations [2-4], we found a strong association of the *HLA-DRB1* region with RA in African-Americans (Figure S1). No other associations reached our genome-wide significance threshold of 5 x 10^{-9}, but several variants were suggestively associated (Figure S1) in phase I.

Trans-ethnic meta-analysis of RA (Phase 2)

The summary statistics from this fixed-effect meta-analysis were then combined with summary statistics from European ancestry and Asian ancestry RA from Okada et al. [3]. Data from Okada included 19,234 European RA and 61,654 Eur controls and

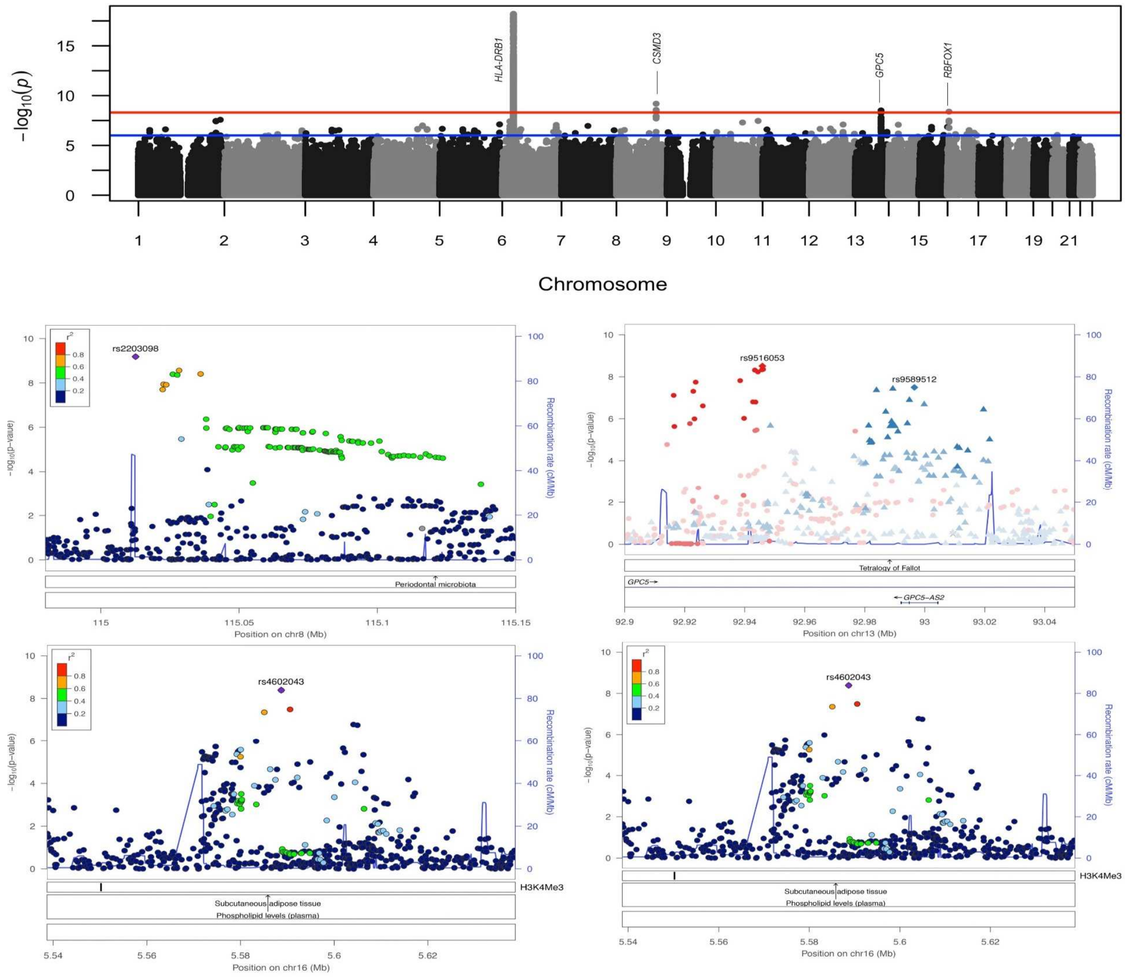


Figure 2A is a manhattan plot of p-values from association testing of our phase II data (African-American GWAS) regression under an additive genetic model. The x-axis indicates chromosome and position, the y-axis indicates –log$_{10}$(p). The horizontal red line is drawn at 5x10^{-9} and is the threshold for genomewide significance, the blue line is for suggestive evidence of association and is drawn at 1*10^{-6}. Zoom plots for each of the associations detected only in African-Americans are presented in Figures 2B-D. Figure 2B-D are Locus Zoom plots of novel associations with RA in African-American populations. Figure 2B shows chr8:114,980,000-115,150,000 (nearest gene: *CSMD3*). Figure 2C plots chr13:92,900,000–93,050,000, in an intronic region of *GPC5*. In this figure, red coloration indicates the variants are more strongly linked to rs9516053, while blue coloration indicates variants more strongly linked to rs9589512. Figure 2D plots chr16: 5,538,689-5,638,689, in an intronic region of *RBFOX1*. In Figure 2E, the independence of the associations of the *PADI2* and *PADI4* loci is shown. On the y-axis is –log10(combined p-value) from the trans-ethnic meta-analysis of all three global populations. The x-axis is genomic position. Red coloration indicates the variants are more strongly linked to rs761426 (the index variant in *PADI2*) than rs2301888 (the index variant in *PADI4*). The blue line indicates genomic recombination rate, which is why independent signals are found so close together.

4,873 East Asian RA and 17,641 East Asian controls. This dataset was then used to perform a genome-wide trans-ethnic meta-analysis (TEMA) using METASOFT [27] [28]. Based on this TEMA, we calculated p-values for all three populations together (p_{TE}), but also for each population individually (p_{AA}, p_{EUR}, and p_{EAS}).

In Phase II, the genome-wide trans-ethnic meta-analysis (TEMA) identified three novel genetic associations with RA in African-American patients and controls ($p_{AA} < 5x10^{-9}$) (Figure 2A). Figure 2A shows a Manhattan plot of genome-wide associations with African-American RA patients and controls ($p_{AA)}$ from this TEMA. These index variants in these novel risk loci are: rs2203098 in *CSMD3*, rs9516053 in *GPC5*, and rs4602043 in *RBFOX1* (see Figures 2B-D; Table 2; Table S3). These associations appear to be specific to African-Americans (M_{AA}=1.0) as they were not found in previous studies of RA in Europeans or Asians (M_{EUR}<0.2; M_{EAS}<0.2). Briefly, M_i is the posterior probability that the effect exists in a study i (see Methods or [28]).

In addition to these findings in African-Americans, we discovered a fourth association in our appears to be present in all three populations ($p_{TE} < 5x10^{-9}$; M_{AA}, M_{EUR}, and $M_{EAS} > 0.8$). The *PADI2* locus was not reported in the prior meta-analysis we used as a reference [3], but it was subsequently reported as such [29]. This association maps to a group of variants in intron 14 of *PADI2*, for which rs761426 was the most strongly associated SNP (OR=0.90; 95% CI: 0.88 - 0.914; p_{TE}=2.48 x 10-10; see Figure 2E). Okada *et al.* performed a conditional analysis of the *PADI4* locus and identified an independent significant association signal at *PADI2* (rs761426, adjusted $P = 2.3 \times 10^{-9}$) [29]. The RA risk T allele of rs761426 has a cis-eQTL effect that increases *PADI2* mRNA expression in whole blood ($P = 4.6 \times 10^{-12}$) [29]. Our data are consistent with rs761426 being both the index variant and the functional variant producing the disease biology in African-American populations as well as others. As PAD2 encodes a peptidyl arginine deiminase, this variant likely exerts its effect by increasing PAD2 expression which increases citrullinated neoantigen production (see Discussion). We performed a conditional analysis of African-Americans with RA to confirm the independence of the association signals in *PADI2* and *PADI4*. We calculated linkage between all strongly associated variants in *PADI2* and *PADI4*, but found 0 pairs of variants across the two loci that had r2 > 0.10. The most strongly linked pair was rs11203290 in *PADI2* and rs12131500 in *PADI4*, which have $r^2 = 0.089$ in European populations. Thus, the association of *PADI2* with RA appears to be independent of the association of *PADI4* in European, Asian, and African-American ancestries.

Assessment of Validated Risk Loci in African-Americans with RA

To assess the effect of established risk loci from other populations, we performed detailed analyses using METASOFT [27, 28]. Given our smaller sample size, the lack of an association of a known risk locus in African-Americans could be due either to insufficient statistical power to detect an effect that exists (type II error); or the true absence of an effect (*Beta* = 0). To address this concern, we first calculated M-values for the 101 index variants identified by Okada *et al.* [3]. The M-value assigned to a genetic variant is similar to a posterior probability that a variant is pathogenic and produces the association signal found in the risk locus (see Methods). Then, to visualize the relationship between M-values and association p-values of these variants, we provide P-

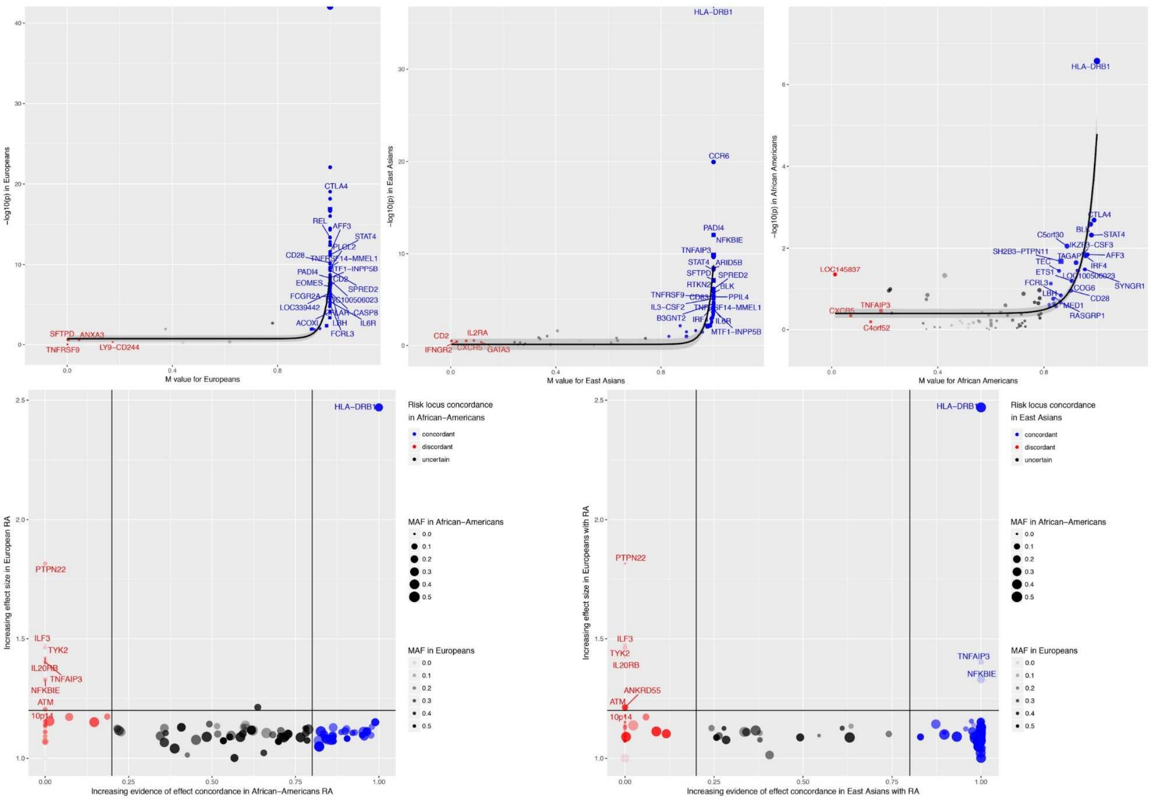


Figure 3A-E. P-M plots in Europeans (3A), Asians (3B) and African-Americans (3C). M-values are plotted on the x-axis. Here, larger values indicate increased likelihood the variant has an effect in the population plotted. Variants with M > 0.8 are colored blue and are considered to have evidence supporting effect replication in that population for the purposes of this study. Variants with M < 0.2 are colored red and are considered to have evidence against replication. Negative log10 association p-value (y-axis) is plotted on the y-axis for the same population. indicate that the index variant plotted like exerts an effect in that population. Further details can be found in the original description of the m-value [28]. Figure 3D-E. Scatter plots of effect size in Europeans with RA versus M-value in East Asians and African-Americans with RA. Here, Europeans are taken as population 1 because studies on this population have been larger than others to date. In these plots, the saturation (alpha) is proportional to MAF in Europeans, while the size of the dot is proportional to the minor allele frequency in African-Americans (Figure 3D) or East Asians (Figure 3E). Here, odds ratios have been coerced to be greater than 1 by first taking the absolute value of beta, then exponentiating. Increasing odds ratio in Europeans (OR_{EUR}) on the y-axis versus M_{AA} or M_{EAS} on the x-axis.

M plots [30] for all three populations in Figure 3A-C. Of the initial 101 RA risk variants studied, 18 could not be assigned an M-value in African-Americans due to low allele frequency. Of the remaining 83 RA variants, 51 variants had intermediate M values (0.2 – 0.8), signifying that there was insufficient evidence to interpret the effect size as either lack of effect or lack of statistical power (see Table S3 for detailed data). Importantly, however, 28 variants had MAA ≥ 0.8 (Table 3), suggesting that the effect size for African-Americans is similar to that in Europeans and East Asians [28]. In comparison, the number of variants showing an effect in East Asians similar to that in Europeans is 59, but this difference likely reflects better statistical power.

There was evidence of effect discordance ($M_{AA} < 0.2$) in 4 loci (Table 3) in African-Americans, and for an additional 18 variants, no M-value was calculated due to differences in allele frequency. In East Asians, the results are similar: there are 5 variants with MEAS < 0.2 (in *CD2*, *IFNGR2*, *CXCR5*, *IL2RA*, and *GATA3*; see Table S3) and no M-value was calculated for 18 variants due to low MAF.

Upon examining the 22 discordant variants (here we refer both to the 4 variants that had $M_{AA} < 0.2$; and 18 that had a MAF too low to test) we noted specific phenotypes compared to other RA risk variants. These discordant alleles were more likely to have large effect size in Europeans ($OR_{EUR} > 1.25$ or $OR_{EUR} < 0.8$) and to be coding variants. In fact, when the index variants were sorted by effect size in the European population, only 2 (*HLA-DRB1* and *NFKBIE*) of the 16 variants with the largest effect size had both MAF > 0.05 and M-value > 0.2 in at least one of the other populations (Table S3).

Overall, we find the risk variants of strongest effect in Europeans are much more likely not to replicate (M-value or MAF < 0.05) in both Asian and African-American populations (here, both Asians and African-Americans are plotted against effect size in Europeans due to the larger study size of the latter population). This is shown in Figure 3D, which plots the effect size for all 101 risk variants in European populations against the M-value for African-Americans with RA (the 83 M-values index variants for which M_{AA} were calculated as well as the 18 index variants not plotted in Fig. 3A-C due to low MAF). Specifically, among African-Americans, the following large-effect loci (in Europeans) are found at low risk allele frequency: *PTPN22*, *TYK2*, *IL20RB*, *ATM*, *10p14*, *DNASE1L4*, and *TNFAIP3* (see Figure 3D). In addition, there is evidence that several variants lack an effect in African-Americans ($M_{AA} < 0.2$): *TNFAIP3*, *CXCR5*, *LOC145837* and *c4orf52*. An analogous plot is presented for East Asian populations in Figure 3E. In East Asians, the following large-effect risk loci (again in Europeans) are found at low risk allele frequency: *PTPN22*, *ILF3*, *TYK2*, *IL20RB*, *ANKRD55*, *ATM*, *10p14*, *DNASE1L4*, and *TNFAIP3* (see Figure 3E). One variant appears to lack a true effect in East Asians (*CXCR5*), and may contribute to RA susceptibility only in Europeans.

Association Enrichment Analysis using MAGENTA

It is thought that the index variant is the disease-producing variant in the locus in only a small minority of cases [31]. Thus, an analysis based only on index variants could suffer from type II error. To account for this, we constructed an association enrichment analysis using MAGENTA [32]. Specifically, we scanned loci not shown to be concordant at the index variants for enrichment of genetic associations across the entire gene. In one case – *IL3/CSF2* – the index variant had low allele frequency and was not associated with RA in African Americans, but both *IL3* and *CSF2* were enriched for genetic associations in a scan conducted using MAGENTA [32] ($p_{AA} = 6.21 \times 10^{-3}$ and $p_{AA} = 7.16 \times 10^{-3}$; see Methods). This indicates that *IL3/CSF2* locus may in fact contribute to risk of RA in African Americans, but perhaps through different variants. This conclusion is generally supported by the results of our trans-ethnic fine-mapping experiments, but an in-depth analysis of this locus is needed to decide this with certainty. For all other loci, neither analysis of the index variant (using M-values) nor analysis of the remainder of the locus (using MAGENTA) provided evidence of an association.

Trans-ethnic fine-mapping and prioritizing candidate pathogenic variants

In Phase 3, we used the association summary statistics from Phase 2 to conduct a trans-ethnic fine-mapping (TEFM) analyses using CAVIARBF and PAINTOR3 [21, 22]. This analysis aimed to identify the variants most likely to be pathogenic within each RA risk locus (see Methods). In general, we found the results produced by the PAINTOR3 algorithm to be more reliable than those computed using CAVIARBF. This is likely because CAVIARBF does not yet integrate data from multiple populations into a single fine-mapping experiment, which is a key point given the strengths of our study. As a result, though we generated estimates using CAVIARBF, we relied on those from PAINTOR3, and those are the values reported elsewhere in the manuscript (e.g. Table 4 and Figure 4). We analyzed 98 non-HLA RA risk loci (those in the HLA region have been extensively studied). Consistent with the report of Kichaev *et al.,* we validated that the following are likely to be pathogenic variants (posterior probability > 0.8): rs2476601 in *PTPN22*, rs7731626 in *ANKRD55*, rs147622113 in *ILF3*, rs909685 in *SYNGR1*, rs1893592 and rs12715125 in *EOMES*, rs968567 in *FADS2*, rs657075 in *IL3/CSF2*, and rs71508953 in *ARID5B* (see Table 4). This overlap between our findings is not surprising granted the large overlap of our datasets, which both used the European and Asian RA samples from Okada *et al.* [3].

Importantly, however, despite being a relatively small proportion of the total dataset, addition of the African-American data enabled identification of 9 additional variants highly likely to be pathogenic (posterior probability > 0.8). These include: rs3087243 in *CTLA4* (see Figure 4), rs72634030 in *C1QBP*, rs34536443 in *TYK2*, rs706778 in *IL2RA*, rs10774624 in *SH2B3/PTPN11*, rs7902146 in *ARID5B*, rs2812378 in *CCL19-CCL21*, rs2233434 in *NFKBIE*, and rs13330176 in *IRF8* (see Table 4). Several of the variants identified are either known to be the pathogenic variant in the locus or are the subject of recent and ongoing studies (see Discussion).

In most risk loci, no single variant with $p_{POST} > 0.8$ was identified. Nevertheless, in some of these cases, the variants in the credible set nevertheless suggest concrete directions. For example, the 80% credible set in the *IFNGR2* locus was comprised only of rs9974603 ($p_{POST} = 0.59$) and rs9975155 ($p_{POST} = 0.24$). Both of these variants lay >25 bp from a conserved transcription factor binding site (TFBS): rs9974603 is ~25 bp from the TFBS of the transcription factor E2F6, and rs9975155 is >10bp from that of *ZBTB7A*, directly within the 5'UTR of *IFNGR2*. A recent study demonstrated that a large proportion of RA pathogenic variants lay within loci occupied by the Epstein Barr viral proteins EBNA2 and EBNA3C [33]. This study suggested that autoimmune risk variants exert their effects by allele-dependent binding events in loci known to interact with

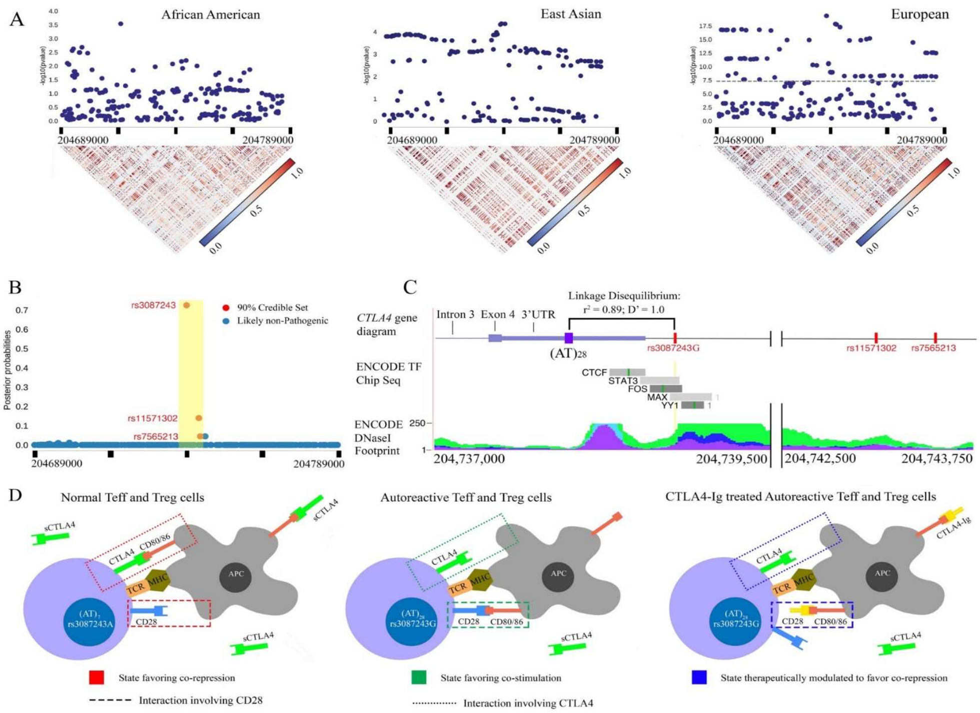


Figure 4. The association and putative biological rationale for the association of rs3087243 in the *CTLA* locus. A. (Top) – Scatter plots of the association (-log10 p-value) of genetic variants in the CTLA4 locus (chr2:204,689,000-204,789,000) in African (left), East Asian (middle) and European (right) to RA susceptibility. Bottom Panels are heatmaps of linkage disequilibrium between variants measured with r^2. B. (Middle, left) - Scatter plot of the posterior probability of pathogenicity for each variant found within chr2:204,689,000-204,789,000. Variants in red belong to the 90% credible set, other variants are colored in blue. C. (Middle, right) - Gene diagram of the region containing the most likely candidate pathogenic variants. The region in C is highlighted in light yellow in panel B. The top track is a gene diagram of the final intron and exon, the 3'UTR and intergenic region downstream of *CTLA4*. There is a dinucleotide repeat in the 3'UTR of *CTLA4* in linkage marked $(AT)_{28}$ with rs3087243-G. The $(AT)_{28}$ variant of this short tandem repeat decreases *CTLA4* mRNA levels in autoreactive T cell lines. The second track shows transcription factor binding sites (TFBS). Variants in red are the same as the credible set found in B. The bottom track shows raw DNAse hypersensitivity signal in Tregs (yellow), Th17 cells (green), Th1 cells (green), and naïve T cells. D. (Bottom) – Model of effects of $(AT)_{28}$ and CTLA4 CT60G genotype. Normal T cells (bottom, left) having $(AT)_7$ and CTLA4 CT60A genotype express more CTLA4 than autoreactive cells (compare red dotted box to green dotted box), tipping the balance toward co-repression. This difference in CTLA4 level results in a normal state in which CD28 is more likely to remain unbound by CD80/86 (green dashed box) than compared to autoreactive T cells (red dashed box). However, the RA therapeutic CTLA4-Ig competes with CD28 to bind CD80/86 (blue dashed box), restoring a balance between co-repression and co-stimulation. This may serve to counteract decreased *CTLA4* expression and levels resulting from a risk genotype or other cause (compare blue dotted and dashed boxes to red and green).

EBNA proteins. How this hypothesis may relate to these variants in *IFNGR2* is treated in the Discussion.

Therefore, although this locus contains 2 variants in the credible set instead of 1, both the position of the implicated variant (in the TFBS marking but outside the consensus binding site) and the biological effect are extremely closely related. Anecdotally, we observed many other loci that contain very interesting results but had several variants in the credible set. We report all the credible sets containing 5 variants or fewer in extended data.

Discussion

Novel associations with RA susceptibility in African-Americans

Our detailed analyses of African-Americans has identified several new RA risk loci (p value below our threshold of 5 x 10-9): *CSMD3*, *GPC5*, *RBFOX1*, and *PADI2*. *CSMD3* is a 73 exon gene stretching ~1.2Mb across 8p23. Interestingly, variants near its homolog *CSMD2* were also suggestively associated with RA (rs55798295, p = 2.84 x 10-7). This family of molecules (*CSMD1*, *CSMD2*, and *CSMD3*) appear to be involved in complement-mediated synapse pruning in the CNS. *CSMD3* is associated with immune phenotypes such as influenza infection and asthma [34, 35]. *CSMD3* might contribute to RA by decrease inhibition of complement activation, but the remarkable pleiotropy of these genes suggests multiple other explanations. *RBFOX1* is a regulator of alternative splicing of mRNA that has been extensively implicated in neural phenotypes [34]. However, there is also evidence that it influences many immune processes, including TCR and BCR receptor signaling, leukocyte migration and differentiation [36].

We also found association of RA with a variant near glypican-5 (GPC5), a pleotropic gene associated with several immune phenotypes, including multiple sclerosis in African-Americans [34]. Glypicans are components of proteoglycans that appear to influence the behavior of the extracellular matrix during development and cellular proliferation. Another glypican molecule, *GPC3*, is a well-studied oncogene that serves as a ligand to *CTLA4* (*CD152*) on the surface of $CD4^{+}CD25^{-}$ T-cells, thereby influencing cellular proliferation and TNF production [37].

GPC5 is also a tumor suppressor gene, but appears to inhibit tumor growth by suppressing WNT/B-catenin signaling [38]. Polymorphisms in *GPC5* have been associated with multiple sclerosis in Norwegian [39], Spanish [40], and African-American populations [41], as well as with IFN-B response in MS [42]. A trans-eQTL in the *GPC5* locus appears to downregulate Proliferating Cell Nuclear Antigen (PCNA) associated factor, which may result in increased proliferation of CD4+ T-cells in the RA synovium, or in synovial fibroblasts [43]. Glypican-5 is also known to modulate blood protein levels through interaction with *PRKCQ* [44], another RA risk gene that is highly expressed on T-lymphocytes. In light of these observations, we speculate that variants in the *GPC5* locus may alter CD4+ T-cell behavior, contributing to dysregulated ECM growth in the RA synovium. Overall, while replication of this locus is needed, prior findings indicate both a known role for this gene in autoimmunity. African-American population with MS and RA appear to share the IFN-B response genes *IRF5*, *IRF8*, and now *GPC5*. While IFN-B never emerged as an effective treatment for RA, findings from MS indicate and that this risk gene may be druggable.

Our trans-ethnic meta-analysis identified an association with *PADI2* in European, Asian, and African-American populations (M-value for each population > 0.95). Ours is the first report implicating *PADI2* as an independent genetic risk factor for RA in African-Americans (Figure 2E). *PADI2*, *PADI3*, and *PADI4* have distinct specificities against cellular substrates, which has important implications regarding autoantigen selection in RA [45]. Both genetic and experimental data suggest the association of *PADI2* with RA is independent of that of *PADI4* [29, 46], a known risk allele for RA and a key enzyme in RA due to its role in citrullination and the generation of the ACPA response [47]. Recent clinical evidence shows that not only PAD4 but PAD2 protein level and activity are increased in synovial fluid of RA compared to those with osteoarthritis [48]. Likewise, a recent study of TNF-induced arthritis in mice showed that *PADI2* deficiency led to decreased numbers of plasma cells, decreased serum IgG levels, and decreased clinical and pathological findings [49]. In addition, citrullination was nearly absent from the ankles of *PADI2*-deficient, but not *PADI4*-deficient, mice [49]. By contrast, in humans with RA, PAD2 level in the joint is elevated, very likely owing to two distinct mechanisms. First, the T allele of rs761426 (which was the index variant in our study) appears to increase expression of *PADI2* mRNA in the whole blood [29]. Second, suppression of *PADI2* expression mediated by microRNA miR-4728-5p (a suppressor of PADI2 expression) appears to be reduced in RA. Of note, not only *PADI2* but the locus containing miR-4728-5p itself is also has a genetic association in European, East Asian [3], and African populations (rs59716545; $p_{TE} = 2.60 \times 10^{-13}$; $M_{EUR} = 1$, $M_{EAS} = 1$, $M_{AA} = 0.95$). Thus, RA risk variants appear to affect expression of *PADI2* in RA targets appear to act co elevated levels of PADI2 in RA.

With regard to *PADI4*, the index variant is rs2301888, is in near total linkage with rs2240335. rs2240335 codes a synonymous mutation of the second amino acid of an alternative transcript for PAD4. This variant is an eQTL for PAD4 expression the A allele of rs2240335. This change was recently shown to increase *PADI4* expression in neutrophils. Of note, the authors of this study also describe a statistically significant, 3-fold enrichment of neutrophil eQTLs among other RA risk variants. Once expressed, *PADI4* citrullinates histones and other proteins that initiate NETosis (a central process in RA pathobiology), particularly when in the presence of RF in an oxidative environment. As a result, increased expression of *PADI4* based on rs2240335 genotype fits well with current understanding of the relationship between PAD enzymes and autoantibody-mediated pathogenicity. Overall, the results of our association testing as well as our trans-ethnic fine-mapping of *PADI2* and *PADI4* loci support the conclusion that genetic variation in these loci and their regulators contributes to RA pathogenesis by increasing expression of these enzymes in all three global populations. Once produced, PAD2 and PAD4 catalyze citrullination, resulting in a greater number of citrullinated neoantigens and thereby driving autoantibody-mediated RA pathobiology.

Insights gained from analysis of previously validated RA risk loci

Examples of effect discordance of RA risk loci in African-Americans compared to other ethnicities has been observed previously (e.g. lack of influence of rs2476601 in PTPN22). our study suggests these variants tend to be easier to isolate using trans-ethnic fine-mapping. Indeed our in-silico approach successfully isolated large effect coding variants with $p_{POST} > 0.99$ in loci such as *PTPN22*, *ILF3*, and *TYK2* (Table 4 and discussed further below). The presence of this discordance may have a future impact on precision medicine. Treatments that target the pathways containing these risk genes might vary in efficacy among persons of various ethnic backgrounds. As Okada *et al.* note, *ILF3* and *TYK2* (the third and fourth strongest genetic risk variants for RA in European and Asian populations) both have protein-protein interactions with the IL-6 receptor (soluble and membrane bound IL-6Rα) which are targeted by RA drugs toclizumab and sarilumab. The RA-associated *ILF3* and *TYK2* variants are uncommon in European populations, but very rare or totally absent in East Asians and Africans. Therefore while our results agree with prior findings that the genetic basis of RA is mostly shared, differences do exist, which may have important clinical consequences.

Candidate pathogenic variants

We used the results from our meta-analysis (Phase 2) to guide our fine-mapping studies (Phase 3). We compared our results to those of Kichaev et al. [21], who also used the European and Asian set analyzed by Okada et al. [3]. Kichaev *et al.* reported a posterior probability of pathogenicity > 0.8 for 12 of the 101 loci associated with RA. Adding our African-American RA dataset to the same dataset of Europeans and Asians from Okada et al. [3] yielded important similarities and differences. We generated posterior probability estimates for the same 12 variants, which were largely concordant with prior results (Table 4). However, the addition of our data, though only 4% of the total data by number of study participants, enabled identification of eight additional novel candidate pathogenic variants.

While some of these variants are not well-studied, several of these variants are known functional polymorphisms. For instance, we obtained a posterior probability estimate of 1 for rs2476601, a well-known autoimmune risk variant, and a posterior probability of 0.99 for rs34536443, which is a loss-of-function variant that alters catalytic ability of *TYK2*. We also identified the recently characterized variant rs909685 in *SYNGR1*. This SNP was recently shown to disrupt the binding site of PITX3, which imparts allele specific expression (ASE) of *SYNGR1* at baseline. Of note, the variant we identified in the 5'UTR of REL also acts by disrupting a conserved residue in the CTCF canonical binding motif other examples can be seen in Table 3B.

Inclusion of data from African-American populations allowed specific insight in two risk loci. These are rs2233434-rs2233424 in *NFKBIE* pPOST = 0.99), and rs3087243 in CTLA4 (pPOST = 0.80). First, *NFKBIE* is the second largest effect non-HLA risk locus that has strong evidence of trans-ethnic support, with an OR near 1.25 in each population. A recent study identified two functional nonsynonymous variants in *NFKBIE*: the G allele of rs2233434 (Val194Ala) G allele and the C allele of rs2233433 (Pro175Leu). This study showed that both of these variants increase NF-κB activity upon stimulation of HEK293A cells with TNF-α, with rs2233433 having a stronger effect, and demonstrated an expression imbalance in individuals heterozygous for rs2233434 [50]. The same study showed dose-dependent inhibition of NFKBIE expression during exposure to vectors carrying the non-risk allele of rs2233434. Thus, functional data suggest that either of these variants could contribute to RA risk in European and Asian populations. However, according to our analysis rs2233434 ($p_{POST} = 0.482$) is >400 times more likely to be the pathogenic variant than rs2233433 ($p_{POST} = 0.001$). We investigated why one variant was so much more strongly prioritized granted that these two variants are strongly linked in European and Asian populations ($r^2 = 1$) and both have been shown to produce functional effects. We found that the T allele of rs2233433 is completely absent from West African populations.

By contrast, rs2233434 appears to be present and account for the association signal in African-Americans. Because rs2233433 is absent but rs2233434 is present in Yoruban populations, but the effect on RA risk is still detected in Yoruban populations, we suggest that rs2233434 exerts the effect on RA risk through allele-specific expression, ultimately resulting in decreased inhibition of NF-κB and increased RA risk in all ethnic groups. However, there is a second key point. Imamura et al. provide further support for the role of rs2233424 in RA. They both knockdown and overexpress *NFKBIE* in human RA synovial cells with and without the Val194Ala genetic variant (rs2233434). Following this, they show that methotrexate derivatives accumulate within the cells overexpressing the Val194Ala mutant allele compared to wild-type NFKBIE due to a decrease in SLC19A1 mRNA [51]. Thus, this variant may affect both NF-kB signaling in the RA and methotrexate treatment response in the rheumatoid synovium. Granted the heterogeneity found among large effect loci in this study, this may be but one of many tantalizing illustrations of ways in which inclusion of African samples can aid precision medicine for individuals of African ancestry as well as for all populations.

Our trans-ethnic meta-analysis confirmed the well-known association of CTLA4 to RA susceptibility (Figure 4A) and our trans-ethnic fine-mapping provided improved insight into this association of CTLA4 to RA susceptibility (Figure 4B). Specifically, our 90% credible set included rs3087243 ($p_{POST} = 0.80$) 3'UTR of CTLA-4 as well as rs11571302

- an eQTL for *CTLA4*. rs3087243 itself lies within the binding site for several transcription factors (see Fig 4C), is associated with lower mRNA levels of soluble CTLA-4 [52], and lies within a region of transcription factor binding (including for *STAT3* - Figure 4C). As a result, it may be a "pathogenic" variant that exerts effects on RA susceptibility directly. However, to the best of our knowledge, there is not yet experimental demonstration of a mechanism through which rs3087243 relates to RA. One recent report indicates that the CT60G allele does not alter binding affinity [53]. It is thus possible that a different variant underlies RA risk. Because of the way trans-ethnic fine-mapping algorithms work, if this is true, such a variant is highly likely to be 1) linked to rs3087243, 2) not directly genotyped and 3) not easily imputed. In fact, there is a dinucleotide repeat lying within the 3'UTR of CTLA4 (see Figure 4C) [54] that normally has 7 repeats - $(AT)_7$ – and a variant allele with 28 dinucleotide repeats - (AT)28. It was recently shown that the $(AT)_{28}$ variant is in strong linkage with the G allele of rs3087243, and transfection of $(AT)_{28}$ allele decreases CTLA4 mRNA and protein levels in Jurkat T cells [54]. Because *CTLA4* serves a co-repressor for T-cell activation, decreased RNA and protein levels would be expected to increase risk of autoimmune disease processes whose pathophysiology has T cell component. Thus, we present a model in which either rs3087243, or $(AT)_{28}$, or a haplotype containing both variants, mediates RA risk by decreasing *CTLA4* expression and protein levels (Figure 4D). In addition to identifying elongated $(AT)_n$ elements, we identified 5 other genetic variants in strong LD ($r^2 > 0.8$) with rs3087243 in one or more ethnic population. Because CTLA-4 can compete with CD28 for CD80/86 binding, reduction of its expression could lead to increased immune response in T cells (Figure 4D). CTLA4-Ig is a therapeutic for RA that acts presumably by binding to CD80 and CD86, thereby preventing their interaction with CD28 (Figure 4D). Therefore, results are consistent with the hypothesis that rs3087243 is a functional variant that produces RA risk, but they are also consistent with the possibility that rs3087243 tags a functional (AT)n repeat element that decreases transcription of CTLA-4. Thus, it is tempting to speculate that the presence of such a variant could predispose an individual to RA.

We recently reported trans-ethnic fine-mapping results on the *AFF3* locus recently based on Immunochip genotyping [55]. The results from the current analysis and the Immunochip analysis agree closely. Our approach once again nominated variants near the 5'UTR of *AFF3*, many of which are eQTLs for *AFF3*. In particular, rs9653442 (pPOST = 0.33) and rs6712515 ($p_{POST} = 0.25$) were again found in the credible set [55]. Finally, TEFM of the *IFNGR2* locus identified 2 candidate variants in the IFNGR2 promoter. EBNA3C is known complex with, and drive overexpression of, both E2F6 and ZBTB7A. Both E2F6 and ZBTB7A are transcriptional repressors, and both act on E2F transcription factors including E2F1. In other words, EBNA3C promotes ZBTB7A and E2F6-mediated inhibition of E2F1, which promotes cell proliferation through several prominent RA risk genes involved in the cell cycle, including *ATM*, *CDKN2*, *CDKN4* [56]. The recent manuscript shows that autoimmune risk variants in loci bound by EBNA proteins tend to exert their effects by altering binding of either viral or human proteins. Because EBNA3C accentuates the effects of E2F6 and ZBTB7A, it seems most likely that the C allele of rs9974603 and the T allele of rs9975155 serve to increase binding of these transcription factors to the IFNGR2 promoter, thereby increasing IFNGR2 expression in RA in the manner our lab has reported previously [57]. Prior evidence suggests that increased expression of *IFNGR2* in RA is most likely to occur in T cells [58], B cells, or macrophages [59]. As such, we recommend assays (e.g. EMSA) to demonstrate if these TFs do in fact display allele specific binding and to begin by looking in these cells for differences in IFNGR2 expression differences in RA versus healthy controls. It may also be important to include EBV infection as a group and covariate in these analyses.

Three fine-mapping studies of RA have obtained very different posterior probability estimates for variants in CD28 and *ANKRD55*. Kichaev *et al.* report rs7731626 and rs72767222 in *ANKRD55* [21], while Westra et al. report rs11377254 in *ANKRD55* and rs117701653 in CD28 [53]. We found rs7736126 was a candidate variant ($p_{POST} = 1.0$), but did not confirm any of the others. The differences in could stem from: 1) different study designs (e.g. use of related traits in Europeans only versus use of RA-only data in a trans-ethnic study) 2) differences in imputation quality, imputation method, and additional QC steps (see Supplementary Methods), 3) differences inclusion of rare variants, indels, etc., 4) differences in preparing samples for modeling more than one causal variant per locus. Granted these discrepancies, we advise caution in interpretation of our fine-mapping results in the *CD28* and *ANKRD55* loci.

Our study has several limitations. First, although the present study represents the largest association study of RA conducted in African-Americans, the sample size limits our statistical power to detect associations of variants with low MAF and small effect size in our association study as well as our meta-analysis. Second, we relied on reference data sets to generate LD matrices for the European and Asian samples since genotypes data were not available. However, in several cases we detected a large difference between reported subpopulation-specific allele frequencies from Okada et al. and the 1000 genomes data. In such cases we had to risk miscalibration of LD matrices or accept data loss. Variants with a difference in minor allele frequency > 0.1 were excluded from the analysis. We used a combination of MAF, strand checks, and comparisons of LD and Z-statistics to unambiguously align most SNPs to the 1,000 Genomes reference. Nevertheless, a fraction of SNPs (in particular A/T and G/C SNPs with AF near 0.5) had to be dropped from the study. In addition, indels, short tandem repeats, and other variants that are difficult to impute are missing from the data as well. In the event that such a variant is in fact the pathogenic variant, the posterior probability of that variant can be misattributed to other variants (typically strongly linked variants) so caution is advised. We refer the reader to the analysis of CTLA4 CT60G and the linked $(AT)_{28}$ repeat for an example.

This study accomplishes several related goals. We begin with the largest genome-wide association study in African-Americans with RA to date. We then conduct a trans-ethnic meta-analysis to find concordant and discordant risk loci. We use these results to fuel a trans-ethnic fine-mapping study using published data from three global populations. The results of these studies contain several valuable suggestions. First, our meta-analysis demonstrates widespread differences in RA disease architecture based on differing allele frequency. Second, the inclusion of African Americna genotypes – even though they comprised only ~4% of the total data – nearly doubled the number of candidate pathogenic variants we identified. The importance of this is well-illustrated by rs2233433 and rs2233434 in *NFKBIE* – 2 variants that are in perfect linkage in Asian and European populations but are absent from African-Americans. Our results may help identify biomarkers both for RA disease risk and for methotrexate responsivity, to key goals in RA precision medicine. Because of this untapped potential, we suggest that the use of trans-ethnic cohorts in the future is likely to greatly aid the goals of precision medicine, both for the underserved and for other populations.

In summary, we present the most complete picture of RA in global populations assembled to date. We present evidence that the PADI2 locus is independently associated with RA Europeans, Asians, and African-Americans. In addition, we present evidence that 28 RA risk loci identified in European and Asian populations likely predispose to RA in African-Americans ($m > 0.8$) while 4 likely do not ($m < 0.2$). Finally, we use a trans-ethnic fine-mapping approach to identify an additional 8 high-confidence (pPOST > 0.8) candidate pathogenic variants in RA risk loci. Overall, our study provides a strong rationale to democratize genetic analysis across global populations, both to understand differences in risk architecture and to aid the goals of precision medicine for RA.

Materials and Methods

A. Study patients and controls.

We genotyped a total of 916 cases of autoantibody-positive (ACPA-positive) African-American patients with RA (from the CLEAR Registry and the VARA Registry) and 1370 African-Americans controls without rheumatic disease. Genotyped controls were from the CLEAR Registry and the local Birmingham area. Out-of-study, previously genotyped controls were from SLEGEN – The International Consrtium on the Genetics of Systemic Lupus Erythematosus) (see:https://slegen.phs.wakehealth.edu/public/index.cfm). All RA cases satisfied the 1987 ACR classification criteria [60]. All subjects were self-declared African-Americans and >19 years of age. Genomic DNA from the CLEAR patients with RA and controls was isolated from peripheral blood using standard techniques as previously reported [7].

1. The CLEAR Registry

The CLEAR registry enrolled African-Americans with RA. CLEAR 1 enrolled African-Americans of ≤ 2 years disease duration (from 2000-2006), and CLEAR 2 enrolled African-Americans with RA (not previously enrolled in CLEAR 1) of any disease duration (2006-2011). Participants were enrolled at academic sites: University of Alabama at Birmingham (Coordinating Center); Grady Hospital/Emory University, Atlanta, GA; University of North Carolina, Chapel Hill, NC; Washington University, St. Louis, MO; and Medical University of South Carolina, Charleston, SC. CLEAR controls were African-Americans without rheumatic disease who were age-, sex-, and geographic location-matched (as a group) to the CLEAR RA patients. Controls were screened to exclude rheumatic disease using the validated Connective Tissue Disease Screening Questionnaire (CSQ) [61, 62]. Human subject protocols were approved by the Institutional Review Boards of the each of the participating institutions. Patient sociodemographics, medical history, medications, co-morbid conditions, disease activity measures, and other variables were collected along with blood for extraction of DNA, serum, etc., and radiographs of the hands and feet for scoring of erosions and joint space narrowing (see below), making the CLEAR registry a valuable resource for the study of RA [7, 63, 64][65, 66]. ACPA status was confirmed on all CLEAR participants as previously reported [67]. Characteristics of the CLEAR RA patients and controls are shown in Table 1. The genotyping arrays on which the all participants were assayed in this study is shown in Table S1.

2. The Veterans Affairs Rheumatoid Arthritis (VARA) Registry

VARA is a prospective, observational, multicenter study that includes 12 VA medical centers [68]. 55 African-Americans with ACPA-positive RA from VARA were included. ACPA status was confirmed using a second generation anti-cyclic citrullinated protein ELISA. Demographic and disease characteristics for African-Americans participating in VARA have previously been reported [69].

3. Birmingham Controls

In addition to CLEAR controls, we genotyped healthy African-Americans from the Birmingham, Alabama area, as previously described [70]. Genotype data from previously reported African-American controls from the SLE Genetics consortium (SLEGEN) study were also included in the analysis [70] (see Table S1A).

B. Genotyping methods

1. Genotyping Arrays

Illumina genotyping for all samples was conducted with standard Illumina Infinium protocols and 500 ng of genomic DNA and clustering was performed with Illumina GenomeStudio software. CLEAR and VARA RA cases, CLEAR controls, and Birmingham controls were genotyped on the Illumina Omni 1M array, the Omni 1S array, or both (Table S1A). These samples were hybridized to Omni 1M (Omni 1M Quad v1.0_B) and Omni 1S (Omni 1S_8 v1_H) arrays. We also genotyped CLEAR RA patients and controls on the Illumina Omni 5M Array.

2. Quality Control and Principal Components Analysis

We performed rigorous quality control on each array separately using the same criteria for all arrays. Samples were excluded for any of the following reasons: (i) a call rate >98.5% of the total number of SNPs on the chip (ii) observed heterozygosity rate ± 3 s.d. from the mean (iii) outliers based on PCA were removed based on visual inspection of PCA plots (iv) mismatches in sex designation as determined by genotype versus reported gender (v) IBD > 0.1875 between samples, in which case the sample with the lower call rate was excluded. Markers were excluded for any of the following reasons: (i) call rate >98.5% on each chip separately (ii) Hardy-Weinberg equilibrium (HWE) p-value > 1x10-5 in control samples (iii) minor allele frequency (MAF) ≥ 0.05. We used EIGENSTRAT [71] to estimate principal components. We ensured the selection of SNPs used to generate prinicipal component loadings did not introduce correlations between principal components and cohort membership, plate membership, and other variables of interest. For the Omni 1M and 1S chips, principal components 1, 2, 4, and 6 were included as covariates due to mild correlation with phenotype status; for the Omni 5M chip, principal components 1, 2, 4, and 8 were included. We used data from HapMap3 Utah Residents with Northern and Western European Ancestry (CEU), Yoruba in Ibadan, Nigeria (YRI), and Han Chinese and Japanese from Tokyo (CHB+JPT) to investigate the clustering of our study data. As expected, African-American samples were admixed between the CEU and YRI samples (Figure S2). To control for batch effect, we included batch ID as a covariate in association testing.

3. Imputation

Imputation was carried out using IMPUTE2 [72] with 1000 Genomes Cosmopolitan Samples. SNPs with high missingness across all studies were preferred, and care was taken to ensure there were no large differences between case-control status and data missingness among SNPs used for imputation. The total number of SNPs after imputation was ~22,000,000. After QC based on imputation quality (Info > 0.5) and expected allele frequency (EAF > 0.05), 8,380,626 SNPs remained for analysis. About 0.6% of SNPs with EAF > 0.05 were excluded due to imputation quality. Exclusion of the MHC yielded a lambda value of 1.031 (Figure S3A), indicating little genomic inflation. Exclusion of the MHC yielded a lambda value of 1.03 for both the 1M and 1S arrays, and for the 5M array as well (Figure S3B).

D. Analysis of Genotyping Data

1. Joint Analysis of Omni 1M, 1S and 5M data

We jointly analyzed all of our GWA data from the Illumina Omni 1M and 1S (phase I) and 5M (phase II) arrays. Logistic regression analysis was performed on both genotyped and imputed SNPs using SNPtest v2.5.1 [73] after including sex, cohort ID, and principal components as covariates under additive and dominant genetic models. Setting an appropriate threshold of genome-wide significance has been the subject of substantial inquiry due to the difficulty of establishing an accurate family-wise error rate in the context of widespread LD. Most studies, particularly studies of Europeans subjects, set this threshold at 5*10-8 when testing markers under an additive genetic model.

However, we felt that a more conservative alpha level was appropriate for our study for two reasons. First, compared to European populations, African/African-American populations have smaller average linkage blocks and weaker average LD. This has resulted in a recommendation of ~1-2*10^{-8} for the alpha threshold for genome wide significance in African/African-American populations [74, 75]. Second, we tested many of our variants under both additive and dominant genetic models so we applied an additional Bonferonni correction as a further conservative measure we applied an additional Bonferonni correction, which yielded alpha = 5 x 10^{-9}. We considered SNPs with joint p-value < 1*10^{-6} suggestively associated, which is conservative compared to many recent reports [76]. QQ plots are given for the Omni 1M and 1S arrays (phase I; Figure S2A) and the Omni 5M arrray (Phase II; Figure S2B) after exclusion of the MHC region. Because we expected that true positive associations would share the same direction of effect (granted the similar disease state and ethnic background) we used a fixed effects meta-analysis for this portion of the study.

2. Meta-analysis of RA GWA studies in Europeans, Asians, and African-Americans

Because there is evidence of heterogeneity of effect in many risk loci previously reported in Asians and Europeans (see Discussion) we tested variants using random effects meta-analysis implemented by Han and Eskin [27, 28]. In addition, we sought to employ a formal test of effect size in order to decide which RA risk loci from studies of other ethnic groups replicated in our study of African-Americans. To do this, we conducted a trans-ethnic meta-analysis using GWA summary statistics from African-Americans (our analysis), and summary statistics from persons of European ancestry and from Asian ancestry kindly provided by Okada and colleagues as used in their RA meta-analysis [3]. All data were combined into one dataset and analyzed using METASOFT, an open-source meta-analysis tool (http://genetics.cs.ucla.edu/meta/) [27, 28]. For each variant in this multi-ethnic dataset, METASOFT calculated: a) association p-values for RA in Europeans, East Asians and in African-Americans (p_{EUR}, p_{EAS}, and p_{AFR}, respectively) and b) m-values (posterior probabilities that a genetic variant has an effect in a given population). p and m values for all the index variants reported by Okada *et al.* are shown in Table 2. m-values were not calculated for 18 of the index variants in African-Americans because of either low minor allele frequency or exclusion based on QC metrics as described above. We categorized results into three strata as used by Han and Eskin [27, 28]. These include: a) SNPs that are predicted to have an effect (m-value > 0.8); b). SNPs that are predicted to not have an effect (m-value < 0.2); and c). ambiguous SNPs (m-value between 0.2 and 0.8) in which it is unknown whether there is an effect. Figure 3 shows a P-M plot to aid with visualization of these results. For loci in which the index variant had an m-value between 0.2 and 0.8 or had very low allele frequency, we conducted an association enrichment analysis using MAGENTA [32]. Specifically, we scanned whether the gene was enriched for associated variants. This was done in order to further assess whether the association was truly absent or merely discordant at the index variant alone.

4. Trans-ethnic fine-mapping to identify candidate pathogenic variants

Because our study has the advantage of inclusion of genetic data from several populations, we sought to leverage this to identify candidate pathogenic variants from among the many RA-associated variants in the previously reported risk loci. To accomplish this, we conducted trans-ethnic fine-mapping in 91 RA risk loci using the PAINTOR algorithm, which calculates a posterior probability that each variant in a risk locus is pathogenic. Results range from 0 indicating very unlikely to 1, indicating highly likely. This analysis included the following steps: Step 1. Alignment of reference and alternate alleles for all populations (Asian, European, and African-American) to match the designations found in the 1000 Genomes Project [22]. Step 2. Generation of LD matrices for each population based on linkage in the 1000 Genomes Project. Step 3. Quantifying the importance of each genomic annotation to RA risk for each of 8,138 genomic annotations in effect size estimates (gamma estimates) RA risk. The initial analysis was performed by running each of the 8,138 genomic annotations in the 91 RA risk loci individually. A list of these annotations may be found here: (https://github. com/gkichaev/ PAINTOR_V3.0/wiki/2b.-Overlapping-annotations) [22]. Step 4. These effect size estimates were sorted in descending order, selecting those that were most informative (most likely to improve the model fit). Step 5. We sequentially compared informative annotations, and if they were correlated, we removed the annotation with the smaller effect size estimate until only weakly correlated annotations remained (r2 < 0.10); Step 6. Trans-ethnic fine-mapping was performed using the 5 most informative uncorrelated annotations.

Within each locus, we summed sorted variants in descending order of posterior probability, until 90% of the probability mass was accounted for. This is known as a 90% credible set [77] of candidate pathogenic variants for each RA risk locus (Table S4). We compared our results to functionally validated SNPs to assess the degree of external validation. In most cases, the precise variant that gives rise to a GWAS association is unknown, however some variants are fairly well-studied. For example, in the *PTPN22* locus we obtained a posterior probability of 1.0 (almost certainly pathogenic) for rs2476601, which is generally accepted as the pathogenic variant in Europeans with RA. We next established internal consistency by benchmarking our results on those reported by Kichaev *et al.* [21], who previously analyzed data on Europeans and Asians with RA. Initially our findings differed from those previously reported. We also ran the analysis in CAVIARBF in order to further assess the stability of probability estimates. Finally, We applied rigorous quality control measures to 1) increase the number of variants we were able to include and 2) improve the accuracy of our LD matrices. These are described in Supplementary Methods. After QC, our data generally agreed strongly with the prior results (see Table 4A).

Supplementary Material – Supplementary Material is available at HMG online.

Acknowledgments - We gratefully acknowledge the CLEAR Investigators: Doyt L. Conn, MD; Beth L. Jonas, MD, Leigh F. Callahan, PhD, Edwin A. Smith, MD, Richard D. Brasington, Jr., MD, and Larry W. Moreland, MD. The assistance of Stephanie

Ledbetter, MS is greatly appreciated, as are helpful discussions with Elizabeth E. Brown, PhD.

Funding – This work was supported by National Institutes of Health grants [R01 AR057202, 5UL1 RR025777, 2P60 AR048095, and N01 AR-6-2278 to SLB, K01 AR060848 to RJR, K23 AR062100 to MID, R37 AI024717 and 1U01 HG008666 to JBH]. JBH was also supported by the US Department of Veterans Affairs.

Conflict of Interest Statement - None of the authors report conflicts of interest with regard to this work.

References

1. Silman, A.J. and J.E. Pearson, *Epidemiology and genetics of rheumatoid arthritis*. Arthritis Res., 2002. 4 Suppl 3: p. S265-S272.

2. McInnes, I.B. and G. Schett, *The pathogenesis of rheumatoid arthritis*. N Engl J Med, 2011. 365(23): p. 2205-19.

3. Okada, Y., et al., *Genetics of rheumatoid arthritis contributes to biology and drug discovery.* Nature, 2014. 506(7488): p. 376-81.

4. Viatte, S., D. Plant, and S. Raychaudhuri, *Genetics and epigenetics of rheumatoid arthritis.* Nat Rev Rheumatol, 2013. 9(3): p. 141-53.

5. Hughes, L.B., et al., *Most common single-nucleotide polymorphisms associated with rheumatoid arthritis in persons of European ancestry confer risk of rheumatoid arthritis in African Americans*. Arthritis Rheum, 2010. 62(12): p. 3547-53.

6. Stastny, P., *Association of the B-cell alloantigen DRw4 with rheumatoid arthritis.* N Engl J Med, 1978. 298(16): p. 869-71.

7. Hughes, L.B., et al., *The HLA-DRB1 shared epitope is associated with susceptibility to rheumatoid arthritis in African Americans through European genetic admixture.* Arthritis Rheum, 2008. 58(2): p. 349-58.

8. Ding, B., et al., *Different patterns of associations with anti-citrullinated protein antibody-positive and anti-citrullinated protein antibody-negative rheumatoid arthritis in the extended major histocompatibility complex region.* Arthritis Rheum, 2009. 60(1): p. 30-8.

9. Gregersen, P.K., J. Silver, and R.J. Winchester, *The shared epitope hypothesis. An approach to understanding the molecular genetics of susceptibility to rheumatoid arthritis*. Arthritis Rheum, 1987. 30(11): p. 1205-13.

10. Raychaudhuri, S., et al., *Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis*. Nat Genet, 2012. 44(3): p. 291-6.

11. Viatte, S., et al., *Association of HLA-DRB1 haplotypes with rheumatoid arthritis severity, mortality, and treatment response*. JAMA, 2015. 313(16): p. 1645-56.

12. Reynolds, R.J., et al., *HLA-DRB1-Associated Rheumatoid Arthritis Risk at Multiple Levels in African Americans: Hierarchical Classification Systems, Amino Acid Positions, and Residues*. Arthritis Rheumatol, 2014. 66(12): p. 3274-82.

13. Richard-Miceli, C. and L.A. Criswell, *Emerging patterns of genetic overlap across autoimmune disorders*. Genome Med, 2012. 4(1): p. 6.

14. Morris, A.P., Transethnic meta-analysis of genomewide association studies. Genet Epidemiol, 2011. 35(8): p. 809-22.

15. Mahajan, A., et al., *Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility*. Nat Genet, 2014. 46(3): p. 234-44.

16. Begovich, A.B., et al., *A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis.* Am.J.Hum.Genet., 2004. 75(2): p. 330-337.

17. Viatte, S., et al., *Investigation of Caucasian rheumatoid arthritis susceptibility loci in African patients with the same disease*. Arthritis Res Ther, 2012. 14(6): p. R239.

18. Govind, N., et al., *Immunochip identifies novel, and replicates known, genetic risk loci for rheumatoid arthritis in black South Africans*. Mol Med, 2014. 20: p. 341-9.

19. Laufer, V.A., et al., *Integrative Approaches to Understanding the Pathogenic Role of Genetic Variation in Rheumatic Diseases*. Rheum Dis Clin North Am, 2017. 43(3): p. 449-466.

20. Chen, W., et al., *Fine Mapping Causal Variants with an Approximate Bayesian Method Using Marginal Test Statistics*. Genetics, 2015. 200(3): p. 719-36.

21. Kichaev, G. and B. Pasaniuc, *Leveraging Functional-Annotation Data in Trans-ethnic Fine-Mapping Studies.* Am J Hum Genet, 2015. 97(2): p. 260-71.

22. Kichaev, G., et al., *Improved methods for multi-trait fine mapping of pleiotropic risk loci.* Bioinformatics, 2017. 33(2): p. 248-255.

23. Ong, R.T., et al., *Efficiency of trans-ethnic genome-wide meta-analysis and fine-mapping.* Eur J Hum Genet, 2012. 20(12): p. 1300-7.

24. Consortium, T.H.I., *A haplotype map of the human genome*. Nature, 2005. 437(7063): p. 1299-320.

25. Hinch, A.G., et al., *The landscape of recombination in African Americans.* Nature, 2011. 476(7359): p. 170-5.

26. Skol, A.D., et al., *Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies*. Nat Genet, 2006. 38(2): p. 209-13.

27. Han, B. and E. Eskin, *Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies*. Am J Hum Genet, 2011. 88(5): p. 586-98.

28. Han, B. and E. Eskin, *Interpreting meta-analyses of genome-wide association studies.* PLoS Genet, 2012. 8(3): p. e1002555.

29. Okada, Y., et al., *Significant impact of miRNA-target gene networks on genetics of human complex traits.* Sci Rep, 2016. 6: p. 22223.

30. Kang, E.Y., et al., *ForestPMPlot: A Flexible Tool for Visualizing Heterogeneity Between Studies in Meta-analysis.* G3 (Bethesda), 2016. 6(7): p. 1793-8.

31. Farh, K.K.H., et al., *Genetic and Epigenetic Fine-Mapping of Causal Autoimmune Disease Variants*. Nature, 2015. 518(7539): p. 337-43.

32. Goodarzi, M.O., et al., *Systematic evaluation of validated type 2 diabetes and glycaemic trait loci for association with insulin clearance*. Diabetologia, 2013. 56(6): p. 1282-90.

33. Harley, J.B., et al., *Transcription factors operate across disease loci, with EBNA2 implicated in autoimmunity*. Nat Genet, 2018. 50(5): p. 699-707.

34. Welter, D., et al., *The NHGRI GWAS Catalog, a curated resource of SNP-trait associations.* Nucleic Acids Res, 2014. 42(Database issue): p. D1001-6.

35. Julia, A., et al., *Genome-wide association study of rheumatoid arthritis in the Spanish population: KLF12 as a risk locus for rheumatoid arthritis susceptibility.* Arthritis Rheum, 2008. 58(8): p. 2275-86.

36. Gorenshteyn, D., et al., *Interactive Big Data Resource to Elucidate Human Immune Pathways and Diseases.* Immunity, 2015. 43(3): p. 605-14.

37. Boswell, S., et al., *Induction of CD152 (CTLA-4) and LAP (TGF-beta1) in human Foxp3- CD4+ CD25- T cells modulates TLR-4 induced TNF-alpha production*. Immunobiology, 2013. 218(3): p. 427-34.

38. Yuan, S., et al., *GPC5, a novel epigenetically silenced tumor suppressor, inhibits tumor growth by suppressing Wnt/beta-catenin signaling in lung adenocarcinoma.* Oncogene, 2016. 35(47): p. 6120-6131.

39. Lorentzen, A.R., et al., *Association to the Glypican-5 gene in multiple sclerosis*. J Neuroimmunol, 2010. 226(1-2): p. 194-7.

40. Cavanillas, M.L., et al., *Replication of top markers of a genome-wide association study in multiple sclerosis in Spain.* Genes Immun, 2011. 12(2): p. 110-5.

41. Johnson, B.A., et al., *Multiple sclerosis susceptibility alleles in African Americans.* Genes Immun, 2010. 11(4): p. 343-50.

42. Cenit, M.D., et al., Glypi*can 5 is an interferon-beta response gene: a replication study.* Mult Scler, 2009. 15(8): p. 913-7.

43. Aterido, A., et al., *Novel insights into the regulatory architecture of CD4+ T cells in rheumatoid arthritis*. PLoS One, 2014. 9(6): p. e100690.

44. Suhre, K., et al., *Connecting genetic risk to disease end points through the human blood plasma proteome*. Nat Commun, 2017. 8: p. 14357.

45. Darrah, E., et al., *Peptidylarginine deiminase 2, 3 and 4 have distinct specificities against cellular substrates: Novel insights into autoantigen selection in rheumatoid arthritis*. Ann Rheum Dis, 2012. 71(1): p. 92-8.

46. Naranbhai, V., et al., *Genomic modulators of gene expression in human neutrophils.* Nat Commun, 2015. 6: p. 7545.

47. Firestein, G. and I.B. McInnes, *Immunopathogenesis of rheumatoid arthritis.* Immunity, 2017. 46(2): p. 183-96.

48. Damgaard, D., L. Senolt, and C.H. Nielsen, *Increased levels of peptidylarginine deiminase 2 in synovial fluid from anti-CCP-positive rheumatoid arthritis patients: Association with disease activity and inflammatory markers*. Rheumatology (Oxford), 2016. 55(5): p. 918-27.

49. Bawadekar, M., et al., *Peptidylarginine deiminase 2 is required for tumor necrosis factor alpha-induced citrullination and arthritis, but not neutrophil extracellular trap formation.* J Autoimmun, 2017. 80: p. 39-47.

50. Myouzen, K., et al., *Functional variants in NFKBIE and RTKN2 involved in activation of the NF-kappaB pathway are associated with rheumatoid arthritis in Japanese*. PLoS Genet, 2012. 8(9): p. e1002949.

51. Imamura, H., et al., *Impaired NFKBIE gene function decreases cellular uptake of methotrexate by down-regulating SLC19A1 expression in a human rheumatoid arthritis cell line*. Mod Rheumatol, 2016. 26(4): p. 507-16.

52. Ueda, H., et al., *Association of the T-cell regulatory gene CTLA4 with susceptibility to autoimmune disease*. Nature, 2003. 423(6939): p. 506-11.

53. Westra, H.-J., et al., *Fine-mapping identifies causal variants for RA and T1D in DNASE1L3, SIRPG, MEG3, TNFAIP3 and CD28/CTLA4 loci*. bioRxiv, 2017.

54. de Jong, V.M., et al., *Variation in the CTLA4 3'UTR has phenotypic consequences for autoreactive T cells and associates with genetic risk for type 1 diabetes.* Genes Immun, 2016. 17(1): p. 75-8.

55. Danila, M.I., et al., *Dense Genotyping of Immune-Related Regions Identifies Loci for Rheumatoid Arthritis Risk and Damage in African Americans*. Mol Med, 2017. 23.

56. Pei, Y., EBV Nuclear Antigen 3C *Mediates Regulation of E2F6 to Inhibit E2F1 Transcription and Promote Cell Proliferation*. 2016. 12(8).

57. Tang, Q., et al., *Expression of Interferon-gamma Receptor Genes in PBMCs is Associated with Rheumatoid Arthritis and Its Radiographic Severity in African Americans.* Arthritis Rheumatol, 2015. 67(5): p. 1165-70.

58. Regis, G., et al., *IFNgammaR2 trafficking tunes IFNgamma-STAT1 signaling in T lymphocytes.* Trends Immunol, 2006. 27(2): p. 96-101.

59. Bach, E.A., M. Aguet, and R.D. Schreiber, *The IFN gamma receptor: a paradigm for cytokine receptor signaling.* Annu Rev Immunol, 1997. 15: p. 563-91.

60. Arnett, F.C., et al., *The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis*. Arthritis Rheum, 1988. 31(3): p. 315-24.

61. Karlson, E.W., et al., *A connective tissue disease screening questionnaire for population studies*. Ann Epidemiol, 1995. 5(4): p. 297-302.

62. Karlson, E.W., et al., *High sensitivity, specificity and predictive value of the Connective Tissue Disease Screening Questionnaire among urban African-American women.* Lupus, 2005. 14(10): p. 832-6.

63. Tan, W., et al., *A functional RANKL polymorphism associated with younger age at onset of rheumatoid arthritis*. Arthritis Rheum, 2010. 62(10): p. 2864-75.

64. Song, J.J., et al., *Plasma carboxypeptidase B downregulates inflammatory responses in autoimmune arthritis*. J Clin Invest, 2011. 121(9): p. 3517-27.

65. Tang, Q., et al., *Expression of Interferon-gamma Receptor Genes in Peripheral Blood Mononuclear Cells Is Associated With Rheumatoid Arthritis and Its Radiographic Severity in African American*s. Arthritis Rheumatol, 2015. 67(5): p. 1165-70.

66. Bridges, S.L., Jr., et al., *Radiographic severity of rheumatoid arthritis in African Americans: results from a multicenter observational study*. Arthritis Care Res (Hoboken), 2010. 62(5): p. 624-31.

67. Mikuls, T.R., et al., *Anti-cyclic citrullinated peptide antibody and rheumatoid factor isotypes in African Americans with early rheumatoid arthritis*. Arthritis Rheum, 2006. 54(9): p. 3057-9.

68. Mikuls TR, R.A., Kerr GS, Cannon GW, *Insights and Implications of the VA Rheumatoid Arthritis Registry.* Fed Pract., 2015. 32(5): p. 24-29.

69. Mikuls, T.R., et al., *The association of race and ethnicity with disease expression in male US veterans with rheumatoid arthritis*. J Rheumatol, 2007. 34(7): p. 1480-4.

70. Freedman, B.I., et al., *End-stage renal disease in African Americans with lupus nephritis is associated with APOL1*. Arthritis Rheumatol, 2014. 66(2): p. 390-6.

71. Price, A.L., et al., *Principal components analysis corrects for stratification in genome-wide association studies*. Nat Genet, 2006. 38(8): p. 904-9.

72. Howie, B.N., P. Donnelly, and J. Marchini, *A flexible and accurate genotype imputation method for the next generation of genome-wide association studies*. PLoS Genet, 2009. 5(6): p. e1000529.

73. Marchini, J., et al., *A new multipoint method for genome-wide association studies by imputation of genotypes*. Nat Genet, 2007. 39(7): p. 906-13.

74. Dudbridge, F. and A. Gusnanto, *Estimation of significance thresholds for genomewide association scans.* Genet Epidemiol, 2008. 32(3): p. 227-34.

75. Risch, N. and K. Merikangas, *The future of genetic studies of complex human diseases.* Science, 1996. 273(5281): p. 1516-7.

76. Wang, Z., et al., *A large-scale genome-wide association and meta-analysis identified four novel susceptibility loci for leprosy*. Nat Commun, 2016. 7: p. 13760.

77. Edwards, W.L., Harold; Savage, Leonard J, *Bayesian statistical inference for psychological research.* Psychological Review. 70(3): p. 193-242.

78. Menard, L., et al., *The PTPN22 allele encoding an R620W variant interferes with the removal of developing autoreactive B cells in humans*. J Clin Invest, 2011. 121(9): p. 3635-44.

79. Lopez de Lapuente, A., et al., *Novel Insights into the Multiple Sclerosis Risk Gene ANKRD55*. J Immunol, 2016. 196(11): p. 4553-65.

80. Couturier, N., et al., *Tyrosine kinase 2 variant influences T lymphocyte polarization and multiple sclerosis susceptibility*. Brain, 2011. 134(Pt 3): p. 693-703.

81. Cavalli, M., et al., *Allele-specific transcription factor binding to common and rare variants associated with disease and gene expression*. Hum Genet, 2016. 135(5): p. 485-97.

82. Jansen, R., et al., *Conditional eQTL analysis reveals allelic heterogeneity of gene expression.* Hum Mol Genet, 2017. 26(8): p. 1444-1451.

83. Corradin, O., et al., *Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits.* Genome Res, 2014. 24(1): p. 1-13.

84. Bokor, S., et al., *Single nucleotide polymorphisms in the FADS gene cluster are associated with delta-5 and delta-6 desaturase activities estimated by serum fatty acid ratios*. J Lipid Res, 2010. 51(8): p. 2325-33.

85. Chen, W.C., et al., rs657075 (CSF2) *Is Associated with the Disease Phenotype (BAS-G) of Ankylosing Spondylitis.* Int J Mol Sci, 2017. 18(1).

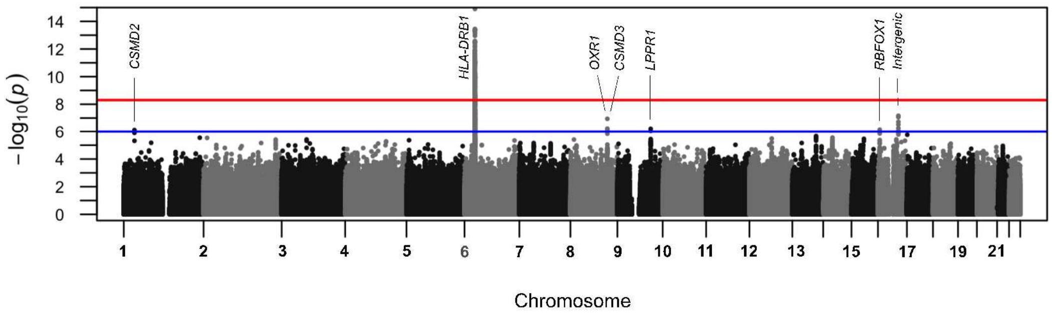


Figure S1 – Manhattan plot of the association summary statistics computed for African-Americans marginally as a part of Phase I. p-values from association testing of our phase II data (African-American GWAS) regression under an additive genetic model. The x-axis indicates chromosome and position, the y-axis indicates $–\log_{10}(p)$. The horizontal red line is drawn at $5*10^{-9}$ and is the threshold for genome wide significance, the blue line is for suggestive evidence of association and is drawn at $1*10^{-6}$. The only association detected was that of the HLA region, but several suggestive associations were detected.

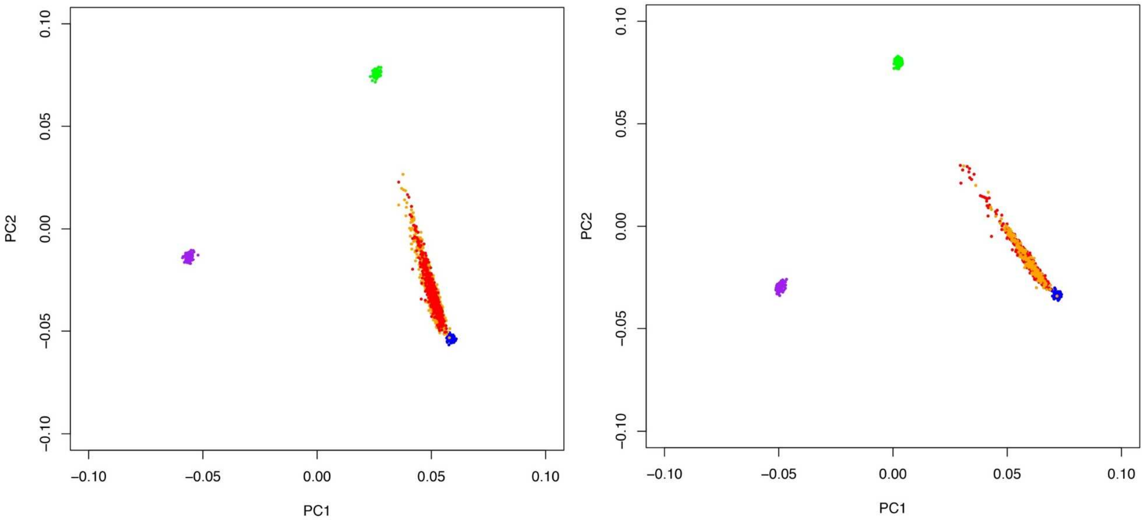


Figure S2 – Quantile-Quantile plot of our association results after removal of the MHC. Scatter plots of principal component loadings. Most of the variation in the study data (red and orange) to lie between ancestral European (green dots) and Yoruban (blue dots) populations. The purple dots signify Asian samples. A – Principal component loadings for the discovery cohort (Illumina Omni 1M and 1S chips). B – Principal component loadings for the replication cohort genotyped on the Illumina Omni 5M chip. C – Principal component loadings for the replication cohort genotyped on the MEGA chip.

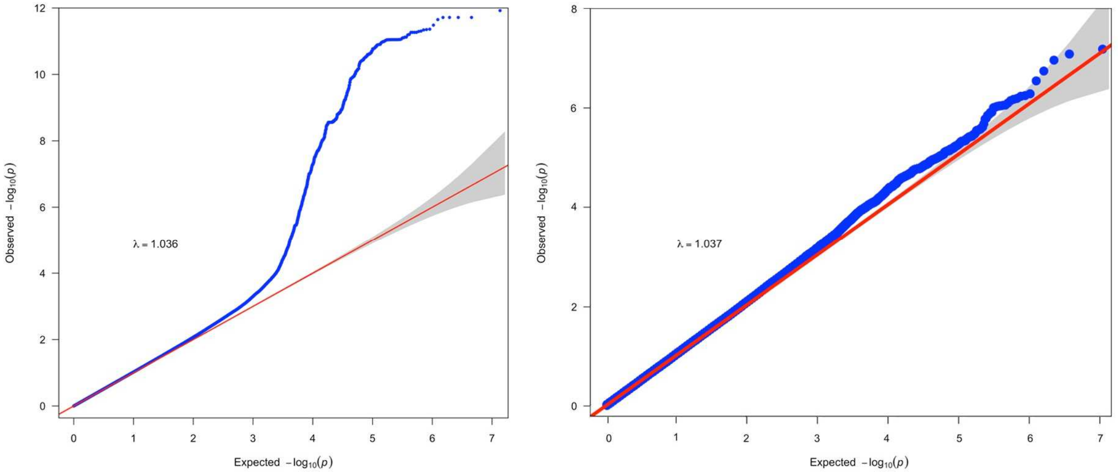


Figure S3 – Zoomplots of the associations of the *GPC5, RBFOX1*, and *CSMD3* with RA in African-Americans. QQ plots after removal of the extended HLA region (chr6:26,000,000-34,000,000). The genomic inflation for each of the three chips in A-C is given by lambda. A – QQ plot for the discovery cohort (Illumina Omni 1M and 1S chips). In figure A, the variant at the top right did not have support from linkage disequilibrium and was not included in the analysis. B – QQ plot for the replication cohort genotyped on the Illumina Omni 5M chip. No excessive genomic inflation of association summary statistics was found ($\lambda_{GC} < 1.05$).

Table 1. Baseline Characteristics of CLEAR Registry

	CLEAR I RA (n=265)	CLEAR I Controls (n=80)	CLEAR II RA (n=597)	CLEAR II Controls (n=194)
Age at enrollment, years, mean (SD)	51.0 (12.7)	54.5 (13.1)	56.0 (11.2)	57.7 (7.61)
Age at RA onset, mean (SD)	49.2 (12.6)	-	45.4 (11.7)	-
Gender (female), %	83.0	72.5	87.2	71.2
Disease duration at enrollment, months, median (IQ 25-75)	12.1 (6.3 - 18.2)	-	101 (36.5-196.0)	-
Family history of RA§ %	30.9	-	38.7	-
Smoking (ever %)	53.3	43.4	51.3	54.4
HAQ Score, median (IQ 25-75)	1.38 (0.70-1.95)	-	1.3 (0.75-1.82)	-
JAM Score, median (IQ 25-75)	4.00 (0.00-12.75)	-	2.30 (0.00-12.3)	-
Number of tender joints* (of the 28 in the DAS28), (IQ 25-75)	6.0 (1.0-15.0)	-	3.8 (1.0-9.0)	-
Number of swollen joints* (of the 28 in the DAS28), (IQ 25-75)	3.0 (1.0-7.0)	-	3.4 (1.0-9.0)	-
Rheumatoid factor, % positive	94.8	15.2	82.7	19.8
Anti-CCP antibody, % positive	97.9	3.8	74.3	2.8
Medications				
DMARDs, ever used %	85.6	-	94.8	-
Methotrexate, ever used %	78.2	-	93.9	-
Biologics, ever used %	4.9	-	7.2	-

African-American Patients with ACPA-positive Rheumatoid Arthritis and Controls Analyzed from the CLEAR registry. HAQ – Health Assessment Questionnaire. IQ 25-75 – interquartile range. JAM – Joint Alignment and Motion Score. *Based on the 28 joints used in the calculation of the DAS28. The CLEAR Registry was started prior to the common use of DAS28 to quantify disease activity.

Table 2. Novel associations with RA reaching genome-wide significance in one or more global population.

SNP ID	Chr	Position	Candidate Gene	A1	A2	P_{EUR}	P_{EAS}	P_{AA}	P_{TE}	M_{EUR}	M_{EAS}	M_{AA}
rs761426	1	17413899	*PADI2*	A	T	1.23E-04	1.26E-06	2.56E-02	2.48E-10*	0.999	1	0.96
rs2203098	8	115012597	*CSMD3*	G	C	6.87E-02	2.85E-01	6.54E-10*	5.47E-08	0	0	1
rs9516053	13	92945884	*GPC5*	C	T	3.24E-01	8.06E-01	3.09E-09*	3.28E-06	0	0	1
rs4602043	16	5588689	*RBFOX1*	C	T	3.31E-01	9.12E-02	4.07E-09*	5.03E-07	0	0.009	1

Abbreviations: Chr – chromosome. Pos – base pair location. P_{EUR}, P_{EAS}, P_{AA} the p-value for the variant calculated using only data from Europeans, East Asians, and African-Americans, respectively. These p-values are the association p-value for the variant in the overall trans-ethnic meta-analysis as calculated by METASOFT using Han and Eskins's Random Effects model, which is optimized to detect potentially heterogenous associations statistic. M_{EUR}, M_{EAS}, and M_{AA} - the m-value for the variant calculated using only data from Europeans, East-Asians, and African-Americans, respectively.

Table 3. Replication of genetic variants previously associated with RA in persons of European / Asian ethnicity.

SNP ID	Chr	Position	Candidate Gene	A1	A2	P_{EUR}	P_{EAS}	P_{AA}	P_{TE}	M_{EUR}	M_{EAS}	M_{AA}
rs9268839	6	32428772	*HLA-DRB1*	A	G	<1E-250	3.7E-134	2.7E-07	<1E-250	1.00	1.00	1.00
rs3087243	2	204738919	*CTLA4*	A	G	9.0E-20	2.2E-04	2.1E-03	9.9E-24	1.00	1.00	0.99
rs11889341	2	191943742	*STAT4*	T	C	6.4E-12	6.3E-09	4.8E-03	5.3E-20	1.00	1.00	0.98
rs2736337	8	11341880	*BLK*	T	C	1.6E-07	2.0E-06	2.6E-03	4.2E-13	1.00	1.00	0.98
rs9653442	2	100825367	*AFF3*	T	C	3.5E-12	4.6E-04	1.4E-02	1.7E-15	1.00	1.00	0.96
rs9378815	6	426155	*IRF4*	G	C	1.6E-07	6.0E-05	1.6E-02	1.0E-11	1.00	1.00	0.96
rs909685	22	39747671	*SYNGR1*	A	T	3.1E-10	3.8E-06	3.3E-02	3.7E-14	1.00	1.00	0.95
rs59716545	17	38031857	*IKZF3-CSF3*	T	G	2.0E-09	9.5E-05	1.4E-02	2.6E-13	1.00	1.00	0.95
rs2233424	6	44233921	*NFKBIE*	T	C	3.3E-08	9.3E-13	4.9E-02	1.6E-19	1.00	1.00	0.93
rs2105325	1	173349725	*LOC100506023*	A	C	1.0E-08	7.6E-03	3.6E-02	9.9E-11	1.00	0.99	0.93
rs2451258	6	159506600	*TAGAP*	T	C	6.4E-10	1.1E-01	2.3E-02	6.2E-11	1.00	0.90	0.92
rs73013527	11	128496952	*ETS1*	T	C	2.0E-06	1.2E-06	6.5E-02	3.1E-11	1.00	1.00	0.91
rs1980422	2	204610396	*CD28*	T	C	6.1E-11	3.4E-02	1.1E-01	3.0E-12	1.00	0.96	0.90
rs9603616	13	40368069	*COG6*	T	C	8.3E-11	1.0E-02	1.1E-01	2.3E-12	1.00	0.98	0.90
rs2561477	5	102608924	*C5orf30*	A	G	5.3E-10	2.1E-01	8.9E-03	5.5E-10	1.00	0.34	0.89
rs2317230	1	157674997	*FCRL3*	T	G	1.0E-05	3.1E-04	1.5E-01	1.1E-08	1.00	1.00	0.86
rs10774624	12	111833788	*SH2B3-PTPN11*	A	G	2.4E-07	-	2.1E-02	8.2E-08	1.00	-	0.86
rs8032939	15	38834033	*RASGRP1*	T	C	2.4E-12	3.1E-05	2.2E-01	3.5E-16	1.00	1.00	0.86
rs1877030	17	37740161	*MED1*	T	C	1.5E-05	3.0E-04	2.2E-01	1.5E-08	1.00	1.00	0.86
rs10175798	2	30449594	*LBH*	A	G	1.4E-07	9.8E-03	2.1E-01	3.4E-09	1.00	0.98	0.86
rs2664035	4	48220839	*TEC*	A	G	2.5E-06	6.0E-01	3.7E-02	4.1E-06	1.00	0.28	0.86
rs2236668	21	45650009	*ICOSLG-AIRE*	T	C	1.2E-05	2.6E-03	2.7E-01	8.0E-08	1.00	0.99	0.85
rs6732565	2	111607832	*ACOXL*	A	G	8.8E-05	8.7E-03	2.6E-01	1.8E-06	1.00	0.98	0.84
rs9372120	6	106667535	*ATG5*	T	G	1.2E-05	1.2E-03	2.3E-01	1.4E-07	1.00	0.99	0.84
rs8133843	21	36738242	*RUNX1-LOC100506403*	A	G	6.0E-09	3.3E-02	1.8E-01	9.7E-10	1.00	0.90	0.84
rs706778	10	6098949	*IL2RA*	T	C	7.1E-12	2.9E-01	7.5E-02	2.1E-11	1.00	0.09	0.83
rs11605042	11	72411664	*ARAP1*	A	G	1.4E-02	9.7E-04	2.8E-01	5.3E-05	0.94	0.99	0.82
rs6479800	10	64036881	*RTKN2*	G	C	2.2E-03	6.9E-07	2.5E-01	8.1E-08	0.78	1.00	0.82
rs17264332	6	138005515	*TNFAIP3*	A	G	6.9E-19	-	3.5E-01	1.3E-17	1.00	-	0.19
rs11933540	4	26120001	*C4orf52*	T	C	9.5E-17	-	6.4E-01	2.2E-15	1.00	-	0.15

rs10790268	11	118729391	*CXCR5*	A	G	3.3E-15	3.2E-01	4.6E-01	1.2E-13	1.00	0.06	0.07
rs8026898	15	69991417	*LOC145837*	A	G	2.4E-17	6.8E-03	4.5E-02	2.5E-17	1.00	0.97	0.01

Chr – chromosome. Pos – base pair location. P_{EUR} – the p-value for the variant calculated using only data from Europeans. P_{EAS} - the p-value for the variant calculated using only data from Asians. P_{AA} the p-value for the variant calculated using only data from African-Americans. These p-values are the association p-value for the variant in the overall trans-ethnic meta-analysis as calculated by METASOFT using Han and Eskins's Random Effects model, which is optimized to detect potentially heterogenous associations statistic. M_{EAS} - the m-value for the variant calculated using only data from Europeans. M_{EUR} - the p-value for the variant calculated using only data from East Asians. M_{AA} - the p-value for the variant calculated using only data from African-Americans.

SNP Information			Z-statistics from GWAS of RA susceptibility in 3 global populations			Unannotated and (Annotated) Posterior Probabilities from Trans-ethnic Fine Mapping		Annotations and citations regarding variant
rsID	Gene	Effect / Alternate Allele	EUR [3]	EAS [3]	AA	European, Asian RA (ref 3; 26)	African-American, Asian & European RA	Description; Reference Numbers
rs2476601	*PTPN22*	G / A	-26.04	-	-	1.00 (1.00)	1.00 (1.00)	R620W [78]
rs7731626	*ANKRD55*	A / G	-9.83	-	-0.3	1.00 (1.00)	1.00 (1.00)	ANKRD55 eQTL [79]
rs147622113	*ILF3*	T / C	-6.13	-	-0.47	1.00 (1.00)	1.00 (1.00)	
rs34536443 & rs74956615	*TYK2*	C / G	-8.12	-	-	-	0.99 (0.99)	P1104A variant – Loss of kinase activity [80]
		A / T	-8.16	-	-			
rs909685	*SYNGR1*	A / T	6.29	4.62	2.13	0.65 (0.84)	0.94 (1.00)	Disrupts PITX3 TFBS [81]
rs1893592	*UBASH3A*	C / A	-5.73	-4.01	-0.44	1.00 (1.00)	1.00 (1.00)	UBASH3A eQTL; [82]
rs72634030	*C1QBP*	A / C	3.5	4.14	0.29	-	0.99 (0.99)	
rs67574266	*REL*	A / G	7.48	-0.32	1.85	-	0.21 (0.96)	Canonical CTCF binding site in *REL* 5'UTR
rs706778	*IL2RA*	T / C	6.86	1.06	2.4	-	0.94 (0.94)	Linked to ASE of IL2RA; PFKFB3 expression [83]
rs10774624	*SH2B3/ PTPN11*	A / G	-5.17	-	-1.38	-	0.96 (0.94)	
rs2233424 & rs2233434	*NFKBIE*	A / G	5.52	7.14	1.97	-	0.51 (0.51)	Val194Ala [50]. Alters MTX uptake [51].
		G / A	5.52	7.13	1.97		0.48 (0.48)	
rs12715125	*EOMES*	G / C	-5.58	-	-0.27	0.95 (0.99)	0.79 (0.83)	
rs3087243	*CTLA4*	A / G	-9.1	-3.7	-2.96	-	0.76 (0.80)	CTLA4 CT60G [54]
rs13330176	*IRF8*	A / T	5.75	3.61	2.06	-	0.79 (0.80)	
rs968567	*FADS2*	T / C	-4.95	-	-	0.29 (0.85)	0.25 (0.65)	AS-ELK1 binding to FADS2 promoter [84]

rs657075	*IL3 / CSF2*	A / G	2.64	4.45	-	0.73 (0.82)	0.52 (0.52)	ASCL6 eQTL [85]
rs71508903	*ARID5B*	T / C	7.26	5.88	-	0.76 (0.93)	0.51 (0.51)	
rs187339910	*MMEL*	A / G	−5.22	−4.18	-	1.00 (1.00)	0.01 (0.01)	
rs72767222	*ANK55RD*	A / C	5.11	-	-	0.99 (0.99)	0.00* (0.00*)	
rs12693993	*CD28*	G / A	−2.74	−1.76	-0.96	0.68 (0.88)	0.00* (0.00*)	

Posterior probability estimates are given in the column “AA, EAS & EUR RA,” those generated by Kichaev *et al.* are in the “EUR and EAS RA” column. Estimates are largely concordant with those of Kichaev *et al.,* with the exception of rs12693993 and rs72767222. Abbreviations – Chr: Chromsome. Pos – Position. A1 – The Effect Allele. A2 – The alternate allele. Z_{EUR}, Z_{EAS}, and Z_{AA} – Z-statistics from the analysis of European, Asian, and African-American populations. P_{POST} – Posterior probability. * indicates it is likely the number of causal variants being modelled differed between studies.

Abbreviations

A1 – Allele 1. In fine-mapping studies, this is the effect allele.

A2 – Allele 2. In fine-mapping studies, this is the alternate allele.

AA – African-American

ACPA – Anti-citrullinated peptide/protein antibodies

AFR – 1000 Genomes super population code for African

BP – Base pair location. All coordinates given according to hg39.

CI – Confidence Interval

CHB – 1000 Genomes population code for Han Chinese in Beijing, China

JPT – 1000 Genomes population code for Japanese in Tokyo, Japan

CLEAR – Consortium for Longitudinal Evaluation of Early Arthritis Registry

EAF – Expected allele frequency

EAS – 1000 genomes population code for East Asian

EUR – 1000 genomes super population code for European

HLA – Human leukocyte antigen

LD – Linkage disequilibrium

MAA – M-Value in the African-American population from phase II of the study

MAF – Minor allele frequency

MEAS – M-Value in the East Asian population from phase II of the study

MEUR – M-Value in the European population from phase II of the study

MHC – Major histocompatibility complex

MTX – Methotrexate

OR – Odds ratio

ORL – The lower bound of the confidence interval on an odds ratio

ORU – The upper bound of the confidence interval on an odds ratio

PAA – P-Value in the African-American population from phase II of the study

MAF – Minor allele frequency

PEAS – P-Value in the East Asian population from phase II of the study

PEUR – P-Value in the European population from phase II of the study

PTE – Trans-ethnic P-Value from phase II of the study

RA – Rheumatoid arthritis

RAF – Risk Allele Frequency

RF – Rheumatoid Factor

SE – Standard Error

SNP – Single nucleotide polymorphism

T1D – Type 1 diabetes

T2D – Type 2 diabetes

TEMA – Trans-ethnic meta-analysis

VARA – Veterans Affairs Rheumatoid Arthritis registry

YRI – 1000 Genomes population code for Yoruban

DISCUSSION

Summary of Key Results

Overview of goals of the present research

Our studies had 3 chief aims. First, we aimed to discover novel associations with RA that have not been found before in other ethnicities. Second, we endeavored to validate in African-Americans known associations identified in genetic studies of RA in Asians and Europeans with RA. Last, we employed trans-ethnic fine-mapping algorithms to isolate candidate causal variants in the loci we identified.

Novel Associations in Rheumatoid Arthritis in African Americans

We found 3 loci (*CSMD3*, *GPC5*, and *RBFOX1*) that have not been previously associated with RA susceptibility in other ethnicities and that appear to be specific to African-American populations. *CSMD3*, or CUB and Sushi Multiple Domains 3 is a large, 73-exon gene stretching ~1.2Mb across 8p23, was associated with RA (rs2203098 G allele; p=6.54 x 10^{-10}). Interestingly, variants near its homolog *CSMD2* were also suggestively associated with RA (rs55798295, p = 2.84 x 10^{-7}). This family of molecules (*CSMD1, CSMD2*, and *CSMD3)* appear to be involved in complement-mediated synapse pruning in the CNS, and *CSMD3* encodes an oligomeric type I transmembrane protein that influences dendrite development [40]. Though the contributions of CSMD genes to neural phenotypes (schizophrenia and autism) is best characterized [40], these proteins are so large and complex that it is scarcely surprising they have been identified in GWAS of numerous different conditions, including immune phenotypes. For example, *CSMD3* is associated with infection and asthma, but this is by no means exhaustive, and *CSMD1* and *CSMD2* have similar profiles [41]. *CSMD3* might contribute to RA by decreasing inhibition of complement activation, but the remarkable pleiotropy of these genes suggests multiple other explanations. Experimental characterization of the protein-protein interactions of *CSMD2* and *CSMD3* might help narrow the range of hypotheses, and genetic fine-mapping studies might also help to pinpoint the right hypotheses to test.

The second novel association we identified is that of *GPC5*, which encodes glypican-5. As described in the introduction, glypicans are components of proteoglycans that are involved in cell signaling, including in the rheumatoid joint. However, the nature of these influences in consistent with a role in RA. Several glypicans are known to influence the behavior of the extracellular matrix during development and cellular proliferation, and as a result some of them (e.g. glypican-3) are well-studied oncogenes [42]. We studied the literature relating to *GPC3* in order to determine if it might help us to understand the association of *GPC5*, this is summarized in the discussion in section IV. Despite the wealth of information on *GPC3*, *GPC5*, and other glypicans, it is unclear how genetic variation in *GPC5* might affect RA risk. However, we offer two tentative rationales: 1) A trans-eQTL in the *GPC5* locus appears to downregulate Proliferating Cell Nuclear Antigen (PCNA) associated factor, which may result in increased proliferation of CD4$^+$ T-cells in the RA synovium, or in synovial fibroblasts [43]. 2) *PRKCQ* was nominally associated with RA in African-Americans, and has been associated with RA susceptibility in other populations. We note that Glypican-5 is known to modulate blood protein levels through interaction with *PRKCQ* [44], another RA risk gene that is also highly expressed on T-lymphocytes. In either case, generally we speculate that variants in the *GPC5* locus may alter CD4$^+$ T-cell behavior, contributing to dysregulated ECM growth in the RA synovium.

It was comparatively difficult to locate evidence regarding the involvement for the last locus we identified, *RBFOX1*, in RA susceptibility. *RBFOX1* is a regulator of alternative splicing of mRNA, and as above with the *CSMD* proteins, it has been extensively implicated in neural phenotypes [41], but there is also evidence that it influences many immune processes. In particular *RBFOX1* appears to influence TCR and BCR receptor signaling, leukocyte migration and differentiation [45], but this evidence comes from databases constructed from large datasets, and specific experiments defining the scope of *RBFOX1*'s contribution to immune regulation are lacking.

Contributions to understanding the genetics of RA in global populations

Although our data was on African-Americans with RA, these studies offer a number of insights into the genetics of RA in all global populations. These fall into several types. First, we asked whether the addition of African-American data would enable identification of any additional risk loci not previously named by Okada *et al.* in their meta-analysis of European and Asian data [24]. We found that *PADI2* is a genetic risk factor for RA independent of any risk conferred by variants in *PADI4*. Our approach showed that both *PADI2* and *PADI4* are risk factors in African-Americans with RA, and corroborated this finding in Europeans and Asians. Also, we found essentially no linkage between associated variants in *PADI2* and *PADI4* in Europeans, Asians, or African-Americans (maximum pairwise $r < 0.15$). Genetic variation in *PADI2* and *PADI4* is of substantial interest to the RA community for both diagnostic and therapeutic reasons. This interest stems mostly from the relationship of PAD enzymes to pathogenic autoantibody production. However, this is addressed in the Results and Discussion sections of Section IV, so further discussion is omitted here.

We also studied the question of whether risk loci in African Americans coincide with those already found in European and Asian populations systematically. Because so many genetic risk factors for RA are weak effect (OR < 1.15), we chose an approach that quantifies the likelihood that a known risk variant is causal in African-Americans. Doing this is crucial in this type of context, otherwise the observer is unable to distinguish a genuine lack of association from a type II error. To do this, as described in the Introduction and in Methods in Section IV, we chose the meta-analytic framework of Han and Eskin, who describe an M-value, which is similar to a posterior probability that a given variant has a true association in a new study [34, 35] . We calculated M-values for the 101 index variants previously identified by Okada *et al*. Using this framework, for 28 variants there was evidence of effect in African-Americans (M > 0.8), for 4 variants there was evidence that an effect does not exist (M < 0.2), and for 51 variants evidence was not strong enough to claim either. For an additional 18 loci we were unable to assign an M-value due to low allele frequency. Thus, overall there were many more risk loci that had evidence of common effect across populations than the opposite, which is consistent with prior reports in RA and other autoimmune diseases. However, we noticed several trends among the variants that do not replicate and those for which no M-value could be calculated due to allele frequency. These trends and their implications are briefly treated in the Discussion of Section IV, and they are discussed in detail below.

Trans-Ethnic Fine-Mapping of the IFNGR2 Promoter

Interestingly, one of the loci that did not replicate in either Asians or African-Americans with RA is a locus we have previously studied, *IFNGR2* (M_{EUR} = 1.00; M_{EAS} = 0.023; M_{AFR} not calculated because MAF of rs73194058 < 0.05 in African-Americans). Interestingly, the variant is common in East Asians, but apparently does not confer risk of RA susceptibility (MAF = 0.48; p = 0.39) despite being present. To try to better understand risk in this locus, we conducted trans-ethnic fine-mapping of the *IFNGR2* locus using all available data (from Europeans, Asians, and African-Americans). We identified 2 candidate pathogenic variants rs9974603 (p_{POST} = 0.59) and rs8126756 (p_{POST} = 0.03) that together comprise a majority of the total posterior probability mass in the locus. rs9974603 lies about 25 base pairs from the conserved TFBS of E2F6, and rs8126756 lies fewer than 10bp from the conserved TFBS of *ZBTB7A*, directly within the 5'UTR of *IFNGR2*.

Consideration of the particular transcription factors implicated strongly suggests a rationale for the involvement of these candidate pathogenic variants. A recent study demonstrated that many candidate variants for RA pathobiology lay within loci that can be occupied by the Epstein-Barr EBNA2 and EBNA3C proteins, in regions that tend to cluster with certain human TFs [46]. This study showed that a large number of candidate "pathogenic" variants may alter gene expression either through or due to EBNA2 and EBNA3C, for example through allele-dependent binding events of these proteins [47]. It is possible that our findings in the *IFNGR2* promoter may represent a finding of this kind. EBNA3C blocks the transcriptional activity of E2F1, thereby preventing E2F1 mediated apoptosis [47]. By contrast, E2F6 serves as a dominant negative repressor of such E2F-targeted pathways, often by competing with other E2F family proteins [47]. EBNA3C appears to enhance E2F6 expression, and also to complex with it, stabilizing E2F6 at its carboxy terminal domain. In doing this, EBNA3C shunts latently infected B cells away from apoptosis and towards cellular proliferation. Thus, E2F6 plays a crucial role in EBNA3C-mediated cell proliferation generally and malignancy specifically. Likewise, EBNA3C can drive overexpression of *ZBTB7A*, which not only leads to aggressive lymphomas but plays a key role in the instruction of early lymphoid progenitors to develop into B lineage by repressing T-cell instructive Notch signaling [47]. Specifically, ZBTB7A abrogates E2F1-dependent *CDKN2A* repression. *CDKN2A* encodes another RA risk gene (CDK2).

It is striking that the only 2 variants accounting for most of the posterior probability mass in our fine-mapping experiment - rs9974603 and rs8126756 – both lie fewer than 25 base pairs from the consensus TFBS of targets of EBNA3C. rs8126756 is also a cis-eQTL for *IFNGR2* in whole blood. In this context, our recent finding that *IFNGR2* expression positively correlates with RA radiographic progression could have several explanations and occur in several different cell types [48]. For instance, because most *IFNGR2* mRNA in whole blood is thought to be NK-cell derived, these findings could possibly indicate that in the context of the active, inflamed RA synovium the increase in *IFNGR2* expression reflects B-cell derived *IFNGR2* overexpression.

As we noted in the Discussion of Section IV, because *EBNA3C* accentuates the effects of E2F6 and *ZBTB7A*, one possible explanation for these findings is that the C allele of rs9974603 and the T allele of rs8126756 serve to increase binding of these transcription factors to the IFNGR2 promoter, thereby increasing *IFNGR2* expression in RA in the manner our lab has reported previously [48].

While it is by no means certain that EBNA proteins relate to the findings in the IFNGR2 promoter, our fine-mapping of this locus strongly suggest that autoimmune risk variants in this locus localize to the 5'UTR of *IFNGR2* and the surrounding region and predispose to RA by increasing expression of *IFNGR2.* Prior studies of RA indicate this expression increase is most likely to occur in T cells, B cells, or macrophages [48]. We suggest that assays capable of demonstrating these variants do in fact result in allele-specific binding could be a helpful place to start. However, we caution that *IFNGR2* expression differences in RA versus healthy controls might not show differences unless EBV infection status can also be ascertained in European patients.

RA Susceptibility versus RA radiographic severity: HLA-DRB1 and CXCR5

While *IFNGR1* expression is associated with RA susceptibility, *IFNGR2* expression is associated with radiographic severity of RA [48]. In Section III we presented at similar findings for several other RA risk loci, including *HLA-DRB1* and *CXCR5*. We return to these findings for two reasons. First, as stated in the introduction, we note that risk factors for closely related conditions may differ substantially. Second, we wish to address the conflicting reports as to whether variants in the HLA region confer risk of radiographic severity of RA or if this association only reflects differences in seropositivity among groups. We note that Viatte *et al.* found that amino acid positions 11, 71, and 74 of HLA-DRB1 are associated with radiographic damage. However, their study included autoantibody positive, autoantibody negative, and inflammatory polyarthritis [49]. As such, it is difficult to be completely certain the signal they are detecting corresponds to radiographic severity as a phenotype or rather to a confounded phenotype such as seropositivity. To address this issue more completely, our study of RA radiographic severity included only RA patients that were seropositive, and stratified them by count and severity of involvement of affected joints (see the Introduction and Methods in Section III). Using this approach, although we detected a strong association of the HLA region to RA susceptibility in African-Americans having a similar magnitude and direction of effect as reported in other ethnicities, there was no association of radiographic damage with any variant in the HLA region. Turning our attention to non-HLA loci, we also report that the locus containing *CXCR5* is unlikely to be associated with RA susceptibility in either East Asians or African-Americans ($M_{EAS} = 0.059$; $M_{AFR} = 0.072$; see Section IV). Nevertheless, we find a nominal association of *CXCR5* in our study of the radiographic severity of RA (section III). Thus, while we caution that it is difficult to completely stratify these related RA phenotypes, study of loci associated with susceptibility to, but not severity of, RA may be very informative to researchers attempting to tease apart factors leading to the instantiation of RA in contrast to its perpetuation and progression, which could contribute to tailored treatment regimens.

Principles of Population Genetics and Trans-ethnic Concordance of RA genetic liability

Above we indicated that we found evidence that 28 of the RA risk loci are similar in African Americans (i.e., we found 28 index variants having $M_{AFR} > 0.8$), but found evidence that risk differs for only 4 ($M_{AFR} < 0.2$). However, we also noted that an additional 18 variants were not assigned an MAFR value due to low allele frequency. In examining these variants, we discovered several trends among them, including several that may have implications for precision medicine in RA.

It is a well-known principle in population genetics that deleterious genetic variants tend to remain at lower allele frequencies [50]. In addition, genetic variants that confer a selective advantage in some circumstances but that decrease fitness in other conditions tend to become fixed at a level that balances these competing forces. Perhaps the best-known and archetypal example of this is the well-known heterozygote advantage for individuals possessing one copy of the sickle cell trait. In regions where falciparum malaria is endemic, this allele is found at a certain frequency, but this allele is essentially absent from populations that do not reside in areas where this pathogen is found. However, examples of this are definitely not limited to Mendelian disease. End-stage renal disease systemic lupus erythematosus (SLE ESRD) is a less famous, but not less compelling, example; certain genotypes bearing coding variants conger protection against *T. b. gambiense* but also predispose to ESRD, and are strongly associated with SLE ESRD [51]. However, the genetic variants that give rise to these *APOL1* haplotypes do not exist outside of regions where *T. gambiense* is exerting a selective pressure.

When we examined the loci that were found in only one or two populations, but not a third, we discovered that such variants were far more likely to have a moderate or large effect size (here, defined as OR > 1.25 or OR < 0.8). They are also much likely to have a lower minor allele frequency: "large-effect" loci had a mean MAF of 0.15, while the remaining loci had mean MAF around 0.30. In addition, the "large-effect" risk loci the former are much more likely to be exonic, while the latter are much more likely to display the autoimmune enhancer phenotype described by Corradin *et al.* (2014) in the multiple enhancer variant hypothesis. For instance, *PTPN22*, *ILF3*, and *TYK2* are the 3 strongest effect non-HLA risk loci, and all are coding variants found in European populations that are very rare or essentially absent from East Asian and African populations.

Granted prior descriptions of the relationship between variant frequency, variant pathogenicity, coding variation, and disease risk known from the field of population genetics, it seems likely that these variant profiles reflect different selective pressures and ultimately different evolutionary forces. Thus, while these observations are not new, there are a number of practical reasons to reflect on these differences in the context of the present studies. First, the presence of large effect loci that are enriched for population specificity (see Figure 3 of Section IV) represents a strong rationale to create multi-ethnic genotyping cohorts to study autoimmune diseases in the future, as it seems that there are frequently large effect loci that are specific to only one or a few populations. For effect sizes in the neighborhood of 1.25 at allele frequency around 0.15 (the values for RA we found empirically) this corresponds to low thousands of samples. This is far, far smaller than the largest GWAS in Europeans in many autoimmune diseases [24]. This means that if we can expect a few loci of large effect per population, then from a cost-efficacy perspective it is far more efficient to design several smaller studies to capture loci of moderate to large effect than it is to attempt to detect loci of weak effect in a well-studied population.

More importantly, a key implication of the claim that allele frequency differs greatly among the largest effect loci in RA is that the risk attributable to those variants also varies across populations [52] (unless the effect size also differs between populations as well, which is usually not observed). As a simple example, consider the difference in population attributable risk that owes to the difference of a single causal variant between two populations. Following Moonesinghe *et al.,* we note that as the difference in risk allele frequency *increases* between two populations, the difference in incidence of that disease phenotype will also tend to increase [53]. In our study, we noted that only 4 loci displayed positive evidence of effect size differences, while 18 were not tested due to very low allele frequency in the untested population. Whatever the above logic may indicate about the biology of a given locus, it is important to note that at the level of population attributable risk, there may in practice be little distinction between loci that do not replicate despite similar allele frequency and those that do not replicate and have differing allele frequency of the "causal" variants. Thus, although 28 variants displayed positive evidence of effect replication and only 4 displayed positive evidence against the same, another way to regard the evidence is that we were able to provide evidence that 28 index variants replicate but 22 differ. However, it is important to be clear that our assessment of that data in a general sense agrees with previous reports that the genetic basis of RA is largely shared across populations (see Section IV, Results).

RA risk variants that differ between populations likely to impact precision medicine in RA

Because the genetic variants with the larger effect sizes for RA tend to be less likely to replicate in other populations, these differences – though a minority of the total number of risk loci, may exert a strong effect on how to deliver precision medicine to patients in the future. The rationale for this claim is drawn from several sources. First, it has been estimated that drug targets that are supported by genetic data from GWAS studies are about twice as likely to gain approval as drugs that do not have direct genetic support for their proposed mechanism of action [54]. This same study observed that genes implicated in complex disease that also produce Mendelian-inherited pathological conditions have the strongest enrichment for successful drug mechanisms (OR = 7.2). We add to these observations that these variants share the characteristics identified among moderate and strong RA risk loci more closely than do weak-effect RA loci (low frequency, discordant among populations, more likely to be coding, higher OR, etc.). Put another way, the greater the contribution genetics plays to disease risk (i.e. the higher the heritability), the more likely successful drugs that target that disease are to be found, and at a locus level, the larger the percentage heritability explained by a given risk variant or risk locus, the more likely that gene is to have been targeted successfully by a therapeutic agent [54]. Because large-effect coding variants are currently better understood than variants within autoimmune gene enhancer and regulatory regions, we would expect this to lead to a disparity in the number of successfully targeted RA risk genes between the moderate to large vs low-risk categories. As an anecdotal example, we note that the candidate variant in the *TYK2* locus is a large-effect coding variant that leads to a loss of function of the *TYK2* kinase. Most would agree this is comparatively easier to identify and conceptualize than the coordinate effect of multiple enhancer variants coordinately altering gene regulation.

In summary, while we agree that the genetic basis of RA is mostly shared between persons of differing global ancestries, we emphasize that there are several reasons to expect that the differences that do exist might be particularly influential to precision medicine in RA.

Trans-Ethnic Fine-Mapping of Candidate Pathogenic Variants in AFF3, CTLA4, NFKBIE, and other RA risk loci

The trans-ethnic fine-mapping (TEFM) of these loci is treated at length elsewhere (see Discussion, Section IV). Here, we only summarize the findings and implications thereof. Because inclusion of African Samples has been shown to greatly increase the power to fine-map pathogenic variants [55], we felt that we were well-positioned to conduct these studies. In total, we fine-mapped >90 RA risk loci.

Prior studies have indicated a role for rs2233434 and rs2233424, two SNVs in *NFKBIE*, but distinguishing between them has been difficult owing to perfect LD in European and Asian populations ($r^2 = 1$). However, our TEFM of this locus indicated that rs2233434 was >400 times more likely than rs2233433 to be the pathogenic variant. Granted the tight linkage described in other populations, we were curious as to why rs2233434 was so much more highly prioritized. We were surprised to learn that rs2233433 is entirely absent from West African populations. Previous studies have suggested that rs2233434 and rs2233433 exert distinct functional effects [56]. The fact that the *NFKBIE* locus is missing rs2233433 but is still associated with RA in African-Americans helps us narrow down not only the variant but potentially also the mechanism by which the pathogenic variant produces risk of RA (see Discussion, Section IV). To this specific example we add the observation that the data from African-Americans with RA enabled identification of 9 *additional* candidate pathogenic variants with posterior probability > 0.8, which is a very large relative gain granted the small amount of data in African-Americans that we included compared to the large number of Europeans and Asians already genotyped (roughly 2,500 : 100,000 or 2.4% of the total dataset).

Therefore, we find that the addition of global samples to existing GWAS is desirable from the perspective of not only distributive justice, but cost efficacy as well.

To date, the greatest emphasis in fine-mapping experiments has been on trimming the size of the credible set down to a small number of variants. However, we noted several different patterns within fine-mapped loci. While many of the loci are dominated by the signal from just one variant (e.g. rs2476601 in *PTPN22*), several others do not display this pattern. In *AFF3* we identified ~5 candidate causal variants, and 4 of them were annotated as *AFF3* eQTLs [57]. We argue that although this story is not as conceptually simple as the case in which there is 1 causal variant, this does not imply the evidence is less compelling. Likewise, in *IFNGR2*, we identified 2 variants just upstream of the conserved TFBS for *ZBTB7A* and *E2F6*, both of which are targeted by EBNA3C in the same fashion.

Alternatively, in *CTLA4*, though just one variant (rs3087243) does account for a majority of the posterior probability, we were unable to determine conclusively whether rs3087243 is the causal variant, or if a $(AT)_{28}$ dinucleotide short tandem repeat (STR) in linkage with rs3087243 is the causal variant (see Figure 4 in Section 4), or potentially both. This ambiguity is difficult to address because the STR is neither genotyped on commercial DNA microarrays nor imputed with high quality. This latter case points to a limitation of not only this study but many contemporary studies of RA.

Limitations of the Current Studies

Limitations of each study are discussed in the Discussion section of each manuscript. In addition, the review article of Section II indicates additional obstacles to progress in the study of RA genetics in general. Additional considerations of a more general nature are presented here.

Statistical Power

In total, the CLEAR dataset included genotyping data for ~2500 RA cases and controls across the Omni 1M, Omni 1S, Omni 5M, and Immunochip study. At this study size, and we expected to be well-powered to detect associations with OR > 1.20 and 0.1 ≤ RAF ≤ 0.9. However, this means we had only moderate power to detect variants with OR around 1.15, which is typical for RA.

We took a multi-faceted approach to address this issue comprehensively. First, after association testing on our African-American samples alone, we ran a joint analysis of European, Asian, and African-American RA. The purpose of this is to increase statistical power to detect associations, as is described in the Section I of this document. While analysis of the African-American data alone identified several variants near our alpha threshold of 5×10^{-9}, we did not detect any loci at a genome wide level of significance until the joint treatment of all three datasets. Second, we anticipated difficulty in distinguishing between variants that have a true association with RA, but do not appear to display one due to statistical noise, from variants lacking a true association altogether. To address this, we employed a formal test of effect concordance to adjudicate whether RA risk variants replicate or not. Despite this, we were unable to make a call for approximately half of the RA risk loci (51 of 102). Both experience with study of Asian populations and theoretical considerations [52] suggest that small effect loci will continue to be discovered even with very large sample sizes. However, initially we believe a treatment of >5,000 RA patients and controls would be sufficient to increase the number of variants that can be assigned as concordant or discordant ($M > 0.8$ or $M < 0.2$) to levels similar to found in East Asians currently (see Section 4 Supplemental Table 3). We also believe that a study of this size in African Americans would further increase fine-mapping accuracy in RA (potentially dramatically), in particular if whole genome sequencing or high-quality imputation is carried out.

DNA Microarray Technology and Array Design

This latter consideration brings us to a discussion of the technologies used. First, nearly all genotyping chips created before 2012 were designed using European genomes and assay many positions not found in persons of African ancestry. Moreover, many sites of common variation in Africans are not assayed by standard genotyping chips. Indeed the first chip that we obtained for CLEAR that was designed with African populations was the MEGA array in 2015. This tends to decrease information capture across the genome for African samples, in particular for the smaller arrays. For the 5M array we achieved on average very high imputation accuracy (>99% of masked genotypes were called accurately) through a combination of conservative settings and deeper imputation of loci we planned to study in depth, as has been recommended in the literature (IMPUTE2 paper).

Nevertheless, we had difficulty in imputing certain variants in which we were interested, such as the $(AT)_n$ dinucleotide repeat described above in the 3'UTR of *CTLA4*. If this occurred only sporadically, it would still be of concern since poorly imputed variants can result in inaccurate estimation of posterior probabilities. However, the real issue is that poor imputation currently occurs systematically, and affects certain classes of genetic variation more frequently than others. For instance, because STR and indels are 1) not frequently genotyped on microarrays and 2) more difficult to call accurately using short-read next generation sequencing (NGS), we caution that any candidate causal variant that is in linkage with such a variant should be scrutinized carefully. Multiple fine-mapping algorithms such as CAVIARBF [36], PAINTOR [37, 38], MANTRA [58], BIMBAM [59], and others [60] handle this issue differently. Regardless of the treatment used, it should be assumed that the credible sets generated are at least partially inaccurate if the true pathogenic variant is not included in the study.

Supporting expression, epigenomic, and experimental data

Finally, several of our results would have benefited from experimental characterization. We took care to generate TEFM results that suggest not only variants to be studied but where possible a biological rationale for how this may occur. For example, in *CTLA4*, one logical follow-up study would be to examine whether rs3087243, the $(AT)_n$ dinucleotide repeat, or both increase *CTLA4* expression in T regulatory cells of RA patients and controls. In addition, a genome editing technology such as CRISPR/CAS9 could be used to abrogate the tight linkage between such variants. Depending on findings, assessment of whether abatacept response correlates with either genotype could be useful granted the clinical interest in that question. RNA-Seq and ATAC-Seq on this same cohort would also greatly increase available options for analytical plans.

Relevance of the current studies in the context of current RA genetic research

At the present time, research into RA genetics is undergoing several important transitions. First, there are very few studies that focus on association testing alone; rather many studies have begun to address the related problems of association testing and fine-mapping together. MANTRA is a good example of a framework that addresses both of these challenges simultaneously [58]. In our study, we draw heavily on METASOFT and PAINTOR, as these enabled us to 1) address concern relating to statistical power that we had and 2) coordinately analyze data from all 3 ethnicities in one fine-mapping experiment. Although in theory the fully Bayesian approach in implemented in CAVIARBF and BIMBAM [36, 59] is capable of integrating multiple datasets from more than one global population, at this time the software would have to run three separate analyses, which of course foregoes the main advantage of being able to triangulate risk by leveraging differences in LD patterns. At any rate, regardless of the specific approaches used, our studies presents a well-defined workflow from GWAS to trans-ethnic meta-analysis of GWA data to trans-ethnic fine-mapping of meta-analysis results. In so doing we position the work as a "post-GWAS" workflow.

We also believe that it is important to address health disparities in genetic research of RA. Since the goal of precision medicine is ultimately to provide tailored an optimized treatment to people based on their unique genetic profile, we believe that if the genetics of RA is insufficiently studied in a given population, this will preclude that population from receiving the benefits of precision medicine. While we therefore believe that these studies are valuable in their own right, we also note that the analysis of this data alongside existing datasets is valuable to the entire community. For example, our study enabled us to greatly improve fine-mapping accuracy, including in loci where the association signal in African-Americans was marginal on its own. To summarize, we believe this study is crucially relevant as an example of how even a small amount of data from a poorly understood population is marginally more valuable than additional samples genotyped in a well-studied population. In this way, the significance of our study extends well beyond the RA literature.

Suggestions relating to optimal design of future genetic research cohorts

We believe future of genetic research into RA and other complex conditions will have less and less to do with association testing, and will come to address the question of "which variants are the functional ones in this disease state and why?" While for the former question it might be defensible to study one or two populations only, for the latter question, a growing body of evidence suggests that trans-ethnic research cohorts will out-perform ethnically homogenous ones [60] . Thus, we believe that in the future, genetic association studies should ideally include individuals from 4 or more global ancestries. For example, in the United States a study admitting Native Americans, African Americans, Indian Americans, Asian Americans and European Americans will be more well-suited to address the goals of the coming age of precision medicine both in terms of information gain and in terms of cost-efficacy.

The Future of Genomic Research into RA and Complex Disease

Three related goals in contemporary complex disease genetic research

I have come to view the present and future of genetic research into complex disease as a progression along a pathway that connects three related goals. The first is the identification of associations and has begun to be addressed by GWAS. Ongoing studies that increase both the size and the completeness of these datasets will be necessary, as current studies do not assay all genetic variants equally well, nor do they comprehensively assess risk in all global populations. The second is the identification of the variants that are pathogenic (or rather, functional in the context of a given disease biology) from among what can be many hundreds of strong associated variants. While fine-mapping experiments of this kind have been ongoing for decades, recent years have increasingly begun to favor the advent of methods that address this question in-silico, in a data-driven, high-throughput fashion [61, 62]. Finally, I am persuaded that the advent of large research cohorts of NGS data will lead us to a point when very plausible genetic variants have been identified in thousands of complex disease loci. The third goal will be to design functional experiments to validate the effect of each candidate pathogenic variant experimentally. Because of the scope of this problem, I believe it is highly likely that ultimately high throughput assays will enable the identification of many such variants simultaneously.

Factors complicating autoimmune disease variant identification

The challenge in characterizing the function of complex disease variants is two-fold. First, subsets of these variants are known to display time-dependent effects, cell-type specific effects, activation state dependency, interaction effects and other complicating factors. This means it is frequently exceedingly easy to generate a null result even if one has an essentially accurate working hypothesis; for example, one could look in the right cell-type and stimulate with the right cocktail of activating factors, but miss the effect by looking at the wrong time-point. The second challenge is the sheer number of variants. Currently there are ~15,000 variants in the NHGRI GWAS catalog, but studies based on polygenic analysis [63], Bayesian inference [52], mixed linear modeling [64], not to mention coalescence theory and population genetics, have suggested that the true number of pathogenic variants for conditions like RA may ultimately number in the thousands *per disease*.

A role for high-throughput assays in overcoming time-dependency, activation-state dependency, cell-type specificity, and other complicating factors

If it is indeed true that such a large number of variants will ultimately be found, then it is likely that the experimental validation of a large majority of GWAS variants will not be carried out in traditional scientific experiments as we have known them. Rather, I suspect they will ultimately be carried out in a high-throughput fashion. This claim would seem to be at odds with the first challenge I mentioned; namely the dependency of the pathogenic effect on one or multiple complicating factors. However, note that techniques have already been developed that can circumvent one or more of these problems. As examples consider the following. STARR-Seq [61] is essentially a massively parallel reporter assay capable of detective enhancer activity quantitatively genome-wide . CRISPRa, or tiled CRISPR activation, can identify stimulus-responsive elements *independently of the cell type in which they act* [62] across genomic segments in excess of 100kb in length. I have also seen data presented recently regarding the construction of databases that catalog temporal course of activity of a large number of signaling molecules and transcription factors. Finally, unifying hypotheses that address specific environmental exposures that may be relevant, will also likely help in functional characterization of categories of these genetic variants once the precise biological rationale can be identified. As an example, consider the recent suggestion that ½ of SLE risk variants map to regions in which the Epstein Barr viral protein EBNA2 sits [46].

Integrating genetic, environmental, and experimental evidence

Thus, while I think there is reason to believe that in the long term it will likely be possible to validate GWAS variants in a high-throughput fashion, in the near-term there are critical unsolved problems. One of the most central of these is to integrate knowledge of diseases like RA that are gained from experimental biology, genetic studies, and epidemiologic studies. While in reality genetics, environment, and biology give rise to one another seamlessly, scholarship on these subjects often belongs to separate literatures and is not well-integrated. While selecting the right concepts to integrate and integrating them in the right way is an art, there are definite steps that can be taken. Systematically coding knowledge gained about a disease state that has accumulated over years based on a particular tool, animal model, or technology into a database (and then publicizing the database) enables researchers of all stripes to rapidly sift through the results of a large literature. In the current study, I leveraged thousands of genome-wide annotations, but considered only a handful of epidemiologic risk factors, for instance. A catalog of the genomic loci affected by, say, smoking or viral infection of a given type could enable enrichment analysis of the same kind based on results from other literatures.

Closing Summary and Future directions

To close, the present studies addressed the genetics of RA in persons of African-American descent. The studies found three novel RA risk loci not previously identified in Europeans or Asians with RA *GPC5*, *RBFOX1* and *CSMD3*. In addition, we validated and disconfirmed >30 RA risk loci in this population. Our trans-ethnic fine-mapping experiments identified 8 additional loci with high confidence.

Two major suggestions of this body of research are:

1) That although the genetic basis of RA is mostly shared in African-Americans with RA, the differences that do exist between populations occur disproportionately among the strongest effect risk loci. For several reasons, these differences are particularly likely to impact precision medicine.
2) Use of trans-ethnic genetic data is not crucial for single variant association testing, except to identify population-specific associations. However, the present studies add to the body of evidence suggesting that trans-ethnic studies of complex diseases easily outperform larger studies in a single ethnically homogenous cohort with respect to fine-mapping candidate variants, a crucial upcoming goal of complex disease genetics. This has clear implications for precision medicine and study design of future GWA studies.

Finally, this work identifies a number of candidate pathogenic variants with plausible mechanisms of action (see the trans-ethnic fine-mapping experiments from Section

IV). Several of these variants are either already known to (i.e. the candidate variant identified in *NFKBIE*) or may possibly (i.e. the candidate variants in *CTLA4*) modulate DMARD activity. The putative effects of several of these candidate variants could be verified using relatively simple functional assays. We therefore suggest these studies indicate clear future directions for further research.

GENERAL LIST OF REFERENCES

1. Lawrence, R.C., et al., *Estimates of the Prevalence of Arthritis and Other Rheumatic Conditions in the United States, Part II.* Arthritis Rheum, 2008. **58**(1): p. 26-35.

2. Myasoedova, E., *Is the incidence of rheumatoid arthritis rising? Results from Olmsted County, Minnesota, 1955-2007.* 2010. **62**(6): p. 1576-82.

3. Lee, D.M. and M.E. Weinblatt, *Rheumatoid arthritis.* Lancet, 2001. **358**(9285): p. 903-11.

4. Wong, J.B., D.R. Ramey, and G. Singh, *Long-term morbidity, mortality, and economics of rheumatoid arthritis.* Arthritis Rheum, 2001. **44**(12): p. 2746-9.

5. Aletaha, D., et al., *2010 Rheumatoid arthritis classification criteria: an American College of Rheumatology/European League Against Rheumatism collaborative initiative.* Arthritis Rheum, 2010. **62**(9): p. 2569-81.

6. Landewe, R. and D. van der Heijde, *Radiographic progression in rheumatoid arthritis.* Clin Exp Rheumatol, 2005. **23**(5 Suppl 39): p. S63-8.

7. Leeb, B.F., et al., *Disease activity measurement of rheumatoid arthritis: Comparison of the simplified disease activity index (SDAI) and the disease activity score including 28 joints (DAS28) in daily routine.* Arthritis Rheum, 2005. **53**(1): p. 56-60.

8. Terao, C., S. Raychaudhuri, and P.K. Gregersen, *Recent Advances in Defining the Genetic Basis of Rheumatoid Arthritis.* Annu Rev Genomics Hum Genet, 2016. **17**: p. 273-301.

9. Karlson, E.W., et al., *A retrospective cohort study of cigarette smoking and risk of rheumatoid arthritis in female health professionals.* Arthritis Rheum, 1999. **42**(5): p. 910-7.

10. Baka, Z., E. Buzas, and G. Nagy, *Rheumatoid arthritis and smoking: putting the pieces together.* Arthritis Res Ther, 2009. **11**(4): p. 238.

11. van Vollenhoven, R.F., *Sex differences in rheumatoid arthritis: more than meets the eye.* BMC Med, 2009. **7**: p. 12.

12. Cutolo, M., et al., *Sex hormones and rheumatoid arthritis.* Autoimmun Rev, 2002. **1**(5): p. 284-9.

13. Ansar Ahmed, S., M.J. Dauphinee, and N. Talal, *Effects of short-term administration of sex hormones on normal and autoimmune mice.* J Immunol, 1985. **134**(1): p. 204-10.

14. Liao, K.P., L. Alfredsson, and E.W. Karlson, *Environmental influences on risk for rheumatoid arthritis.* Curr Opin Rheumatol, 2009. **21**(3): p. 279-83.

15. Chan TD, Brink R. *Affinity-based selection and the germinal center response.* Immunological reviews 2012;247:11-23.

16. Tarlinton DM. *Evolution in miniature: selection, survival and distribution of antigen reactive cells in the germinal centre*. Immunology and cell biology 2008;86:133-8.

17. Zhang X, Ing S, Fraser A, et al. *Follicular helper T cells: new insights into mechanisms of autoimmune diseases*. The Ochsner journal 2013;13:131-9.

18. Lee SK, Bridges SL, Jr., Kirkham PM, Koopman WJ, Schroeder HW, Jr. *Evidence of antigen receptor-influenced oligoclonal B lymphocyte expansion in the synovium of a patient with longstanding rheumatoid arthritis*. The Journal of clinical investigation 1994;93:361-70.

19. Tangye SG, Ma CS, Brink R, Deenick EK. *The good, the bad and the ugly - TFH cells in human health and disease*. Nature reviews Immunology 2013;13:412-26.

20. Okada, Y., et al., *Significant impact of miRNA-target gene networks on genetics of human complex traits.* Sci Rep, 2016. **6**: p. 22223.

21. Naranbhai, V., et al., *Genomic modulators of gene expression in human neutrophils.* Nat Commun, 2015. **6**: p. 7545.

22. Chu Y, Wang F, Zhou M, Chen L, Lu Y. *A preliminary study on the characterization of follicular helper T (Tfh) cells in rheumatoid arthritis synovium*. Acta histochemica 2014;116:539-43.

23. Firestein, G. and I.B. McInnes, *Immunopathogenesis of rheumatoid arthritis.* Immunity, 2017. **46**(2): p. 183-96.

24. Okada, Y., et al., *Genetics of rheumatoid arthritis contributes to biology and drug discovery.* Nature, 2014. **506**(7488): p. 376-81.

25. Sherk, H.H., *Commentaries on the history and cure of diseases. Digitorum Nodi by William Heberden MD.* Clin Orthop Relat Res, 2004(427 Suppl): p. S3-4.

26. Deighton, C.M. and D.J. Walker, *The familial nature of rheumatoid arthritis.* Ann Rheum Dis, 1991. **50**(1): p. 62-5.

27. Knevel, R., et al., *Genetic predisposition of the severity of joint destruction in rheumatoid arthritis: a population-based study.* Ann Rheum Dis, 2012. **71**(5): p. 707-9.

28. Stastny, P., *Association of the B-cell alloantigen DRw4 with rheumatoid arthritis.* N Engl J Med, 1978. **298**(16): p. 869-71.

29. Gregersen, P.K., J. Silver, and R.J. Winchester, *The shared epitope hypothesis. An approach to understanding the molecular genetics of susceptibility to rheumatoid arthritis.* Arthritis Rheum, 1987. **30**(11): p. 1205-13.

30. Raychaudhuri, S., et al., *Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis.* Nat Genet, 2012. **44**(3): p. 291-6.

31. Yamamoto, K., et al., *Genetic studies of rheumatoid arthritis.* Proc Jpn Acad Ser B Phys Biol Sci, 2015. **91**(8): p. 410-22.

32. Shi, J. and S. Lee, *A novel random effect model for GWAS meta-analysis and its application to trans-ethnic meta-analysis.* Biometrics, 2016. **72**(3): p. 945-54.

33. Farh, K.K.H., et al., *Genetic and Epigenetic Fine-Mapping of Causal Autoimmune Disease Variants.* Nature, 2015. **518**(7539): p. 337-43.

34. Han, B. and E. Eskin, *Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies.* Am J Hum Genet, 2011. **88**(5): p. 586-98.

35. Han, B. and E. Eskin, *Interpreting meta-analyses of genome-wide association studies.* PLoS Genet, 2012. **8**(3): p. e1002555.

36. Chen, W., et al., *Fine Mapping Causal Variants with an Approximate Bayesian Method Using Marginal Test Statistics.* Genetics, 2015. **200**(3): p. 719-36.

37. Kichaev, G. and B. Pasaniuc, *Leveraging Functional-Annotation Data in Trans-ethnic Fine-Mapping Studies.* Am J Hum Genet, 2015. **97**(2): p. 260-71.

38. Kichaev, G., et al., *Improved methods for multi-trait fine mapping of pleiotropic risk loci.* Bioinformatics, 2017. **33**(2): p. 248-255

39. Nielsen, J.B., et al., *Genome-wide association study of 1 million people identifies 111 loci for atrial fibrillation.* bioRxiv, 2018.

40. Mizukami, T., T. Kohno, and M. Hattori, *CUB and Sushi multiple domains 3 regulates dendrite development.* Neurosci Res, 2016. **110**: p. 11-7.

41. Welter, D., et al., *The NHGRI GWAS Catalog, a curated resource of SNP-trait associations.* Nucleic Acids Res, 2014. **42**(Database issue): p. D1001-6.

42. Li, L., et al., *Oncogenic activation of glypican-3 by c-Myc in human hepatocellular carcinoma.* Hepatology, 2012. **56**(4): p. 1380-90.

43. Aterido, A., et al., *Novel insights into the regulatory architecture of CD4+ T cells in rheumatoid arthritis.* PLoS One, 2014. **9**(6): p. e100690.

44. Suhre, K., et al., *Connecting genetic risk to disease end points through the human blood plasma proteome.* Nat Commun, 2017. **8**: p. 14357.

45. Gorenshteyn, D., et al., *Interactive Big Data Resource to Elucidate Human Immune Pathways and Diseases.* Immunity, 2015. **43**(3): p. 605-14.

46. Harley, J.B., et al., *Transcription factors operate across disease loci, with EBNA2 implicated in autoimmunity.* Nat Genet, 2018. **50**(5): p. 699-707.

47. Pei, Y., et al., *EBV Nuclear Antigen 3C Mediates Regulation of E2F6 to Inhibit E2F1 Transcription and Promote Cell Proliferation.* PLoS Pathog, 2016. **12**(8): p. e1005844.

48. Tang, Q., et al., *Expression of Interferon-gamma Receptor Genes in PBMCs is Associated with Rheumatoid Arthritis and Its Radiographic Severity in African Americans.* Arthritis Rheumatol, 2015. **67**(5): p. 1165-70.

49. Viatte, S., et al., *Association of HLA-DRB1 haplotypes with rheumatoid arthritis severity, mortality, and treatment response.* Jama, 2015. **313**(16): p. 1645-56.

50. Maher, M.C., *Population genetics of rare variants and complex diseases.* 2012. **74**(0): p. 118-28.

51. Dummer, P.D., et al., *APOL1 kidney disease risk variants – an evolving landscape.* Semin Nephrol, 2015. **35**(3): p. 222-36.

52. Stahl, E.A., et al., *Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis.* Nat Genet, 2012. **44**(5): p. 483-9.

53. Moonesinghe, R., et al., *Estimating the contribution of genetic variants to difference in incidence of disease between population groups.* Eur J Hum Genet, 2012. **20**(8): p. 831-6.

54. Nelson, M.R., et al., *The support of human genetic evidence for approved drug indications.* Nat Genet, 2015. **47**(8): p. 856-60.

55. Spain, S.L. and J.C. Barrett, *Strategies for fine-mapping complex traits.* Hum Mol Genet, 2015. **24**(R1): p. R111-9.

56. Myouzen, K., et al., *Functional variants in NFKBIE and RTKN2 involved in activation of the NF-kappaB pathway are associated with rheumatoid arthritis in Japanese.* PLoS Genet, 2012. **8**(9): p. e1002949.

57. Danila, M.I., et al., *Dense Genotyping of Immune-Related Regions Identifies Loci for Rheumatoid Arthritis Risk and Damage in African Americans.* Mol Med, 2017. **23**: p. 177-87.

58. Morris, A.P., *Transethnic Meta-Analysis of Genomewide Association Studies.* Genet Epidemiol, 2011. **35**(8): p. 809-22.

59. Servin, B. and M. Stephens, *Imputation-based analysis of association studies: candidate regions and quantitative traits.* PLoS Genet, 2007. **3**(7): p. e114.

60. Asimit, J.L., et al., *Trans-ethnic study design approaches for fine-mapping.* Eur J Hum Genet, 2016. **24**(9): p. 1330-6.

61. Arnold, C.D., et al., *Genome-wide quantitative enhancer activity maps identified by STARR-seq.* Science, 2013. **339**(6123): p. 1074-7.

62. Simeonov, D.R., et al., *Discovery of stimulation-responsive immune enhancers with CRISPR activation.* Nature, 2017. **549**(7670): p. 111-115.

63. Purcell, S.M., et al., *Common polygenic variation contributes to risk of schizophrenia and bipolar disorder.* Nature, 2009. **460**(7256): p. 748-52.

64. Yang, J., et al., *Advantages and pitfalls in the application of mixed-model association methods.* Nat Genet, 2014. **46**(2): p. 100-6.

APPENDIX: IRB APPROVAL FORMS

Project Revision/Amendment Form

Form version: June 26, 2012

In MS Word, click in the white boxes and type your text; double-click checkboxes to check/uncheck.

- **Federal regulations require IRB approval before implementing proposed changes. See Section 14 of the IRB Guidebook for Investigators for additional information.**
- **Change means any change, in content or form, to the protocol, consent form, or any supportive materials (such as the Investigator's Brochure, questionnaires, surveys, advertisements, etc.). See Item 4 for more examples.**

1. Today's Date	12/5/2013

2. Principal Investigator (PI)

Name (with degree)	S. Louis Bridges, Jr.MD,PhD	**Blazer ID**	lbridges
Departmen t	Rheumatology	**Division (if applicable)**	Clinical Immunology and Rheumatology
Office Address	178 Shelby	**Office Phone**	4-4616
E-mail	Lbridges@uab.edu	**Fax Number**	4-1564

Contact person who should receive copies of IRB correspondence (Optional)

Name	Stephanie Ledbetter	**E-Mail**	sledbetter@uab.edu
Phone	4-7423	**Fax Number**	4-4616
Office Address (if different from PI)	177 F Shelby		

3. UAB IRB Protocol Identification

3.a. Protocol Number	**X061215004**
3.b. Protocol Title	Continuation of the Consortium for the Longitudinal Evaluation of African-Americans with Rheumatoid Arthritis, Coordinating Center

3.c. Current Status of Protocol—Check ONE box at left; provide numbers and dates where applicable

□	**Study has not yet begun**		**No participants, data, or specimens have been entered.**	
□	**In progress, open to accrual**		**Number of participants, data, or specimens entered:**	
□	**Enrollment temporarily suspended by sponsor**			
□	**Closed to accrual, but procedures continue as defined in the protocol (therapy, intervention, follow-up visits, etc.)**			
	Date closed:		**Number of participants receiving interventions:** **Number of participants in long-term follow-up only:**	
☒	**Closed to accrual, and only data analysis continues**			
	Date closed:		**Total number of participants entered:**	1063 RA, 550 controls

4. Types of Change

Check all types of change that apply, and describe the changes in Item 5.c. or 5.d. as applicable. To help avoid delay in IRB review, please ensure that you provide the required materials and/or information for each type of change checked.

☐ **Protocol revision (change in the IRB-approved protocol)**
In Item 5.c., if applicable, provide sponsor's protocol version number, amendment number, update number, etc.

☐ **Protocol amendment (addition to the IRB-approved protocol)**
In Item 5.c., if applicable, provide funding application document from sponsor, as well as sponsor's protocol version number, amendment number, update number, etc.

☒ **Add or remove personnel**
In Item 5.c., include name, title/degree, department/division, institutional affiliation, and role(s) in research, and address whether new personnel have any conflict of interest. See "Change in Principal Investigator" in the IRB Guidebook if the principal investigator is being changed.

- ☐ **Add graduate student(s) or postdoctoral fellow(s) working toward thesis, dissertation, or publication**
  In Item 5.c., (a) identify these individuals by name; (b) provide the working title of the thesis, dissertation, or publication; and (c) indicate whether or not the student's analysis differs in any way from the purpose of the research described in the IRB-approved HSP (e.g., a secondary analysis of data obtained under this HSP).

☐ **Change in source of funding; change or add funding**
In Item 5.c., describe the change or addition in detail, include the applicable OSP proposal number(s), and provide a copy of the application as funded (or as submitted to the sponsor if pending). Note that some changes in funding may require a new IRB application.

☐ **Add or remove performance sites**
In Item 5.c., identify the site and location, and describe the research-related procedures performed there. If adding site(s), attach notification of permission or IRB approval to perform research there. Also include copy of subcontract, if applicable. If this protocol includes acting as the Coordinating Center for a study, attach IRB approval from any non-UAB site added.

☐ **Add or change a genetic component or storage of samples and/or data component—this could include data submissions for Genome-Wide Association Studies (GWAS)**
To assist you in revising or preparing your submission, please see the IRB Guidebook for Investigators or call the IRB office at 934-3789.

☐ **Suspend, re-open, or permanently close protocol to accrual of individuals, data, or samples (IRB approval to remain active)**
In Item 5.c., indicate the action, provide applicable dates and reasons for action; attach supporting documentation.

☐ **Report being forwarded to IRB (e.g., DSMB, sponsor or other monitor)**
In Item 5.c., include date and source of report, summarize findings, and indicate any recommendations.

☐ **Revise or amend consent, assent form(s)**
Complete Item 5.d.

☐ **Addendum (new) consent form**
Complete Item 5.d.

☐ **Add or revise recruitment materials**
Complete Item 5.d.

☐ **Other (e.g., investigator brochure)**
Indicate the type of change in the space below, and provide details in Item 5.c. or 5.d. as applicable.
Include a copy of all affected documents, with revisions highlighted as applicable.

▶

5. Description and Rationale

In Item 5.a. and 5.b, check Yes **or** No **and see instructions for** Yes **responses.**
In Item 5.c. and 5.d, describe—and explain the reason for—the change(s) noted in Item 4.

□Yes □ No	**5.a. Are any of the participants enrolled as normal, healthy controls?** If yes, describe in detail in Item 5.c. how this change will affect those participants.
□Yes □ No	**5.b. Does the change affect subject participation, such as procedures, risks, costs, location of services, etc.?** If yes, FAP-designated units complete a FAP submission and send to fap@uab.edu. Identify the FAP-designated unit in Item 5.c. For more details on the UAB FAP, see www.uab.edu/cto.

5.c. Protocol Changes: In the space below, briefly describe—and explain the reason for—all change(s) to the protocol.

▶ **Vincent A. Laufer, B.A.**, is a current PhD student and participant in the NIH Medical Scientist Training Program (MD/PhD Program. As part of this program, Vincent is studying statistical genetics, complex disease, and rheumatology to gain additional research training and experience. He will be undertaking analysis of both CLEAR Whole Genome and CLEAR GWAS data in order to elucidate genetic factors associated with Rheumatoid Arthritis as part of the proposed research described in the IRB-approved HSP for this protocol.

5.d. Consent and Recruitment Changes: In the space below,
(a) describe all changes to IRB-approved forms or recruitment materials and the reasons for them;
(b) describe the reasons for the addition of any materials (e.g., addendum consent, recruitment); and
(c) indicate either how and when you will reconsent enrolled participants or why reconsenting is not necessary (not applicable for recruitment materials).

Also, indicate the number of forms changed or added. For new forms, provide 1 copy. For revised documents, provide 3 copies:

- **a copy of the currently approved document (showing the IRB approval stamp, if applicable)**
- **a revised copy highlighting all proposed changes with "tracked" changes**
- **a revised copy for the IRB approval stamp.**

▶

Signature of Principal Investigator____ Date____________________

FOR IRB USE ONLY

☐ Received & Noted ☐ Approved Expedited* ☐ To Convened IRB

____ ____________________

Signature (Chair, Vice-Chair, Designee) Date

DOLA ____________________

Change to Expedited Category **Y / N / NA**

*No change to IRB's previous determination of approval criteria at 45 CFR 46.111 or 21 CFR 56.111

Investigator’s Progress Report

Form version June 26, 2012

In MS Word, click in the white boxes and type your text; double-click checkboxes to check/uncheck.

☒ **Continuing Review (Complete Items 1-11)** **—OR—** ☐ **Final Report—all protocol-related activities are complete, including data analysis (Complete Items 1-10, and Item 12)**	**—FOR—**	☒ **Expedited Review** **—OR—** ☐ **Convened (Full) Review**

1. Dates

Today’s Date	12/09/2014	**To help avoid delay, respond to all required items in the format provided, and include requested materials.**
Starting Date of Project	06/01/2012	**If previous approval expires before approval is officially re-issued by the Office of the IRB, all work on the protocol must cease.**
Date of Last IRB Approval	03/04/2014	**The IRB recommends applying for continuing review 4-6 weeks before expiration of current approval. (See schedule.)**

2. Principal Investigator (PI)

Name (with degree)	Maria I. Danila, MD, MSc	**Blazer ID**	mdanila
Department	Medicine	**Division**	Clin. Immunol./Rheumatol.
Office Address	FOT 858A	**Office Phone**	5-1961
E-mail	mdanila@uab.edu	**Fax Number**	4-4198

PI Contact who should receive copies of IRB correspondence (Optional)

Name	Stephanie Ledbetter	**E-mail**	sledbetter@uab.edu
Phone	4-7423	**Fax Number**	4-1564
Office Address (if different from PI)	177 F SHEL		

3. UAB IRB Protocol Identification

Protocol Number	X120308011
Protocol Title	Genetic Architecture of Rheumatoid Arthritis in African Americans
Study Sponsor(s)	NIH
OSP Proposal Number (9 digits)	000407797
Note. If the source or amount of funding for this project has changed, include the new or revised funding application and provide the new OSP Proposal Number:	

4. Purpose

In two or three sentences, briefly summarize the purpose of this protocol, and related studies if applicable. Please use non-technical language, and write more for adults with general knowledge than for specialists.

▶ This protocol is funded by an NIH sponsored career development award. Specifically, this research project proposes to identify genetic associations with the risk and severity of RA in African Americans, by analyzing clinical and genotyping data available from previous protocols, such as CLEAR (Consortium for the Longitudinal Evaluation of African Americans with Rheumatoid Arthritis), and building predictive models for RA risk and outcomes. Only existing data records will be used for this study; no participants will be enrolled under this protocol.

5. Screened, entered, or otherwise accessed by the UAB Investigator(s). Include numbers for individuals, specimens, data records, charts, etc., as applicable to the protocol.

5.a. Number screened for study entry since the start of the project? (See 5.d.i.)	2700
5.b. Number entered in study since the start of the project? (See 5.d.ii.)	2700
5.c. Number entered in study since the last IRB review?	0

5.d. Complete the grids below to show how many have been screened and entered, along with their age or age range, gender, and race/ethnicity. Copy/paste the grids to repeat them for additional groups (e.g., controls, sub-studies) if needed.

Note. If the research involves minors (<19 years of age), the PI must provide a separate, signed memorandum that either (a) confirms the previously assigned Children's Risk Level (CRL) number or (b) reassigns it and gives the reasons it has changed.

5.d.i. Number Screened (Totals = 5.a.)

Race / Ethnicity	Male		Female	
	Age Range	Number Screened	Age Range	Number Screened
Caucasian				
African American				
Native American				
Asian				
Hispanic				
Other				

5.d.ii. Number Entered (Totals = 5.b.)

Race / Ethnicity	Male		Female	
	Age Range	Number Entered	Age Range	Number Entered
Caucasian				
African American				
Native American				
Asian				
Hispanic				
Other				

☒ Check the box at the left if the demographic information was not available (e.g., not collected for screening; collecting only specimens or data records and did not have access to the information)

6. Protocol Staff Listing

For each individual currently involved in the design, conduct, and reporting of the research, list the person's name, role in research, and CIRB status in the table below. Copy/paste the table for each individual.

Financial Interests Related to the Research—Conflict of Interest (COI)

Human subjects research involving a disclosed financial interest on the part of any UAB employee or their immediate family is subject to IRB review following review by the UAB Conflict of Interest Review Board (CIRB). The following definitions apply: *Immediate family* means spouse or a dependent of the employee. *Dependent* is any person, regardless of his or her legal residence or domicile, who receives 50% or more of his or her support from the public official or public employee or his or her spouse or who resided with the public official or public employee for more than 180 days during the reporting period. *Financial Interest Related to the Research* means financial interest in the sponsor, product or service being tested, or competitor of the sponsor.

If one of the four items listed below is marked for an individual, a financial interest disclosure must be submitted to or currently on file with the CIRB. The IRB must receive a completed CIRB Evaluation before it will conduct its review.

COI 1 An ownership interest, stock options, or other equity interest related to the research of any value.

COI 2 Compensation related to the research unless it meets two tests:

- Less than $10,000 in the past year when aggregated for the immediate family.
- Amount will not be affected by the outcome of the research.

COI 3 Proprietary interest related to the research including, but not limited to, a patent, trademark, copyright, or licensing agreement.

COI 4 Board of executive relationship related to the research, regardless of compensation.

FULL NAME	CONFLICT OF INTEREST (COI)
Maria Ioana Danila	☒None, or □1 □ 2 □3 □4 If any, is MOU in place? □ Yes □No
S. Louis Bridges	☒None, or □1 □ 2 □3 □4 If any, is MOU in place? □ Yes □No
Donna K. Arnett	☒None, or □1 □ 2 □3 □4 If any, is MOU in place? □ Yes □No
David B. Allison	☒None, or □1 □ 2 □3 □4 If any, is MOU in place? □ Yes □No
Hemant Tiwari	☒None, or □1 □ 2 □3 □4 If any, is MOU in place? □ Yes □No
Krish Ramen	☒None, or □1 □ 2 □3 □4 If any, is MOU in place? □ Yes □No
Peter Gregersen	☒None, or □1 □ 2 □3 □4 If any, is MOU in place? □ Yes □No
Degui Zhi	☒None, or □1 □ 2 □3 □4 If any, is MOU in place? □ Yes □No
Andy Westfall	☒None, or □1 □ 2 □3 □4 If any, is MOU in place? □ Yes □No
Vincent Laufer (being added)	☒None, or □1 □ 2 □3 □4 If any, is MOU in place? □ Yes □No

7. Information Since the Date of Last IRB Review

- **Mark at least one checkbox to indicate the type(s) of information received since the Date of Last IRB Review.**
- **Please summarize each type of information, and provide details and copies as requested.**

7.a. You received multi-center trial reports that you have not previously forwarded to the IRB. Attach a copy and, in the space below, provide the date	□ Yes ☒ No Multi-Center Trial Report

and source of report, and summarize the findings and any recommendations: ▶	
7.b. You received data and safety or other monitoring reports (e.g., DSMB, sponsor site visit). Even if you have already forwarded a copy to the IRB, attach a copy and, in the space below, provide the date and source of report, and summarize the findings and any recommendations: ▶	□ Yes ☒ No Data Safety or Other Monitoring Report
7.c. You learned of literature published about this research. Attach the publication or provide its web address, and summarize the published findings here: ▶	□ Yes ☒ No Published Literature
7.d. You learned of other relevant information regarding this research, especially about risks associated with the research. Attach a copy of the source and/or summarize below, and check "Other Information" at right. Check "Affects Willingness" also if this information might affect a participant's willingness to continue in the research, and describe the effects on participants here: ▶	□ Yes ☒ No Other Information □ Yes ☒ No Affects Willingness
7.e. You have received another type of information. Summarize the information here, including details relevant to participants: ▶	□ Yes ☒ No Other Type of Information

8. Events Since the Date of Last IRB Review

Mark at least one checkbox to show event(s) that have occurred since the Date of Last IRB Review. Please summarize all events, and provide specific details and/or copies as requested.

8.a. One or more "reportable events" have occurred, which may constitute unanticipated problems involving risks to participants or others. Attach UAB Problem Report even if already reported to the IRB; attach UAB Problem Summary Sheet; provide brief narrative summary (2-3 sentences) of any trends or increases in frequency or severity noted, or enter "None noted" here: ▶	□ Yes ☒ No Reportable Events (Table A)
8.b. Participants have experienced harms (expected or unexpected, serious or not serious) that do not meet the UAB IRB criteria for "reportable events." Attach UAB Problem Summary Sheet; provide brief narrative summary (2-3 sentences) of any trends or increases in frequency or severity noted, or enter "None noted" here:	□ Yes ☒ No Other Events (Table B)

▶

8.c. You have had one or more problems obtaining informed consent.
Briefly describe the problems here:

☐ Yes ☒ No
Consent Problems

▶

8.d. You have received complaints about the research.
Briefly describe the number and nature of the complaints:

☐ Yes ☒ No
Complaints

▶

8.e. One or more participants withdrew, or were withdrawn from, the research.
Indicate here the number of withdrawals and the reason for each:

☐ Yes ☒ No
Withdrawals

▶

8.f. Participants have experienced research-related benefits.
For example, “60% of participants in the treatment group appear to have reduced symptoms or reduced severity of symptoms, compared with 10% in the placebo group.”
Briefly describe the benefits here:

☐ Yes ☒ No
Benefits

▶

8.g. The risks, potential benefits, or both of this research have changed.
Briefly describe the changes here:

☐ Yes ☒ No
Change in Risk or Benefit

▶

8.h. Events have occurred that relate to participant safety but do not fit into the categories listed above.
Briefly describe the events here:

☐ Yes ☒ No
Other Events

▶

9. Protocol and/or Informed Consent Modifications

Check the applicable boxes to indicate modifications made since Date of Last IRB Review (Yes to 9.a.) or requested with this renewal (Yes to 9.b.). Please provide the details and materials requested.

9.a. Previous Modifications

Since the last IRB review, have you made modifications to the protocol, consent process, or consent document?

☐Yes ☒No

If Yes, have the modifications been approved by the IRB?

☐Yes—Provide a copy of each amendment form stamped “Approved” by the IRB during this approval period.

☐No—In the space below, justify making the modification without prior IRB approval:

▶

9.b. Modifications Requested With This Renewal

Are you requesting IRB review of changes to the protocol (e.g., procedures, personnel, recruitment)? If

☒ Yes ☐ No
Protocol Changes

so, check “Yes” and describe them in the space below. If adding personnel, indicate role in research, provide full name and UAB department/division, and address conflict of interest.

▶ Vincent A. Laufer, B.A., is a current MD/PhD student and participant in the NIH Medical Scientist Training Program. As part of this program, Vincent is studying statistical genetics, complex disease, and rheumatology to gain additional research training and experience. He will be undertaking analysis of both of data in order to elucidate genetic factors associated with Rheumatoid Arthritis as part of the proposed research described in the IRB-approved HSP for this protocol. Robert Plenge is being removed from this protocol, as he is no longer an investigator on this project, due to change in his employment.

Are you requesting IRB review of changes to the consent process and/or form(s)? If so, check the applicable “Yes” box and, in the space below, describe the changes.	□ Yes ☒ No Consent Process Changes □ Yes ☒ No Consent Document Changes

If the changes affect the consent form(s), indicate the number of consent-assent forms used for this protocol, and describe the changes to each form:
(a) describe all changes to IRB-approved forms and the reasons for them;
(b) describe the reasons for the addition of any materials (e.g., addendum consent); and
(c) indicate either how and when you will reconsent enrolled participants or why reconsenting is not necessary.

Also, indicate the number of forms changed or added. For new forms, provide 1 copy. For revised documents, provide 3 copies:

- **a copy of the currently approved document (showing the IRB approval stamp, if applicable)**
- **a revised copy highlighting all proposed changes with “tracked” change**
- **a revised copy for the IRB approval stamp.**

▶

10. Gene Therapy, Gene Transfer, Recombinant DNA

If this study involves	□ Gene therapy	□ Gene transfer	□ Recombinant DNA	☒ None of these
	Complete this item, and include memorandum with original signatures of Gene Therapy Review Panel addressing the risk-benefit ratio, any recommendations, and the CRL if applicable.			**Go to Item 11.**

10.a. Has the Panel's assessment of the risk-benefit ratio of this project changed? If yes, please explain below.	□ Yes □ No Risk-Benefit Change

▶

10.b. Does the Panel have any recommendations regarding the protocol or the consent form? If yes, please explain below.	□ Yes □ No Panel Recommendations

▶

Note. If the research involves minors (<19 years old), the panel's memo must either confirm the previously assigned CRL number or reassign it and give the reasons it has changed.

11. Continuing Review—Complete only if you want to renew IRB approval so that protocol-related activities can continue.

11.a. Accrual Status—Indicate whether the study is "NOT YET OPEN," "OPEN," or "CLOSED" (described below) and provide the details requested for that accrual status.

NOT YET OPEN: No individuals have been screened or entered.	□ Not Yet Open
OPEN: The study could still enroll more individuals, add more specimens, review more records, etc. • **Attach a copy of the most recently approved consent form(s) OR note in the space below that the IRB has waived informed consent and/or use of a consent form.** • **Describe plans for future accrual and/or enrollment here:**	□ Open

▶

CLOSED: No more individuals will be enrolled, no more specimens or records will be added.	☒ Closed
If the study is closed, is a consent form being submitted for review? If "Yes," explain why in the space below.	□ Yes ☒ No Closed & Consent Form
• **Indicate the date closed to accrual:**	**12/31/2011** Date Closed
• **Choose one status to describe accrued participants, specimens, records:**	Check **ONE** Status Below:
One or more is still receiving procedures as defined in the protocol (therapy, intervention, follow-up visits, etc.)	□On protocol procedure
All are off protocol-driven procedures, in long-term follow-up only	□In long-term follow-up
All are off protocol-driven procedures, in data analysis only	☒In data analysis

▶ This study uses existing clinical and genotyping data from CLEAR for data analysis.

▶ All data collection from the contributing study (CLEAR) was closed on 12/31/2011, and is in data analysis.

11.b. Describe any interim findings from this research. Please note that the IRB expects to receive findings on any protocol approved for 5 years.

▶ No interim or significant findings to be reported at this time.

12. Final Report—Complete only if you want to end IRB approval after all protocol-related data analyses are complete and no further work on the protocol will be done.

12.a. On what date were the final data analyses completed?	Final Date

12.b. Summarize the final findings from this protocol:

▶

12.c. Who will be responsible for managing and storing the data records, including any and all research-related electronic files and paper documents?

Name	
UAB Dept/Div, or Employer	
Work Address	
Daytime Telephone	

12.d. Describe the storage plan. How will data records be stored—on paper, computers, or both? How will they be protected from damage, unauthorized release, loss, and theft? How long will the data be stored?

▶

12.e. At the end of the storage period, will the data records be destroyed, archived, or transferred? Describe the plan in detail.	□Destroy □Archive □Transfer

▶

Note. **Specimens may be stored only if/as described in the IRB-approved protocol. Data records must be stored as described in the sponsor's protocol or contract if applicable, and/or in the UAB Health System Record Retention Policy. Anyone wishing to use these data or specimens for secondary research purposes or for purposes preparatory to secondary research must obtain prior IRB review and approval.**

Signature of Principal Investigator: ______________________________

Date: ______________________________

FOR IRB USE ONLY – Expedited Review

Change to Expedited Category Y / N

No change to IRB's previous determination of approval criteria at 45 CFR 46.111 or 21 CFR 56.111 □

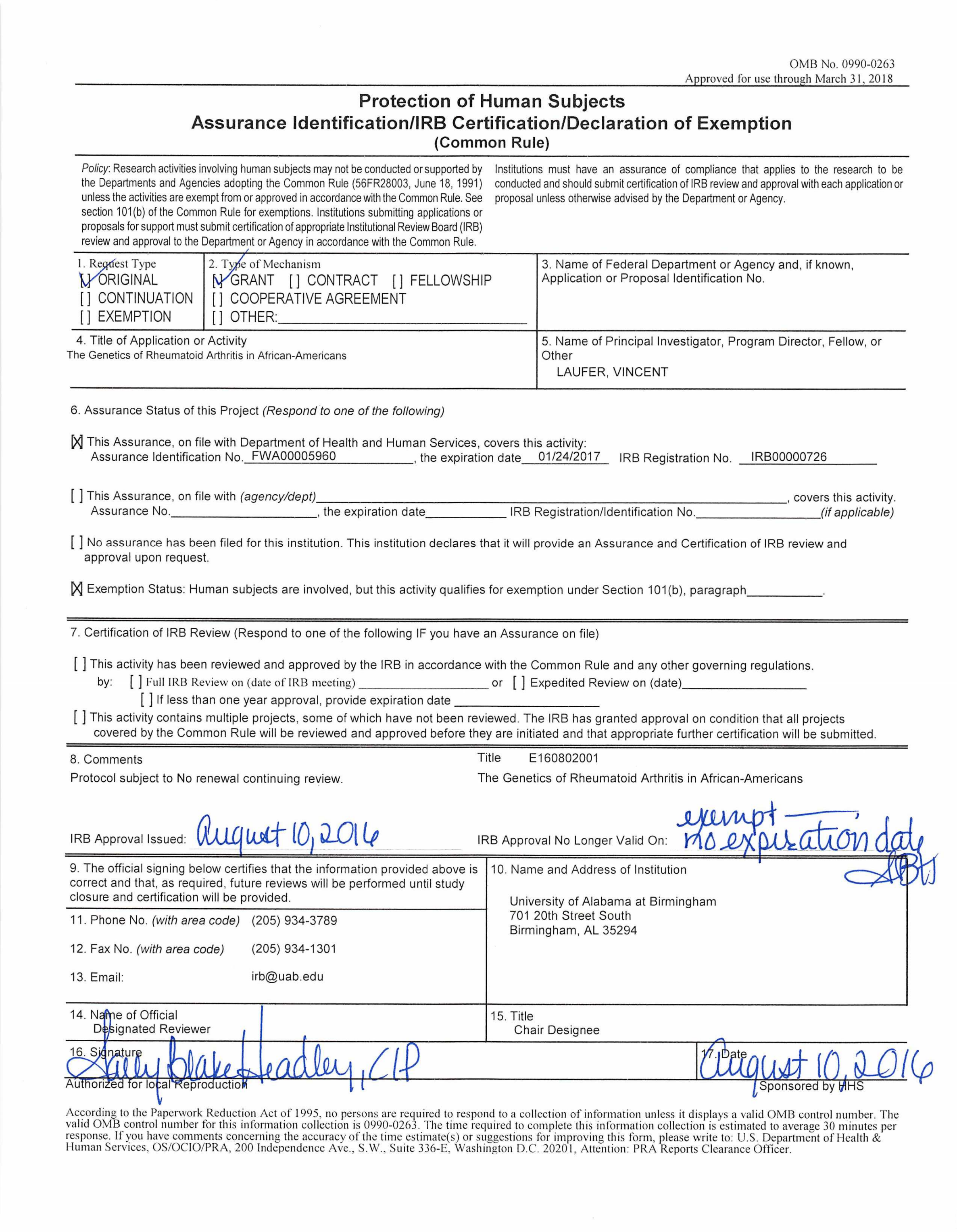
____

____

____

Signature (Chair, Vice-Chair, Designee)

Date



OMB No. 0990-0263
Approved for use through March 31, 2018

Protection of Human Subjects
Assurance Identification/IRB Certification/Declaration of Exemption
(Common Rule)

Policy: Research activities involving human subjects may not be conducted or supported by the Departments and Agencies adopting the Common Rule (56FR28003, June 18, 1991) unless the activities are exempt from or approved in accordance with the Common Rule. See section 101(b) of the Common Rule for exemptions. Institutions submitting applications or proposals for support must submit certification of appropriate Institutional Review Board (IRB) review and approval to the Department or Agency in accordance with the Common Rule.

Institutions must have an assurance of compliance that applies to the research to be conducted and should submit certification of IRB review and approval with each application or proposal unless otherwise advised by the Department or Agency.

1. Request Type
[✓] ORIGINAL
[ ] CONTINUATION
[ ] EXEMPTION

2. Type of Mechanism
[✓] GRANT [ ] CONTRACT [ ] FELLOWSHIP
[ ] COOPERATIVE AGREEMENT
[ ] OTHER:______________________

3. Name of Federal Department or Agency and, if known, Application or Proposal Identification No.

4. Title of Application or Activity
The Genetics of Rheumatoid Arthritis in African-Americans

5. Name of Principal Investigator, Program Director, Fellow, or Other
LAUFER, VINCENT

6. Assurance Status of this Project *(Respond to one of the following)*

☒ This Assurance, on file with Department of Health and Human Services, covers this activity:
Assurance Identification No. FWA00005960, the expiration date 01/24/2017 IRB Registration No. IRB00000726

[ ] This Assurance, on file with *(agency/dept)*______________________, covers this activity.
Assurance No.______________, the expiration date__________ IRB Registration/Identification No.______________ *(if applicable)*

[ ] No assurance has been filed for this institution. This institution declares that it will provide an Assurance and Certification of IRB review and approval upon request.

☒ Exemption Status: Human subjects are involved, but this activity qualifies for exemption under Section 101(b), paragraph__________.

7. Certification of IRB Review (Respond to one of the following IF you have an Assurance on file)

[ ] This activity has been reviewed and approved by the IRB in accordance with the Common Rule and any other governing regulations.
by: [ ] Full IRB Review on (date of IRB meeting) ______________ or [ ] Expedited Review on (date)______________
[ ] If less than one year approval, provide expiration date ______________
[ ] This activity contains multiple projects, some of which have not been reviewed. The IRB has granted approval on condition that all projects covered by the Common Rule will be reviewed and approved before they are initiated and that appropriate further certification will be submitted.

8. Comments
Protocol subject to No renewal continuing review.

Title E160802001
The Genetics of Rheumatoid Arthritis in African-Americans

IRB Approval Issued: August 10, 2016

IRB Approval No Longer Valid On: exempt — no expiration date
SBH

9. The official signing below certifies that the information provided above is correct and that, as required, future reviews will be performed until study closure and certification will be provided.

11. Phone No. *(with area code)* (205) 934-3789

12. Fax No. *(with area code)* (205) 934-1301

13. Email: irb@uab.edu

10. Name and Address of Institution
University of Alabama at Birmingham
701 20th Street South
Birmingham, AL 35294

14. Name of Official
Designated Reviewer

15. Title
Chair Designee

16. Signature
Sally Blake Headley, CIP

17. Date
August 10, 2016

Authorized for local Reproduction

Sponsored by HHS

According to the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number. The valid OMB control number for this information collection is 0990-0263. The time required to complete this information collection is estimated to average 30 minutes per response. If you have comments concerning the accuracy of the time estimate(s) or suggestions for improving this form, please write to: U.S. Department of Health & Human Services, OS/OCIO/PRA, 200 Independence Ave., S.W., Suite 336-E, Washington D.C. 20201, Attention: PRA Reports Clearance Officer.