The search for the genetic contribution to autoimmune thyroid disease: the never ending story?

Matthew J. Simmonds and Stephen C.L. Gough

Abstract
Unlocking the genetic contribution to autoimmune thyroid disease (AITD) will hold one of the keys to understanding disease pathogenesis and developing improved treatments. Significant increases in our understanding of the human genome combined with methodological advances in our ability to search for genetic variation have transformed the way in which we screen the genome for susceptibility loci. From early linkage analysis through to candidate gene studies and most recently genome-wide association screening, each methodology has revealed important insights into not just the heritability of AITD but also the best way of identifying disease causing DNA variants. This review will examine each of the different genome screening techniques, highlighting the successes and failures of each methodology and the lessons learnt which have helped inform the next phase of the disease-gene identification process. We will also look to see where we should be focusing our research efforts in the future.

Keywords: autoimmune thyroid disease; Graves’ disease; Hashimoto’s thyroiditis; linkage analysis; candidate gene studies; genome-wide association screens

HOW DO YOU SOLVE A PROBLEM LIKE AUTOIMMUNE THYROID DISEASE?
Unlocking the genetic contribution to complex diseases such as autoimmune thyroid disease (AITD), including Graves’ disease (GD) and Hashimoto’s thyroiditis (HT), is vital to our understanding of disease pathogenesis and the development of improved therapeutic interventions. Although the desire to identify disease causing DNA variants has not changed in over 50 years, the ways in which we go about screening for susceptibility loci has evolved (Figure 1 timeline of genome-wide screening technologies), with breakthroughs in our understanding of the human genome and improved genotyping technologies helping to inform future screening methodologies. We will focus on the successes and failures of genome-wide screening over the years, what they have taught us about AITD heritability and where to focus future efforts into finding the remainder of disease susceptibility loci.

ONCE UPON A TIME... LINKAGE STUDIES IN AITD
The first approach employed to screen the genome for the genetic contribution to AITD was that of genome-wide linkage analysis. Linkage analysis involved the use of a series of microsatellite markers (short segments of DNA anywhere from 1–6 bp repeated in tandem) spread across the genome looking for co-segregation with disease, in family based data sets. Once a region of linkage was detected a ‘reverse genetics’ approach was employed to work out which...
gene within the region was linked with AITD, before determining the functional consequences [1]. Linkage analysis had previously proved highly successful in locating rare variants with large effects responsible for monogenic diseases, where the disease is caused by one gene. It was hoped that these successes in monogenic disease could be replicated in complex disease such as AITD where disease was thought to be the result of many genetic factors.

The first genome linkage scan in AITD was performed in 1999 in 56 multiplex families (containing unaffected and GD or HT affected family members). Three areas of linkage to GD, designated GD-1, GD-2 and GD-3 on chromosomes 14q31, 20q11.2 and Xq21, respectively, two areas of linkage to HT, designated HT-1 and HT-2 on chromosome 13q32 and 12q22, respectively, and linkage to AITD in general, AITD-1 on chromosome 6p11, distinct from the previously associated HLA affects within the region, were reported [2]. Increasing the number of families screened to 102 rather than providing support for these regions, produced no support for GD-3 or HT-1, with only weak evidence for GD-2 found after removing a subset of 29 Italian families [3]. Some evidence for association of AITD-1 was reported however, the position of this association was put into doubt as a stronger association was detected ~30 cm away from the original signal [3]. At the same time, new areas of linkage were also detected including linkage of GD to chromosome 7q and linkage of chromosome 6p, 8q and 10q to AITD in general [3]. Lack of replication between data sets however raised concerns over the reproducibility and indeed the robustness of linkage studies. Independent studies in a small White Caucasian family cohort showed some support for GD-3, albeit with association now at Xp11 [4], and detected linkage to chromosome 18q21 [4, 5]. Results from the largest AITD genome linkage screen conducted in 1119 Caucasian AITD affected-sib-pairs although identifying three new regions of linkage to GD on chromosomes 18p11, 2q36 and 11p15, and suggestive linkage to GD and HT at several other regions, failed to replicate any previously detected linkage results [6]. More interestingly this screen also failed to detect association of the HLA region, CTLA-4 and PTPN22, which have been unequivocally associated with GD by candidate gene studies [7–9], suggesting that linkage may not be suitable for finding susceptibility loci for AITD.

Genome-wide linkage studies in Oriental AITD populations showed a similar story to that seen in
Caucasians. A genome-wide screen in 123 affected Japanese sib pairs revealed that chromosome 5q31-33 produced their strongest linkage to AITD in general, with evidence of strong linkage also reported to chromosome 8q23-24 in HT, similar to that detected in a previous Caucasian linkage screen to AITD [3, 10]. Disappointingly, no other previously detected regions showed any linkage to disease in this Japanese cohort [10]. Some support for linkage to chromosome 5q31 was obtained from a genome linkage screen of 54 Chinese Han multiplex families, which also reported linkage to chromosome 1p33 and 1q42 [11]. Beyond replication of chromosome 5q31 within two Oriental populations and potentially some evidence for linkage of chromosome 8q23-24 to AITD in general, no other regions of linkage have been consistently replicated [3, 10, 11].

Even for replicated regions of linkage, such as 5q31-q33, these regions cover many hundreds of possible candidate genes making it difficult, if not impossible, to localize individual susceptibility loci. To narrow down association, positional cloning techniques have been employed whereby additional microsatellite markers and single nucleotide polymorphisms (SNPs) have been used to try to localize the signal to a specific gene/s. Only after nearly 10 years and numerous failed attempts was a GD susceptibility locus, SCGB3A2, located within the 5q31 region [12, 13]. Additional screening of the thyroglobulin gene within the 8q24 region of linkage showed some association with GD, although further replication is still required. Whether either of these associations accounts for the linkage signals within these regions is still open to debate. Even for the established AITD genes, TSHR, located in the GD-1 region of linkage and CD40, located in the GD-2 region, it is not clear whether it is these genes driving the respective linkage signal [9].

Several technical issues were proposed to explain why linkage failed in AITD. When linkage was first undertaken no dense microsatellite maps of the human genome were available to inform linkage study design, affecting genome coverage and mapping resolution [14, 15]. In contrast to monogenic disease, where disease is the result of single rare variants of large effect size it became apparent that in complex diseases, such as AITD, we were looking for a series of genes with much smaller effect size. Many of the first linkage studies were conducted on a small number of families which can be significantly affected by population substructure within the data set. This was clearly evident with GD-2 which only showed association in a cohort of 102 families when 29 Italian families were removed [3]. Although positional cloning techniques were employed to try and narrow down linkage signals to a specific gene/s, many signals covered regions encompassing hundreds of genes making it difficult to further localize the association detected.

Lack of success of linkage studies has been seen for almost all complex diseases, suggesting that although successful for monogenic diseases, linkage analysis has major limitations in complex disease. Although linkage has not delivered novel susceptibility loci for AITD, some conclusions, into the genetic contribution to AITD, could be drawn:

(i) No large genetic effects were discovered for AITD, confirming that disease susceptibility was more likely to be due to several genes of smaller effect;
(ii) There are differences between the genetic contribution to Oriental and Caucasian subjects, suggesting that a different set of genes may explain disease susceptibility in different environments;
(iii) Greater understanding of the genome and new technological breakthroughs were required to perform comprehensive screening.

CANDIDATE GENE STUDIES TO THE RESCUE

After the failure of linkage to detect replicable susceptibility loci, candidate gene studies, which were used in parallel with linkage, were more heavily relied upon to detect AITD genetic effects. Candidate gene studies involve investigating specific genes, which are selected based on a prior hypothesis about their role in disease. They can be conducted in either family cohorts by comparing transmission of ‘disease’ alleles from parents to affected and un-affected offspring or in case–control cohorts of non-related individuals with disease and controls without disease to look for differences in allele frequency between cohorts.

With approximately 20 000–25 000 genes in the human genome, many of whose function is still unknown, where do you start looking? As disruption to the immune system is key to AITD onset, this provided an obvious starting point, with a role for several immune response genes subsequently
Table I: Genes shown to be associated with GD using candidate gene studies

<table>
<thead>
<tr>
<th>Associated gene</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consistently replicated using candidate gene studies alone</td>
<td></td>
</tr>
<tr>
<td>HLA class II</td>
<td>Ban et al. [56], Chen et al. [57], Simmonds et al. [58], Wongsurawat et al. [59]</td>
</tr>
<tr>
<td>CTLA-4</td>
<td>Furugaki et al. [60], Ueda et al. [61], Vaidya et al. [62], Yanagawa et al. [63]</td>
</tr>
<tr>
<td>PTPN22</td>
<td>Heward et al. [64], Smyth et al. [65], Velaga et al. [66]</td>
</tr>
<tr>
<td>TSHR</td>
<td>Brand et al. [67], Dechaire et al. [67], Hirata et al. [68]</td>
</tr>
<tr>
<td>IL-2RA/CD25</td>
<td>Brand et al. [69]</td>
</tr>
<tr>
<td>CD40</td>
<td>Simmonds et al. [23], Houston et al. [69], Jacobson et al. [70, 71], Tomer et al. [72]</td>
</tr>
<tr>
<td>SCGB3A2</td>
<td>Song et al. [13], Simmonds et al. [12]</td>
</tr>
<tr>
<td>Genes detected using candidate gene studies awaiting confirmation</td>
<td></td>
</tr>
<tr>
<td>FCRL3*</td>
<td>Kochi et al. [20], Simmonds et al. [21], Owen et al. [73]</td>
</tr>
<tr>
<td>HLA Class I*</td>
<td>Farid et al. [74], Grumet et al. [75], Huang et al. [76], Matser et al. [77], Simmonds et al. [78]</td>
</tr>
<tr>
<td>IFIH1*</td>
<td>Sutherland et al. [79]</td>
</tr>
<tr>
<td>ADRB2</td>
<td>Chu et al. [80]</td>
</tr>
<tr>
<td>DIO2</td>
<td>Chistiakov et al. [81]</td>
</tr>
<tr>
<td>FCGR8a</td>
<td>Yeomim et al. [82]</td>
</tr>
<tr>
<td>IL-23</td>
<td>Huber et al. [83]</td>
</tr>
<tr>
<td>IL-3</td>
<td>Chu et al. [84]</td>
</tr>
<tr>
<td>IRF1</td>
<td>Yang et al. [85]</td>
</tr>
<tr>
<td>LMP7</td>
<td>Heward et al. [86], Vives-Pi et al. [87]</td>
</tr>
<tr>
<td>LMP2</td>
<td>Heward et al. [86], Vives-Pi et al. [87]</td>
</tr>
<tr>
<td>PDCD1</td>
<td>Newby et al. [88]</td>
</tr>
<tr>
<td>TNFα</td>
<td>Nakkuuot et al. [89], Simmonds et al. [90]</td>
</tr>
<tr>
<td>TCR-β</td>
<td>Demaine et al. [91], Pickrell et al. [92], Zhang et al. [93]</td>
</tr>
<tr>
<td>TPO</td>
<td>Kotani et al. [94, 95], Ludgate et al. [96], Pirro et al. [97]</td>
</tr>
<tr>
<td>UGRP1</td>
<td>Yang et al. [85]</td>
</tr>
<tr>
<td>Inconsistent findings</td>
<td></td>
</tr>
<tr>
<td>FOXP3</td>
<td>Ban et al. [98], Owen et al. [99]</td>
</tr>
<tr>
<td>IL-13</td>
<td>Bednarzuk et al. [100], Hiromatsu et al. [101], Simmonds et al. [102]</td>
</tr>
<tr>
<td>IL-4</td>
<td>Heward et al. [103], Yang et al. [85]</td>
</tr>
<tr>
<td>IL-1RN</td>
<td>Blakemore et al. [104], Muhlberg et al. [105]</td>
</tr>
<tr>
<td>Tg</td>
<td>Collins et al. [106, 107], Tomer et al. [108, 109]</td>
</tr>
<tr>
<td>TNFβ</td>
<td>Nakkuuot et al. [99]</td>
</tr>
<tr>
<td>Vitamin D receptor</td>
<td>Ramos-Lopez et al. [113], Stefanic et al. [114]</td>
</tr>
</tbody>
</table>

This table shows the susceptibility loci detected for GD via candidate gene studies split into consistent findings, findings awaiting confirmation and inconsistent finding. Asterisk indicates the findings which were later confirmed by GWAS. Adapted from Zeitlin, et al. [18].

identified using this approach, including the HLA class II and class I region, CTLA-4, and PTPN22 (odds ratio (OR) $\geq 1.50$) (Table 1 GD associations detected by candidate gene studies) [7, 8].

Candidate gene studies were, however, not without their limitations. Most researchers employed the use of polymerase chain reaction-restriction fragment length polymorphism (RFLP-PCR) technology to type for specific DNA variants, often only looking at SNPs in coding regions and 3’- and 5’-ends of genes. These incomplete approaches lead to some genes being excluded as being associated with GD in early candidate gene studies. A good example of this is the TSHR, which was later shown to be strongly associated with GD, once a more comprehensive strategy was embarked upon [16, 17].

An appreciation for the need to screen all common variation within a gene combined with increases in our understanding of the linkage disequilibrium (LD) between SNPs enabled us to use specific tag SNPs to capture a series of SNPs in strong LD without having to individually screen each variant [18]. The replacement of PCR–RFLP with fluorescent based genotyping platforms also increased the speed at which a specific gene region could be screened. This strategy allowed us to generate evidence of association of CD25, FCRL3 and CD40 with GD (OR $\leq 1.30$) [19–23]. As time progressed it became obvious that the size of effect on the development of disease, or OR, for susceptibility loci now being detected was becoming smaller and smaller with gene effects of OR $\geq 1.50$ giving way to effects of OR $< 1.30$. This necessitated the collection of larger data sets to provide sufficient power to detect, and exclude, effects of this magnitude (Figure 1).

Although the usefulness of candidate gene studies to follow-up on associated gene regions was clear, it was questionable whether an efficient method of genome-wide screening it could ever deliver. This was because for every successful gene detected by this approach, tens or hundreds of other genes investigated failed to show replicable association with disease [18]. A novel genome-wide screening method was, therefore, required and it was this idea that lead to the development of genome-wide association studies (GWAS) in the mid-2000s that have revolutionized the field of complex disease.

THE NEW KID ON THE BLOCK—GWAS IN AITD

Before GWAS could be contemplated, the limitations encountered with both linkage and candidate gene studies, including, genome coverage, data set
size and high throughput genome-wide screening technology needed to be addressed.

Due to the results obtained from linkage and candidate gene studies, it was believed that complex disease heritability was due to numerous common variants [minor allele frequency (MAF) >5%], known as the ‘common variant, common disease’ hypothesis [14, 24]. The Hapmap project was set up with the aim of cataloguing all common variation in the human genome, including non-coding regions, in four different ethnic populations; Caucasians, Africans, Chinese and Japanese, to test this hypothesis [25, 26]. Three million common SNPs were identified, on which LD analysis was performed, to determine how many SNPs would need to be directly screened to capture all common variation in the human genome. It was calculated that >80% of the three million common variants could be captured by only 500,000 SNPs [27, 28]. Advances in screening methodologies also meant that specially designed SNP genotyping chips could be used to perform dense genotyping on 500,000 or more variants in a fast and relatively cost efficient manner [1]. Candidate gene studies had demonstrated that the size of gene effects likely to be detected for complex disease were low (OR <1.30) highlighting the requirement for large samples sizes (>2000 cases and controls) and replication data sets of equal or greater size to confirm such effects. As it was assumed that any variant of interest detected would be consistent across the whole population investigated, careful matching for geographical variation and ethnicity would also be required to reduce the chances of false positive results [14]. As GWAS of 500,000 SNPs has the potential to generate a large number of false positives, power calculations suggested that only variants producing \( P \leq 10^{-7} \) should be considered significant at a genome-wide level [29, 30].

Numerous \( \geq 500,000 \) SNP GWAS have been completed for several autoimmune disease. To date, however, no such screens have been completed in AITD. Smaller screens using many of the concepts of larger GWAS studies have been completed for GD (Figure 2). As part of the Wellcome Trust Case Control Consortium (WTCCC) our group investigated a genome-wide set of 14,436 non-synonymous (nsSNPs) comprising all experimentally validated nsSNPs with a MAF >1% in western European samples in a data set of 1000 UK Caucasian GD cases, as well as 1000 breast cancer, ankylosing spondylitis (AS) and multiple sclerosis (MS) cases and 1500 controls [31]. Due to the reduced number of tests being performed a \( P < 10^{-5} \) was considered to be significant at a genome-wide level [32]. Unsurprisingly the largest association detected with GD was the HLA region \( P < 10^{-20} \), with the maximal signal concentrated around previously detected GD HLA class I and class II associations [8, 31]. Due to such high \( P \)-values within this region even very distant SNPs in modest LD showed indirect evidence of association, making it difficult to determine additional HLA loci. Analysis of data from this region is ongoing to disentangle any additional associations.

None of the other nsSNPs reached a \( P < 10^{-5} \) but several variants showed association at \( P < 10^{-4} \), which just fell short of this cut off value. Interestingly association of a nsSNP in FCRL3 \( (P = 2.1 \times 10^{-4}) \), which has previously been proposed to be associated with GD in Oriental and Caucasian populations, was detected [20, 21, 31]. Association of this nsSNP was believed to be the result of LD with the previously screened FCRL3 associations [20, 21]. To determine if larger GD associations existed within FCRL3, tag SNP screening was performed using 7 tag SNPs in an extended Caucasian data set, which revealed a stronger association of another SNP within this region, rs11264798 \( (P = 1.6 \times 10^{-5}, \ OR = 1.22) \). Additional nsSNP associations were also found within neighboring FCRL5 \( (P = 1.3 \times 10^{-4} - 5.0 \times 10^{-5}) \), with tag SNP analysis of FCRL5 and logistic regression analysis with previously typed FCRL3 tag SNPs, revealing that any associations at either gene was due to LD with FCRL3 rs11264798 and rs10489678 SNPs [33]. Ten other novel GD associations were detected at \( P < 10^{-4} \) but as these regions showed no prior evidence of association the results were viewed with caution [31]. A replication study of these loci in an independent collection of 1578 UK GD patients and 1946 matched controls showed no evidence of association with GD, although when combined with the original WTCCC study group, minor associations \( (P \geq 10^{-3}) \) remained for four of the nsSNPs, present within HDLBP, TEKT1, JSRP1 and UTX. Tag SNP screening of these regions in a UK Caucasian GD collection of 2478 samples and 2690 controls, only found minor additional associations within HDLBP and TEKT1 \( (P = 0.042–0.002) \) [34]. With the exception of minor associations at HDLBP and TEKT1, the remainder of the novel loci detected failed to be
replicated, suggesting that these nsSNPs do not have a significant effect on the risk to GD in the UK [34]. A second analysis of the nsSNP GWAS data undertaken comparing the 1000 GD against all non-AITD diseases cases and controls (4500 samples), detected genome-wide association of a nsSNP within TSHR \( (P = 2.1 \times 10^{-5}) \) [31]. This provided further evidence for association of this region with GD, which has now been refined to two variants, rs179247 and rs12101255 \( (P \leq 10^{-7}, \text{OR} = 1.53 - 1.55) \) present within a 40 kb region of TSHR intron 1 [17].

The main limitation of the nsSNP GD screen was that of genome coverage. Two of the most well known GD associations, CTLA-4 and PTPN22, were not detected due to lack of representation on the genotyping chip [31]. Similarly for the other established GD susceptibility loci, CD40 and CD25 had only limited SNP coverage on this chip and SCGB3A2 had no coverage at all, explaining why these associated variants were not detected. This may also explain why other biologically plausible candidates were not picked up in this screen. This suggests that much of the remainder of the genetic contribution to GD may lie outside coding regions and may instead be due to common variation in non-coding regions as has been seen for many of the susceptibility loci detected to date. This serves to highlight the need for further larger GWAS in AITD encompassing common variation in non-coding regions. Whilst data from such screens is awaited, AITD has been shown to co-cluster with other autoimmune diseases, including T1D, MS and RA, suggesting the sharing of susceptibility loci between autoimmune diseases [35]. It has been well documented that although each autoimmune disease has a set of disease specific susceptibility loci, such as the TSHR in GD, they also...

**Figure 2:** Genes associated with GD detected either through candidate gene studies and/or GWAS. This figure shows a diagrammatical representation of the chromosomes within the human genome and on them highlighted the name and location of GD susceptibility loci confirmed by candidate gene studies alone, candidate gene studies and GWAS and loci found by GWAS in other autoimmune diseases, including MS and T1D and confirmed in GD by candidate gene studies.
share a series of common autoimmunity susceptibility loci, such as the HLA region, CTLA-4 and PTPN22 [7, 36]. With data now generated from several GWAS studies, including MS, AS, T1D, RA and the aforementioned nsSNP GWAS in GD, it is now possible to determine the similarities and differences in the genetic architecture between these diseases [37]. After analyzing a set of 15 SNPs associated with all five autoimmune diseases, the diseases were shown to cluster within two disease groups with MS and AITD having similar but distinct genetic profiles from RA and AS [37]. T1D also showed strong similarity to AITD [37]. This suggests that data obtained for other GWAS studies, in particular T1D and MS, may also provide potential insights into the aetiology of GD and could help explain why different autoimmune diseases co-occur in the same individual.

INSIGHTS FROM GWAS IN OTHER AUTOIMMUNE DISEASES

Due to the long established link between T1D and GD many of the T1D GWAS loci detected have also been studied in GD. Two ≥500 000 SNP screens have been performed for T1D [38, 39]. The first was performed on 550 000 SNPs in a cohort of 563 European T1Ds and 1146 controls and 483 family trios [39] and the second, as part of the WTCCC, investigated seven different complex disease cohorts of 2000 cases each and 3000 shared controls [38]. Interestingly both screens identified association of previously detected T1D loci, with the 550 000 SNP screen identifying association of several SNPs within the HLA region, PTPN22 and INS (P < 10−6) and the WTCCC detecting association of the HLA region and PTPN22 (P < 10−26) as well as providing some evidence of association of CTLA4 (P = 10−5), CD25 (P = 10−5) and IFIH1 (P = 10−5) [38, 39]. Both scans also detected novel susceptibility loci. The 550 000 SNP scan found strong replicable association of three variants within KIAA0350 on chromosome 16p13 (P = 6.7 × 10−11) [39]. The WTCCC also reported genome-wide significance for seven novel T1D signals (P < 5 × 10−7) throughout the genome including further independent replication of KIAA0350 and associations at chromosomal locations 12q13, 12q24, 4q27 12p13, 18p11 and 10p15 [38]. Replication of the T1D WTCCC GWAS and a previous 13 378 nsSNP GWAS performed on 3400 cases and 3300 controls, in an additional 4000 cases and 5000 controls and a family data set of 2997 parent child trios was performed [40]. Eleven SNPs showing association with T1D (P ≤ 5 × 10−7) within the 500 000 WTCCC screen were also screened, with four of these regions showing convincing replication (P < 1.15 × 10−14) including KIAA0350 on chromosome 16p13 and CD25 on 12q24, ERBB3 on 12q13, and PTPN2 on 18p11 (P = 1.14 × 10−14–1.52 × 10−20) [40]. Fourteen nsSNPs from the T1D nsSNP screen that showed a trend towards association (P > 1 × 10−3) were also followed up and in addition to confirming previously reported associations at known T1D susceptibility loci IFIH1 and PTPN22, associations were also detected for CD226 on chromosome 18q22 (P = 1.38 × 10−8) and some evidence for nsSNPs within CAPSL, C20orf168, IL7R and MED12L, although they did not reach genome-wide significance (P < 8.25 × 10−5) [40].

Thirteen of the associated replicated T1D loci were also screened in 2200 UK Caucasian GD and 3600 controls with some evidence of association of IFIH1, PTPN2, CAPSL, CD226 and loci at chromosome 4q27 and 2q11 (Figure 2). Additional meta-analysis of the T1D GWAS data are helping to identify numerous additional susceptibility loci that will in the future also be screened in GD. Like T1D, several MS GWAS have been performed and a series of susceptibility loci identified. Unlike T1D there has been less of a drive to replicate these effects in AITD, but some MS loci have previously been detected through GD candidate gene studies (Figure 2). Determining which susceptibility loci cross over between each autoimmune disease will provide important insights into disease genetic susceptibility. The WTCCC Immunochip project is currently underway, using a 200 000 SNP custom array to provide dense SNP mapping within all significant GWAS loci from a series of autoimmune disease, including GD, T1D, AS, MS, RA and systemic lupus erythematosus (SLE), and will provide previously unimaginable insights into the genetics of not just each disease itself but to the autoimmune disease process in general.

THE MORAL OF THE GWAS STORY IN AITD

Although complete GWAS screening has not been performed in AITD, insights into AITD genetic susceptibility have been achieved through the
nsSNP study and studies in other related autoimmune diseases. At this stage, therefore, we can conclude that:

(i) GWAS has not identified any large effects (OR > 1.40) not previously detected by candidate gene studies;
(ii) Newly detected associations are modest (OR < 1.30) and are likely to only contribute to a fraction of the genetic heritability;
(iii) Many GWAS associations, although better than linkage studies, do not survive replication, demonstrating the need to screen large replication cohorts;
(iv) Susceptibility loci detected for AITD are not all in coding gene regions, suggesting that focusing on coding SNPs may miss true disease causing DNA variants.

Even though GWAS signals have been defined to specific genes we cannot be certain that the gene named actually contains the aetiological variant/s attributable to these regions of association. As such determining the aetiological variants represented by these associations remains an important challenge.

THE SEQUEL: SEARCHING FOR THE ‘MISSING’ HERITABILITY

In T1D, GWAS has increased the number of confirmed susceptibility loci from between 5 and 6 to over 40 [41]. However, even with the large number of new susceptibility loci identified it is believed that only a proportion of T1D genetic heritability is accounted for. T1D is not unique in this regard, with similar findings being reported in most complex diseases [42]. It has previously been suggested, by some, that genetic heritability in complex disease had been overestimated and that the ‘missing’ heritability does not exist. Although new methods based on GWAS results are being developed to better estimate heritability, it remains the belief that current heritability measures are more or less accurate [42], which leaves us with the question of where is the ‘missing’ heritability? Whilst GWAS have screened the majority of common variation (MAF >5%) within the human genome, other types of variants may hold the key to finding the missing heritability, with other approaches required to identify these effects including:

Trans-ethnic mapping

GWAS in data sets of different ethnicity may detect regions of association that have only appeared weakly associated in a specific population but by exploiting the differences in SNP allele frequencies and LD patterns between different ethnic groups, may be easier to detect in another ethnic group [42, 43].

Copy number variants

Copy number variants (CNV) are structural variations within the genome including deletions, insertions, inversions and translocations [1]. Depending on their location CNVs may influence gene number and if located in a significant part of the gene or its regulatory region may also alter the gene product quantity. CNVs were not routinely captured during the first phase of GWAS. Improved cataloguing and CNV screening methodology have enabled more extensive screening of CNVs to be carried out [44]. Recently 3432 non-duplicated, multi class CNVs were screened in 2000 cases from eight common complex diseases, including T1D and RA and 3000 shared controls [45]. Three replicable CNV loci were identified including a CNV within the HLA region associated with T1D ($P = 8 \times 10^{-153}$, OR = 0.20) and RA ($P = 1.4 \times 10^{-39}$, OR = 1.77). However, no evidence could be found for association of this CNV independent of previously detected strongly associated common variants within this region [45]. Although not all known CNVs, in particular rare CNVs, could be captured in this study, the results suggest that common CNVs are unlikely to play a major role in disease either through a specific CNV having a large effect (OR > 1.30) or through numerous CNVs having a small effect (OR < 1.20) and are unlikely to explain the missing heritability [45].

Rare and low minor allele frequency variants

Low minor allele frequency variants (MAF = 0.5–5.0%) or rare variants (MAF < 0.5%) are common in the human genome and have been shown to influence complex diseases including autism, epilepsy and schizophrenia [42, 46]. It has been proposed that rare variants with large effects, similar to those previously found to explain many monogenic diseases
and which were originally discounted for complex
disease, may also play a role in complex disease
[42, 46, 47]. A lack of information on low minor
allele/rare variants within the genome and high
throughput technologies to screen such variants has
hampered previous efforts to look at the role of these
variants. Insights from the 1000 genomes project,
which was established with the aim of sequencing
1500 individuals of different ethnicity, is extending
our knowledge of human variation within the
genome from all known common variation down
to all low and rare variants in the genome with an
MAF $= 1–5\%$ (http://www.1000genomes.org). So
far the pilot study has sequenced 172 individuals at
low depth and identified 11 million new low and
rare variant SNPs [14]. Combined with the devel-
opment of high throughput next generation sequen-
cing tools able to screen 55 billion bases in 10 days,
the search for rare variants in complex disease is
becoming a real possibility [48]. Interestingly some
evidence for a role of rare variants has already been
found in the T1D susceptibility loci, IFHI1, where
four additional rare variants were found independent
of the common GWAS signal originally detected
[49]. This supports the idea that rare variants are
likely to occur in regions already associated with dis-
 ease and are proposed to either enhance/modulate
previous detected associations or through LD could be
accounting for the previous associations within the
region [42, 46, 47]. It has also been suggested
that linkage results in complex disease, which
showed great success at finding rare variants in
monogenic disease, may in fact be able to inform
the search for rare variants in complex disease [15].
While still very much in its infancy it will be
interesting to see the impact of rare variants upon
complex disease susceptibility.

**COULD GENE INTERACTION
EXPLAIN THE MISSING
HERITABILITY?**

Screening further variation in the genome could
produce important insights into locating the missing
heritability for complex disease. However, it has
been suggested that how susceptibility loci interact
with each other, the environment, regulatory ele-
ments within the genome and parent of origin/
trans-generational effects may also be accounting
for some of the missing heritability.

**Interaction with regulatory elements
within the genome**

Variation in how genes interact within the genome
and in particular how they are regulated will un-
doubtedly add another level of complexity to the
genetic architecture of disease. It has been suggested
that multiple common variants could affect protein–
protein or protein–DNA interactions, potentially
altering gene regulation and protein production
[47]. These effects are not as simple as one effect
causing a change in expression and will be more dif-

cult to access. It is envisaged that advances in our
ability to screen for non-coding DNA with regulat-
dory potential including sequence conservation,
histone modifications, DNAse I hypersensitivity and
cis-regulatory mutations, can, in the future, be com-
bined with GWAS data to uncover if interaction of
multiple variants will help explain disease [47, 50]. It is
also believed that this approach could inform fine
mapping of already identified genetic associations.
Combining fine mapping of a specific association
with information on non-coding DNA with regula-
dory potential within the region may help refine as-
sociation and identify the etiological variants present.

**Gene–gene and gene–environment
interaction**

Gene–gene interaction or epistasis has been proposed
to occur between genes that cluster within specific
immune pathways enhancing their effect on disease
susceptibility [47]. Some examples already exist in
GD including a possible interaction between HLA
class II molecules and thyroglobulin [51]. This
suggests that once additional GWAS have been
performed in AITD, gene–gene interaction studies
will be vital to determine the true association of dif-
f erent susceptibility loci with disease. Similarly, inter-
action of specific environmental factors with specific
genes has also been proposed to modulate or en-
hance their effect on disease but has been difficult
to take into account in GWAS. This is due, in part,
to relatively little being known about the environ-
mental factors linked to complex disease and the
difficulties associated with relying on patients retro-
spective re-call. Future advances will hopefully make
it easier to quantify these effects [47].

**Different genetic effects for different
clinical phenotypes**

The use of large data sets comprising more than 2000
cases and 2000 controls and in many cases data sets of
more than 10,000 cases and matched controls to perform GWAS and replication studies have afforded the power to detect a plethora of small effect common variants (OR = 1.10–1.20) in complex disease [38]. As complex diseases tend to consist of heterogeneous combinations of different sub-phenotypes, which can vary from patient to patient, one potential downside of using such large data sets is that they could be diluting down genetic effects which only predispose to a specific sub-phenotype [42]. Screening large data sets of ‘super’ cases, consisting of a data set enriched for a specific clinical sub-phenotype, will hopefully reveal whether any of the missing heritability is due to sub-phenotype effects.

**Trans-generational and parent of origin effects**

Studies on mice have suggested a role for trans-generational genetic effects in which phenotypic variation in one generation results from genetic variation in inherited epigenetic markers, such as small RNA, RNA binding proteins, micro RNA and DNA methylation, that are passed down through several generations [52]. These effects can alter genotype–phenotype correlations in offspring, suggesting that the genotype of individuals in previous generations is a better predictor of a disease phenotype than those in the present generation [47, 52]. It is unknown if trans-generational genetic memory occurs in humans. However, a recently identified variant in KCNQ1 on chromosome 11p15 has been shown to either confer or reduce risk to type 2 diabetes depending on whether the allele is inherited paternally or maternally, respectively [53]. These types of variants are difficult to detect, with large family studies required to make this possible. It has also been suggested that variation in the microbiota within the human gut, which is believed to interact with the immune system to help prevent infection, could also play a role in autoimmune disease onset [54, 55]. The microbiota is laid down during the first few years of life and can be influenced by the mother’s microbiota, passed on to the offspring through breast feeding [54, 55]. Such studies could be of particular interest in AITD where there is a strong female preponderance and where family studies have shown that AITD patients are more likely to have a female relative with AITD than a male relative [35]. Such studies may not only explain some of the missing heritability but also the strong female preponderance seen in AITD.

**CONCLUSION**

The different approaches employed to screen the genome for susceptibility loci have revealed key insights into the role played by genetic heritability in AITD. Greater understanding of the type and size of associations likely to be detected combined with increases in our understanding of and ability to screen the human genome has enabled the development of new and improved genome-wide screening methodologies. Although the most recent genome screening approach of GWAS is still in its infancy in AITD, some key insights have already been achieved in our understanding of the genetic susceptibility to AITD. Combined with replication of GWAS performed in other autoimmune diseases, several novel susceptibility loci have been identified for AITD. The challenge that now faces us is the need to refine association of the susceptibility loci already identified, including performing fine mapping of GWAS signals to determine the aetiological variants driving these effects, as well as undertaking additional screens to detect other possible associated variants. Lessons from other complex diseases inform us that the search for genetic heritability is far from over and that many potential avenues still need to be explored to help us fully understand the genetic contribution to AITD pathogenesis.

**FUNDING**

This work was supported by the Wellcome Trust grant 084702.

---

**Key Points**

- Linkage failed to detect susceptibility loci for complex disease such as AITD.
- Candidate gene studies successfully detected association of several AITD susceptibility loci, revealing associations with OR > 1.50 for the HLA region, CTLA-4, PTPN22 and TSHR, later giving way to effects of OR < 1.30 for CD40, FCRL3 and CD25. This suggests that much of the remainder of the genetic contribution to AITD is likely to be due to small effects of this order of magnitude.
- Knowledge from both linkage and candidate gene studies provided insights into disease susceptibility and has helped inform the design and implementation of genome wide screening.
- GWAS have revealed additional susceptibility loci through replication of gene effects found in other autoimmune diseases, such as T1D.
- Further screening for different types of DNA variants and methods to account for gene-gene/environmental interaction, genomic interaction and/or trans-generational/parent of origin effects will also need to be undertaken to determine the complete genetic contribution to AITD.
References

38. WTCCC. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007;447:661–78.


