ABSTRACT
Sjögren’s syndrome is a complex autoimmune disease that involves dysregulation of immune responses that preferentially target exocrine glands. Systemic manifestations vary and may involve nearly every organ system. Genetic studies to date are in their infancy relative to other autoimmune diseases such as systemic lupus erythematosus and rheumatoid arthritis, each with more than 100 genetic associations now established. However, recent work in SS has successfully established associations that shed light on pathophysiology and implicate aberrant innate and adaptive immune responses. In this review, we provide an overview of genetic approaches used to identify risk variants in SS, discuss major findings and their relevance to SS, and describe the future directions that are likely to lead to understanding fundamental causes of this disease and new opportunities for improving clinical care.

Introduction
Sjögren’s syndrome (SS) is one of the most common autoimmune rheumatic diseases, second only to rheumatoid arthritis (RA) in prevalence (1). Data demonstrating the autoimmune nature of the disease is convincing. Patients commonly have antibodies binding self in their sera. The most common of these are antinuclear antibodies as well as anti-Ro (or SSA) and anti-La (or SSB) (2). In addition, a number of other specificities are found including rheumatoid factor, anti-centromere and anti-muscarinic receptor 3 (3). Another feature of disease implicating an autoimmune etiology is the presence of a lymphocytic infiltrate in the commonly involved organs, namely, the salivary and lacrimal glands. This characteristic infiltrate is a critical factor in diagnosis and research classification (4-7). Nonetheless, in almost every aspect, SS is less well studied than other rheumatic illnesses such as RA, systemic lupus erythematosus (SLE), systemic sclerosis (SSc) and idiopathic inflammatory myositis (IIM). The realm of genetics is no exception. For these other diseases, family and twin studies definitively indicate that there is a genetic component to individual risk. Family and twin studies in SS are much less robust (8). For example, Lee and Yoo reported identical twins, one with definite SS and one with possible or early SS (9). Bolsted et al. described monozygotic twin sisters and their mother, all with SS (10), while monozygotic twins concordant for SS and highly similar immunochromical properties of anti-Ro have been reported (11). On the other hand, concordant dizygotic twins have also been reported (12). Monozygotic twins, one with SS and one with SLE were encountered in another report (13). Familial SS is also uncommonly reported (14, 15). Thus, estimates of the genetic contribution to SS based on twin concordance cannot be made. As noted by Anaya et al., the evidence for a genetic basis for SS prior to the era of genome-wide association studies was poor (16). In fact, genetic studies on SS were carried out largely on the basis of extrapolation from other related illnesses such as SLE and RA. Nonetheless, such studies were performed and demonstrate the genetic structure of the disease. This paper will review these results for the non-geneticist.

Genetic studies: a brief primer
Genetic association
Genetic association is generally studied in cohorts of subjects with a particular disease and controls who do not have this disease (case-control design), but can also be studied in families (family-based). Controls must be carefully matched in terms of ethnic and geographic origin in order to avoid population stratification, which is a systematic

Competing interests: R.H. Scofield has received honoraria and travel expenses from Eisai Korea for talk to Korean Rheumatology Association. The other co-authors have declared no competing interests.
difference in allele frequency between subpopulations with different ancestry. Thus, investigators are attempting to find a genotype that correlates with a phenotype more than expected by chance alone. Genetic association exists when a variation in the genetic makeup is found in a statistically significant degree more commonly among those with the disease in question compared to those without said disease.

**Single nucleotide polymorphisms (SNPs)**

The human DNA complement contains about 3 billion base pairs distributed among 23 pairs of chromosomes. These include 22 autosomal chromosomes and two sex chromosomes, either XX in women, or XY in men. Each base can be either a thymidine (T), cytosine (C), guanine (G) or adenine (A). At the site of a single nucleotide polymorphism (SNP), some individuals have one of these four while other individuals have another. SNPs are not mutations, but instead common variants found in at least a few percent of all humans. There are upwards of 300 million such variants in the human genome, accounting for a large amount of the genetic variation between individuals, but only 15 million or so are found in >1% of the population (17). During the early part of the 21st century, the International HapMap Project identified a large number of SNPs in the human population (18). This identification led to the ability to perform large scale genetic studies of the entire human genome.

**Linkage disequilibrium**

In his studies that form the basis of genetics, Mendel demonstrated that inherited phenotypes in pea plants segregated independently. That is, the genes were inherited separately. This is definitely the case for genes that reside on different chromosomes, but not necessarily the case for genes found on the same chromosome, which could be inherited together. However, genetic crossover during meiosis between sister chromosomes results in genetic variants residing on the same chromosome that frequently sort in an independent manner. However, the tendency for variants that are physically close to each to be inherited together is known as linkage disequilibrium. Regions of the genome that tend to be passed to offspring intact are referred to as haplotype blocks and genetic variation within haplotype blocks are in linkage disequilibrium.

**Imputation**

Haplotype blocks and the linkage disequilibrium of these genomic regions is used to advantage in genetic studies. Such studies do not need to study all SNPs and other variants. Instead, a single or a few SNPs within a haplotype block give information about inheritance and genetic association of other (not tested) SNPs within this same haplotype block. Thus, many more can be characterised than were actually studied by use of software that infers data for SNPs not studied but in strong linkage disequilibrium with ones that were directly genotyped (19).

**Genotype array**

A genotype or SNP array allows for determination of alleles for many (up to millions of) SNPs from a single individual in a single, simultaneous experiment. This technology takes advantage of fastidious DNA hybridisation and fluorescent microscopy. The array consists of immobilised allele-specific short sequences of DNA (oligonucleotides). Labelled, fragmented DNA from the test subject then is incubated with the array. The detection system identifies hybridisation at each of the oligonucleotides, which is read out as a specific allelic variant for each SNP.

**Genome-wide association study**

The first successful genome-wide association study (GWAS) was published in 2002 (20). Through 2017, this technology has been applied to the study of about 2000 different human diseases in >3000 studies, but the number changes rapidly. As mentioned above, the most common study design is case-control, but for a number of reasons, these studies require large numbers of subjects in both categories. The analyses of the data produced is perhaps deceptively simple. The size of the genetic effect of any given SNP is reported as the odds ratio—the odds of a diseased individual having a specific allele compared to the odds of a control individual having the same allele. Odds ratios above 1.0 indicated an associated allele. In addition, p-values are calculated by chi square or regression analyses. However, many thousands if not millions of tests have been performed; thus, the statistical analyses must take into account correction of multiple comparisons. Given the size of the human genome and the number of common variant SNPs, a p-value of 10^-8 (0.00000001) is the usual agreed upon level of statistical significance.

There are a number of limitations and problems of GWAS. Control groups that are not well matched to those with the disease create false positive results on the basis of population stratification. These problems may be subtle and difficult or nearly impossible to identify. For example, a cohort of patients derived from White Southern Americans who are largely descendants of English, Irish and Scottish immigrants may not match well to a control American population that includes those with Scandinavian or Southern European heritage, even though all are culturally identified as White Americans. Because of unanticipated population stratification and the large number of tests performed, many of the results of GWAS may be false positives, or at least extremely difficult to reproduce. Only genetic variation that is common in the population can be studied by GWAS. That is, alleles commonly identified as mutations (present in <1% of individuals) are not generally studied by this approach. The SNPs studied in GWAS are almost always a marker, and not the causative polymorphism. Instead, the genetically associated SNPs are in linkage disequilibrium with the polymorphism that supplies the pathophysiological impetus for the disease. Thus, fine mapping is needed. These studies may require very large numbers of subjects because hundreds of genetic variants may be highly correlated; that is, in tight linkage disequilibrium.

**Candidate gene studies**

Instead of investigating the entire genome without an a priori hypothesis,
candidate gene studies investigate a single or small number of polymorphisms in a gene hypothesised to be involved in the pathogenesis of the disease. These studies have the advantage of performing a small number of comparisons. Nonetheless, such studies are generally hampered by small numbers in the cohorts and failure to take into account the large number of allelic variations that might be tested. Thus, many such studies likely result in false positives as well as false negatives.

**MicroRNA**

Not much more than a generation ago, geneticists determined that only a small portion of the human, or any mammals, genome encoded proteins. These areas that did not encode protein were termed ‘junk DNA’. Not unsurprisingly, most of this DNA turns out not to be junk after all. Some DNA encodes for microRNA, which are short (~22 nucleotides) non-coding RNAs that are present in animals, plants and even viruses. These microRNAs bind specific regulatory sequences (3’ untranslated regions) of messenger RNA and by this binding regulate gene expression after transcription. MicroRNAs are an only recently recognised component of genetic variation within a species.

**Epigenetic (methylation) studies**

Expression of genes is silenced by methylation of specific nucleotide sequences found in the promoter regions upstream from the protein translation start site.

**Genetic studies in SS**

**Candidate gene studies**

There have been many candidate gene studies in SS with several reviews summarising these studies (21-23). Since the latest of these reviews, a number of studies have taken place (Table I). For the latest of these reviews, a number of studies have taken place (Table I). For the latest of these reviews, a number of studies have taken place (Table I). For the latest of these reviews, a number of studies have taken place (Table I). For the latest of these reviews, a number of studies have taken place (Table I).

<table>
<thead>
<tr>
<th>Gene (ref)</th>
<th>Allele/variant</th>
<th>Number studied</th>
<th>Ethnicity</th>
<th>Odds ratio / p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>TNFAIP3 (24)</td>
<td>F127C</td>
<td>327</td>
<td>Greek</td>
<td>2.4/0.04</td>
</tr>
<tr>
<td>MTHFR</td>
<td>(25) CG67T</td>
<td>356</td>
<td>Greek</td>
<td>2.13/0.03*</td>
</tr>
<tr>
<td>A1298C</td>
<td>356</td>
<td>600</td>
<td>Greek</td>
<td>0.31/0.01*</td>
</tr>
<tr>
<td>CD28 (26)</td>
<td>GC haplotype</td>
<td>111</td>
<td>138</td>
<td>Mexican</td>
</tr>
<tr>
<td>CTLA4 (26)</td>
<td>CAG</td>
<td>111</td>
<td>138</td>
<td>Mexican</td>
</tr>
<tr>
<td>CGA</td>
<td>111</td>
<td>138</td>
<td>Mexican</td>
<td>11.38/0.001</td>
</tr>
<tr>
<td>H19 (27)</td>
<td>rs3741219</td>
<td>800</td>
<td>350</td>
<td>Chinese</td>
</tr>
<tr>
<td>rs2839698</td>
<td>800</td>
<td>350</td>
<td>Chinese</td>
<td>not significant</td>
</tr>
<tr>
<td>HOTAIR</td>
<td>(27) rs920778</td>
<td>800</td>
<td>350</td>
<td>Chinese</td>
</tr>
<tr>
<td>IKZF1 (28)</td>
<td>rs4917129</td>
<td>665</td>
<td>863</td>
<td>Chinese</td>
</tr>
<tr>
<td>rs4917014</td>
<td>665</td>
<td>863</td>
<td>Chinese</td>
<td>0.76/p=3x10^-3</td>
</tr>
<tr>
<td>HIFA (66)</td>
<td>rs11549465</td>
<td>110</td>
<td>141</td>
<td>Mexican</td>
</tr>
<tr>
<td>ANKA (66)</td>
<td>rs1087595</td>
<td>110</td>
<td>141</td>
<td>Mexican</td>
</tr>
</tbody>
</table>

*for MALT lymphoma versus no MALT lymphoma.

of rs35593994 and rs3164696.

of rs5742990, rs231775, and rs3087243.

Recent candidate gene studies in Sjögren’s syndrome.

Table I. Recent candidate gene studies in Sjögren’s syndrome.

In general, these candidate gene findings are of interest, but because of the inherent limitations of such studies, must be viewed with a skeletal eye until confirmed in other and larger cohorts.

**GWAS**

GWAS has confirmed that the strongest genetic risk factor for SS lies on chromosome 6 in the major histocompatibility complex. However, there are important differences between the associations found in European-derived and Asian cohorts of patients (29). GWAS has been performed in European-derived (30) as well as Chinese Han (31) populations, but not in other ethnicities. Given that causal alleles have not been firmly established for any Sjögren’s genetic association, analyses of pathways and networks may be the most useful approach in considering these data (32). Table II compiles the data concerning pathways implicated by GWAS in SS. Risk of SS varies according to Human Lymphocyte Antigen (HLA) Class II alleles and is consistent with the genetic association found for this region of the genome (33). These data support a role in SS for antigen processing in immune cells, including B cells. Serological status influences genetic association at HLA Class II with anti-Ro/La-positive subjects to be much more likely to harbour the associated HLA alleles than those without these autoantibodies (34). However, while one might conclude that changes in antigen processing and presentation led to autoantibody production, other pathogenic mechanisms, such as increased mRNA expression, may be in play (33).

Another pathway identified in Sjögren’s GWAS is innate immunity, more specifically interferon. Increased expres-
sion of interferon regulated genes is found in peripheral blood cells of these patients and is correlated with the presence of anti-Ro/La (35). Genetic association has been found for multiple genes involved in or regulating the interferon response, including STAT4, IRF5, IL12A, and OAS1 (30, 36, 37). Further investigation of the effect at OAS1 has been performed (38). The risk allele leads to alternative splicing of the gene resulting in an expression of isoform that is unresponsive to type I interferon. Nonetheless, while these data certainly provide important information, the mechanism by which OAS1 genetic variants predispose to SS is incompletely understood.

Genes involved in acquired immunity have also been identified in the SS GWAS. In particular, regulation and activation pathways of B lymphocytes contain SS risk alleles (Table II). Again, while there is available information about the function of most of these genes, the specific relationship to SS aetiopathology is yet to be elucidated. For instance, the genes FAM167A and BLK contain risk alleles for the disease, are encoded in opposite directions from a common promoter region and have inversely correlated protein levels, at least in pulmonary tissue (39). What expression of protein in the lung has to do with SS is unknown. Nonetheless, the protein product of BLK is a major component of B cell signalling and activation; thus, is involved in B cell tolerance (40).

MicroRNA studies
Involvement of microRNAs in SS pathogenesis was recently reviewed (41), but there are few data. Expression of particular microRNAs may be altered either in the serum (42, 43) or the salivary glands (44) of SS patients. Autoimmune disease has been reported targeting microRNAs (45). Further, microRNA may target the genes of the Ro and La protein and change the expression of these canonical autoantigens of SS (46). There are no studies of the genetics of microRNAs among subjects with Sjögren’s syndrome. Thus, the study of microRNA in SS is a new and emerging field.

X chromosome studies
We have studied X chromosome aneuploidies among Sjögren’s syndrome subjects. These studies are not impeached by population stratification because the incidence of X chromosome aneuploidy is not known to be different in human subpopulations (47-54). Klinefelter’s syndrome (male 47,XXY) occurs in about 1 in 500 live male births, but even in the 1st world most of these men are undiagnosed (55). Among 136 men with SS, we found 4 (1 in 34 or 3.0%) had a karyotype 47,XXY (56). This was highly similar to our finding in SLE (57), although we have not found a 46,XX man with Sjögren’s as has been reported in SLE (58, 59). Using a population based survey of all patients diagnosed with Klinefelter’s syndrome, Seminog et al. also found excess SS among these men compared to matched controls (60). We also studied 47,XXX among women with SS and found this abnormality of X chromosome number statistically increased over controls (56). We interpret these data to mean that the number of X chromosomes, not phenotypic sex, explains at least in part the strong sex bias of the disease. We also found extremely rare, if not unprecedented, partial triplication of the distal Xp in two SS patients, one of which had a triple mosaic of 45,X/46,XX/47,XXX0 (61). Thus, the causative genetics of the X chromosome dose effect may lie within distal Xp.

The toll-like receptor 7 (TLR7) and CXorf21 genes both lie on distal Xp and both escape X inactivation in immune cells (62, 63). Over-expression of TLR7 in female immune cells induces changes in interferon production between male and female B cells, monocytes and dendritic cells (62). CXorf21 is only expressed in this same group of immune cells. Knockdown of CXorf21 disrupts TLR7 signalling with loss of production of interferon, IL-6 and TNF-α (unpublished data, Harris, Scolfield). Furthermore, the CXorf21 protein is involved in regulation of endolysosomal pH and its over-expression in these particular immune cells results in pH differences in this cellular compartment when comparing male and female cells (64). Endosomal TLR signalling as well as antigen processing are both highly sensitive to and dependent on lysosomal pH. Another recent study showed CXorf21 expression is regulated by type 1 interferon, and the excess expression in female cells is increased by immune challenge. Furthermore, the CXorf21 protein co-localises with
TLR7 (65). This study also identified a potential causative haplotype (65). Thus, CXorf21 is a candidate to mediate the X chromosome dose effect. Other X chromosome abnormalities such as acquired X monosity in peripheral blood cells and skewed X chromosome inactivation have not been studied in SS to our knowledge.

Future directions
One of the next obvious steps towards understanding genetic risk factors in SS is to expand the sample sizes for performing large scale GWAS. These studies are ongoing through a large international consortium, the Sjögren’s Genetics Network (SGENE). Expansion of study size is critical for increasing statistical power to map new genetic risk loci. This will also enable well powered studies aimed at understanding how genetics are related to various clinical subphenotypes, such as development of lymphoma, production of anti-Ro or anti-La, and fatigue. Detailed functional studies, typically comparing a variety of cellular functions among patient groups divided by risk genotype, are needed to understand the biological consequences of specific risk alleles. Long term goals include development of polygenic risk scores that will help identify patients at increased risk for SS diagnosis, poor prognosis and increased disease activity so that interventions can be implemented to disrupt the chronic, progressive immune-mediated damage that results in disease.

References
Genetics in Sjögren’s syndrome: where we are and where we go / VM. Harris et al.


45. GOURZI VC, KAPSOGEOUROU EK, KYRIAKIDES NC, TZIOUFAS AG: Study of microRNAs (miRNAs) that are predicted to target the autotigens Ro/SSA and La/SSB in primary Sjögren’s Syndrome. Clin Exp Immunol 2015; 182: 14-22.

46. LUE Y, JENTSCH JD, WANG C et al.: XXY mice exhibit gonadal and behavioral phenotypes similar to Klinefelter syndrome. Endocrinology 2005; 146: 4148-54.


53. ABRAMSKY L, CHAPPLE I: 47.XXY (Klinefelter syndrome) and 47.XYY: estimated rates of and indication for postnatal diagnosis with implications for prenatal counseling. Prenat Diagn 1997; 17: 363-8.


