Approaches and advances in the genetic causes of autoimmune disease and their implications

Jamie RJ Inshaw [1], Antony J Cutler[1], Oliver S Burren [2], M Irina Stefana[1], John A Todd [1]

[1] JDRF/Wellcome Diabetes and Inflammation Laboratory
Wellcome Centre for Human Genetics
Nuffield Department of Medicine
NIHR Oxford Biomedical Research Centre
University of Oxford
Roosevelt Drive
Oxford
OX3 7BN

[2] Department of Medicine
University of Cambridge
Addenbrooke's Hospital
Cambridge
CB2 OQQ

Summary

Genome-wide association studies are transformative in revealing the polygenetic basis of common diseases, with autoimmune diseases leading the charge. Despite the field being just over ten years old, advances in understanding the underlying mechanistic pathways of these conditions, which are the result of a dense multifactorial blend of genetic, developmental and environmental factors, have already been informative, including insights in therapeutic possibilities. Nevertheless, the challenge to identify the actual causal genes and pathways and their biological effects in altering disease risk for many identified susceptibility regions remains. It is this fundamental knowledge that will underpin a revolution in patient stratification, therapeutic target discovery and clinical trial design in the next 20 years. Here we outline recent advances in analytical and phenotyping approaches and the emergence of large cohorts with standardised gene expression data and other phenotypic data that are fuelling a bounty of discovery and improved understanding of human physiology.

Introduction

We are enjoying a transformative era of big data, large consortia and cohorts, landscape-changing phenotyping tools, including single-cell genomics, genome-wide association studies (GWAS) of numerous traits and diseases and a multitude of statistical and computational methods and approaches. Here we consider some aspects of analysis, causal gene discovery and highlight some exceptional recent advances, taken from the perspective of a classic autoimmune disease, type 1 diabetes (T1D). The field of the genetics of common multifactorial diseases is moving rapidly towards causal gene and pathway identification and away from worrying about 'missing heritability', which for T1D was a red herring since a polygenic model fits well, explaining most of familial clustering (in terms of sharing of disease diagnosis of siblings, $\lambda_{S,}$ which is the ratio of disease risk for an individual who has a sibling with the disease to the risk of the general population for the disease) in terms of

hundreds or thousands of associated regions across the genome [1]. Explaining λ_s took an exciting leap forward from analyses in families showing that untransmitted parental alleles can significantly impact the characteristics of the children and in determining the sharing of traits between siblings [2]. Exome and whole genome sequencing of vast numbers of patients and population cohorts will reveal very rare (< 0.1%) disease-associated variants that will help us pin down which genes are causal.

The importance of adequately large sample size

Increasing sample sizes from large consortia have enabled discovery of hundreds of chromosome regions associated with autoimmune diseases that were previously not possible [3,4] as well as reduced the overestimation of effect sizes, the so called winner's curse (Box 1) owing to greater statistical power. For example, in the latest rheumatoid arthritis (RA) genetic analysis, over 100,000 individuals were analysed, of which 29,880 were cases, identifying 101 regions associated with RA [2] and for inflammatory bowel disease (IBD), an analysis of 95,000 individuals (42,950 cases) identified 38 novel risk loci [5], bringing the total number of regions associated with IBD to 232. A smaller study in IBD of almost 60,000 individuals (25,042 cases) [6], but with a higher proportion genotyped using a genome-wide platform rather than the custom immune-disease ImmunoChip, identified further associated regions, taking the total number up to 240. GWAS is a triumph of statistical correction for multiple testing where reported regions must pass a p-value threshold of $<5x10^{-8}$, ensuring that the majority of declared regions are true effects. The same rigor is not yet applied to biology and omics/genome-wide phenotyping studies since we do not understand yet the complexity and extent of gene expression, epigenetic modification including DNA methylation, non-coding RNA expression, gene splicing, or post-translational protein modification, all of which could generate over 50,000 functional DNA elements and over 250,000 proteins. This huge statistical space is amplified even more by the proven but largely unexplored interactions between the human genome, its physiology and variation in microbial metabolism and immune reactivity, which could herald a new era of irreproducibility [7].

Each region might have more than one causal gene and variant, which could be a single nucleotide polymorphism (SNP), a repeat sequence or a structural variant such as an inversion. In T1D, a combination of GWAS [8] and use of ImmunoChip [9] in less than 10,000 cases has identified 58 associated regions (www.immunobase.org). We can see how increasing the number of cases in type 2 diabetes (T2D) from 1,924 (Reference[10]) to 74,124 (Reference [11]) has increased the number of associated regions from 3 to 243, so we can expect to uncover many new regions and explain a higher proportion of T1D genetic variance by genotyping more T1D cases. With GWAS data from population-based cohorts available now, controls are not limiting (although caution is recommended – see below). The index or lead variants in these regions can be combined in a genetic or polygenic risk score (GRS or PRS). The GRS can be used to identify those at highest risk of disease, which can be used to select children from the general population for prevention trials [12], distinguishing T1D diagnosed later in life from T2D [13], or to predict disease severity [14]. In the UK Biobank [15] for example, plasma and DNA samples from individuals selected for the very most extreme GRS for the most common diseases but not yet affected by the disease under analysis could identify early precursors of the disorder (or genetically-validated biomarkers).

Even the sizes of the largest disease GWAS are modest compared with population cohort resources such as the UK Biobank and the 23&me cohort. [16] [17], which will be crucial in establishing disease causality [18]. With hundreds of thousands of individuals in these collections, linking medical records with lifestyle with biomarkers with GWAS data, there is incredible scope to make many new discoveries, resulting in an avalanche of new results and preprints in BioRxiv (https://www.biorxiv.org) and the top journals [19]. For some diseases, such as T1D, these research returns will be tempered by a low disease prevalence and/or ascertainment biases exacerbated by disease severity.

However, with large sample sizes also comes a risk of detecting false positive associations; if not accounted for, seemingly small confounding effects can

make a huge difference. For example, the UK Biobank found that almost 150,000 individuals were related to at least third degree [20]. Using linear mixed models employing a genetic relatedness matrix as a random effect can be an effective way of analysing data from related individuals and should be considered in these circumstances [21]. Another consideration when using resources of this kind, particularly when attempting to identify disease-associated variants, is the ratio of cases to controls. Even 'common' autoimmune diseases are generally quite rare, compared to say, T2D, and there is likely to be a large imbalance in the number of controls to cases, which can lead to high type 1 error (false positive) rates[22]. Including only those controls that have been very precisely matched to cases based on a number of demographic and clinical characteristics would reduce heterogeneity and ensure a more balanced design. The multiple testing ogre raises its head again and must be accounted for appropriately when testing thousands of phenotypes in a genome-wide fashion.

Fine mapping candidate causal variants

Identifying the causal variant driving the disease association is complicated by linkage disequilibrium (LD). This correlation between nearby SNPs, which occurs as a result of SNPs being passed down from parent to offspring on the same haplotype, (see [23] for a more detailed explanation and example) makes it difficult to determine which SNP out of an LD block is the causal one. Large sample sizes are critical to reduce the number of DNA variants that could be causal within a region, referred to as the credible set. The more cases and controls analysed, the more rare recombinant chromosomes are analysed and they drive the resolution of the fine mapping. Ethnically diverse populations can be an advantage here because they have different patterns of LD or arrangements of alleles on individual chromosomes or haplotypes. Nevertheless, considerable care must be taken when meta-analysing diverse populations and several methods have recently been developed [5,24,25].

When access to individual-level genotype data is available, and there is still an urgent need to provide these data to allow the accurate assembly of haplotypes, and the population being studied is homogeneous with respect to

ancestry, a reasonable starting point to fine mapping is to use a simple stepwise linear regression approach. This was performed in the T1D ImmunoChip analysis, in which three regions were found to have two independent signals associated with T1D (near *INS*, *PTPN2* and *TYK2*), whilst two regions were found to have (at least) three independent signals (near *IL2RA* and *IFIH1*) [9]. However, there are two potential problems with this approach: the first is identifying a p-value threshold at which to stop the procedure, which is not directly comparable between studies as the p-value is heavily influenced by the power of the dataset. Bayesian approaches such as Probabilistic Identification of Causal SNPs (PICS) [26] can help overcome this problem by generating posterior probabilities of association that are comparable across datasets. The second challenge comes from instances where the index SNP could be tagging two disease-associated haplotypes; in such cases, the stepwise regression approach will identify the SNP that tags the two haplotypes as the lead or index variant, while the two risk haplotypes might not be detected, leading to a misleading number of signals detected in the region. For this reason, approaches that search for combinations of SNPs with the most evidence of association rather than assuming the index SNP is associated, are sometimes more appropriate. GUESSFM [27] is one such approach, and has provided evidence of two independent SNP associations at the *IL2RA* region associated with multiple sclerosis (MS), rs61839660 and rs41295055, whereas stepwise regression only detects one signal, rs2104286, the more commonly reported SNP for this region, which likely tags the two risk haplotypes, and therefore is not a primary candidate for a causal variant. In contrast, rs61839660 almost certainly is a causal variant altering expression of *IL2RA*, encoding the CD25 subunit of the IL-2 receptor [23,28], and altering the sensitivity of T regulatory cells (T_{reg} cells) to IL-2, which is required for their maintenance and function in preventing autoimmunity. When only summary association statistics and not raw data are available, a number of methods, including CAVIARBF [29], FINEMAP [30] and JAM [31] have been developed to fine map regions although there remain challenges in these approaches [32].

Pleiotropy: genetic variation can affect multiple diseases

Since autoimmune and other immune diseases share many common variants, referred to as pleiotropy, several studies have combined data from multiple diseases in order to increase power to detect novel loci. GWAS joint meta-analysis of asthma, hay fever and eczema in 180,129 cases and 180,709 controls as allergic diseases that share a common genetic aetiology identified 99 regions (73 were novel, and only 4.4% were specific to one of the three diseases) [17]. Notably, the study named six candidate causal genes, including *CCR7*, that indicated possible repositioning of existing drugs for other diseases for which the effect on gene expression of the allergy-protective allele and existing drug matched. This pleiotropy extends to T1D and several of their candidate genes are shared, including the *CLEC16A-DEXI* gene region, *ERBB3, IL2RA, BACH2, IL7RA, FASLG, SH2B3, TNFAIP3* and *PTPRK*. The *PTPRK-THEMIS* region, for example, was recently associated specifically with T1D diagnosed under age 5 years and has a function in the thymus [33], in T cell development early in life during the production of recent thymic emigrants [34]. Once again, however, this region shows the now commonly-expected complexity with an association in MS with an index SNP, rs802719 [35] that is not in LD with the lead T1D age-at-diagnosis SNP, rs72975913, notwithstanding the results from T1D indicating more than one association signal in the region [33]. The functional missense SNP in the *IL6R* gene was also shared but in the opposite allelic direction, which illustrates yet another valuable outcome of GWAS: the *IL6R* allele that protects from RA, T1D and cardiovascular disease predisposes to allergic disease [36], indicating a possible contraindication for the use of existing anti-IL-6R antagonist drugs in certain patients (e.g. children with a very high GRS for allergic disease). Another example of opposite associations can be seen at *IL7RA* between autoimmune disease and allergic disease, and relevant phenotypes such as eosinophil count, is illustrated in UK Biobank using the GWAS results from https://biobankengine.stanford.edu versus its association with a number of autoimmune diseases, including MS [37]: the minor C allele of rs6897932 predisposes to autoimmunity and protects from allergic disease.

Typically, only 22/132 (16.7%) of index or sentinel SNPs associated with allergic diseases were coding [17]. The importance of epigenetic variation,

either associated with a disease-associated variant or independently, was highlighted by their finding that for 36 candidate genes CpG methylation (associated with less promoter activity) was found to influence transcription independently of genetic effects. We note, however, that methylation may not be the primary event in the altered transcriptional regulation but instead it may be a consequence of transcription factor binding, where methylation has a role in the maintenance of a transcriptional state [38].

Mendelian Randomisation

Mendelian Randomisation (MR) studies make use of information from different GWAS studies and combine them to draw conclusions about causality of one trait on another (Box 2). With so many publically available summary statistics available for a wide range of potential intermediate phenotypes and diseases (https://www.ebi.ac.uk/gwas/, http://www.gwascentral.org) it is possible to perform MR studies for many different intermediate traits on a multitude of different diseases. An example of MR analyses applied to an autoimmune disease is the link between vitamin D and MS, with two separate publications demonstrating that decreased vitamin D levels are causally associated with increased risk of MS [39] [40], thus providing evidence to support the epidemiological observation that low vitamin D levels may increase MS susceptibility.

However, MR uses three central assumptions, which must all be considered carefully before undertaking MR studies [41]. Firstly, that the SNP(s) of interest are associated with the exposure variable (e.g. vitamin D levels). This assumption is usually not violated and often a reason to conduct the MR study in the first place. Secondly, that the SNP(s) of interest are not associated with any confounding variables. This assumption has been put under threat by the finding that untransmitted alleles in families can impact characteristics and phenotype of the children, such as educational attainment [2], a point echoed in [42]. In addition, we now know that methylation can alter the function of a SNP [43], which is another way in which a SNP under consideration could be associated with a confounding variable and violate the second assumption. Finally, the third assumption is that the SNP(s) of interest impact the outcome (e.g. MS) only through the exposure variable (e.g. vitamin D levels). This is

where pleiotropy can become problematic, since the SNP(s) under consideration could be impacting a number of traits that are, in turn, impacting disease risk. There are two types of pleiotropy, vertical and horizontal. Horizontal pleiotropy, when the SNP(s) impact two or more separate traits, is the kind that can violate MR assumptions. Vertical pleiotropy refers to situations when the variant affects one trait and this results in changes to other traits. An additional concern when using MR is it has been shown that if there is measurement error in the phenotype measurement, the effect direction from MR analyses can appear the opposite from what is the truly the case, though methods have been developed to counter this potential problem [44]. Caution should therefore be taken in undertaking MR studies [45] [46] and there should always be very careful consideration that the assumptions of MR are not violated when using the approach.

Decoding the non-coding genome

Since the majority of disease-associated SNPs identified by GWAS are located in non-coding regions (which includes non-coding RNAs), the challenge remains to identify the target gene(s) or non-coding RNA, tissue specificity and mechanisms via which causal variants drive disease susceptibility, even when disease association has been narrowed down to a single variant in a region [28]. Technological advances in DNA and RNA sequencing and analytics combined with novel methods have provided the platform for large efforts such as ENCODE [47], BLUEPRINT [48], ROADMAP [49] and FANTOM [50] consortia. Functional annotation of the genome, including but not limited to, mapping of histone post-translational modifications, transcription factor binding sites, CpG methylation and areas of open chromatin and the use of massively parallel reporter assays has revolutionised our ability to understand the 'grammar' of the non-coding genome and the mechanisms underpinning variation in gene expression [51,52]. Landmark studies detailing genome-wide screens of open regulatory areas of chromatin, DNaseI hypersensitivity sites (DHS), provided evidence that common variants could alter chromatin accessibility and local gene expression [53] and that autoimmune disease-associated SNPs were enriched in DHS's [54]. Subsequently an enrichment of autoimmune disease-associated

SNPs was detailed in enhancers and clusters of enhancer elements, or 'super-enhancers', in T cells [55] [56]. The enrichment of disease-associated SNPs in enhancers, key regulatory elements that govern the lineage and functional state of cells through tissue-specific and temporal control of gene transcription, provided the initial biological framework linking common disease-associated SNPs, enhancers and regulation of gene expression. Methods have been developed to integrate functional annotation data with GWAS summary statistics to prioritise variants that lie in annotations that are enriched for disease hits in disease-relevant cell types [26] [57] [58] [59]. The application of this information to understand the genetic etiology and cell specificity of autoimmune disease association have been reviewed elsewhere [60]. These methods are effective to guide the most likely variants on average, but also require cautious interpretation since not all disease-associated SNPs will lie in the same type of annotation, the same cell type or the same cellular activation or differentiation state (Fig. 1).

Once regions have been identified as disease-associated, follow-up studies are launched to identify the mechanism through which the variant is altering disease susceptibility. It is common for analyses of expression of quantitative trait loci (eQTL) studies to be carried out, using many fewer individuals than in GWAS, but ideally from purified relevant primary cell populations. The latter is challenging and hence eQTL studies have, to date, been performed on a limited number of purified cell types, most commonly the common blood cells, $CD4^+$ and $CD8^+$ T cells and monocytes [61]. Thus, the chance of identifying a candidate causal gene from a GWAS overlapping with an eQTL depends to varying degrees on: (a) the appropriate cell type and its state of activation being measured, which could be a population of cells in a mixture of cells e.g. B cells in peripheral blood mononuclear cells, and (b) having sufficient individuals in the study to detect the difference (Box 3). Nevertheless, large resources such as the Genotype-Tissue expression (GTEx) project (https://www.gtexportal.org/home/), whole blood analyses [https://molgenis58.target.rug.nl/biosqtlbrowser/] and Immune Variation (ImmVar) [62] have been and will be extremely useful in generating and collating eQTL data for multiple tissues and under multiple conditions, but

there are still many different cell types under many different conditions that can be studied. For example, a particular variant associated with an autoimmune disease might increase disease susceptibility in activated $CD4^+$ T cells but not in non-activated $CD4^+$ T cells (Fig. 1). Furthermore, occasionally results may differ between studies in terms of the direction of effect of the eQTL, such as results for the T1D candidate gene *UBASH3A*, with differences arising from having very low amounts of mRNA in a particular cell type when a complex mixture of cells is analysed (such as T cell-specific gene in whole blood) [63]. It is apparent that establishing the causal tissue or cell type is an important step before declaring the candidate causal genes based solely on eQTL or allele-specific expression studies [64]; the eQTL must be in the cell type that is altering disease susceptibility in that region. The recent ability to examine gene expression in single cells is also going to significantly boost allele-specific expression analyses as well as help define the heterogeneity and functions of cell types within tissues [65] [66]. Recent developments in proteomic technologies now offer the ability to measure the levels of thousands of intracellular and plasma proteins [67,68], providing an emerging wealth of genetic associations with protein-abundance and epitope availability, protein QTLs [67]. Another other important aspect of fine-mapping and causal gene/variant identification is the integration of expression data or functional readouts associated with the credible SNPs, an approach to used to help dissect the *IL2RA* region [27]. Underpinning this integration lies the ability to obtain reliable genotype-to-phenotype data from specific cell types and this goal is greatly enhanced by the availability of large bioresource cohorts of genotyped volunteers willing to recalled for further study based on their risk genotypes and haplotypes [69] [70].

While the decrease in sample size seen in functional studies is a necessary compromise for feasibility, studies would benefit greatly from carrying out pre-study power calculations for the main outcome of interest based on small pilot studies, thus advancing one step further on the road to tackle the widespread crisis of reproducibility in experimental biology (Box 3).

Physical linking of credible risk variant candidates to candidate causal genes

The development of chromosome conformation capture (3C) techniques to examine the 3D structure of chromatin structure [71] has allowed the linkage of regulatory elements such as enhancers containing disease-associated SNPs to their target genes. A seminal study investigating obesity-associated variants in the *FTO* region on chromosome 16q12.2 [72] and application of the targeted 3C technique to the T1D region 16p13.13 [73] identified the genes *IRX3* and *DEXI*, respectively, as candidate causal genes alongside *FTO* and *CLEC16A*, previously considered as the candidate genes simply because they were nearest or contained the most disease-associated SNPs. The application of 3C therefore challenges the practice of nominating candidacy to the closest or most biologically relevant candidate in a disease-associated region. Iterative improvements in genome-wide chromosome conformation map resolution using Hi-C [74] [75] have enabled resolution of chromatin interactions to 750 bp [76] and have provided insight into the principles of chromatin organisation and control of gene expression. However, application of Hi-C to link regulatory DNA elements containing disease-associated SNPs to their targets is confounded by the onerous number of cells and billions of sequencing reads per sample to attain the high-resolution maps required for interpretation. Alternative genome-wide methods have therefore been developed and utilised to identify candidate genes using a targeting approach such as capture Hi-C [76], promoter capture Hi-C (PCHi-C) [77] [78], capture-C [79] [80] ChIA-PET [81] and HiChIP [82]. The network of interactions in each cell type confirmed the role of distal enhancers in transcriptional control of gene promoters in determining hematopoietic cellular identity [77], $CD4^+$ T cell activation [78] and differentiation state [82]. Confirming enrichments of disease-associated SNPs in T cell enhancers, T1D-associated SNPs were enriched in promoter-interacting regions in T cells [77], activation-induced interactions in $CD4^+$ T cells [78] and in enhancer interactions in T_H17 and T_{reg} cell subsets [82]. The complexity of gene regulation was highlighted in each experimental approach where, for example, enhancers can interact with multiple gene promoters, ‘skip’ multiple gene promoters and switch target promoters upon activation or differentiation. Integration of the T1D genome-wide analysis of

both the fine mapping ImmunoChip [9] and imputed GWAS [8] datasets with the PCHi-C datasets [77] [78] demonstrated the utility of linking promoters with their regulatory partners in extending our understanding of the genes involved in T1D. Novel candidate genes were prioritised using this approach in 29 T1D regions, in addition to those already nominated as candidates (Table 1). Genes in four regions that previously had no named candidate gene (e.g. 22q12.2 and *LIF*) and multiple genes with good candidacy in regions with an established candidate gene were prioritised (e.g. 1q32.1 [78]). New plausible candidate genes in regions where the biological relevance of the incumbent gene to the etiology of T1D was not appreciated are also identified [9] (e.g. 10q23.31; new candidate gene *PTEN*). In-depth exploration of each region and different experimental approaches [78] [23] [83] [84] will be required to dissect regions where multiple genes have been prioritised in different cell subsets, differentiation and/or activation states and different cell types [84].

One alternative approach, for example, that does not require chromosome conformation information, used a statistical model involving MR and information about open chromatin and gene expression from 100 genotyped individuals. This study identified over 15,000 putatively causal interactions between distal regions of open chromatin and over 60% of these interactions were over distances of less than 20 kb. Because the authors could infer the direction of causal interactions, the model also significantly improved the ability to fine map: when applied to an eQTL data set, the number of variants in the 90% credible set size was reduced by half [85].

Infection, microbiome and genes

Environmental exposures such as infection have been proposed to elicit the development of autoimmune disease in at-risk individuals [55] [86] [87] [88] [89] [90] and therefore it is of interest to compare GWAS results from infectious disease. Selective pressure by infectious diseases has driven the development of high inter-individual variability in immune genes, especially in the highly polymorphic human major histocompatibility complex (MHC) region [91]. HLA class II and I genes in the MHC region have by far the greatest genetic effect on the risk of developing T1D [92] [93] and, as one might expect, the HLA region

is associated with susceptibility to common infection [87]. In contrast to the GWAS bonanza in autoimmune disease, the GWAS approach until recently has yielded few susceptibility regions in infectious disease. However, the very large data-rich population cohorts[15] [94] [95] are transforming the ability to gain insight into the overlap between the genetics of common infection and autoimmune disease. For example, a number of variants associated with susceptibility to common infection overlapped with T1D disease candidate causal gene associations (*HLA, FUT2, SH2B3*) and candidate genes (*IKZF1, SBK1*) [16].

Notably, the FUT2 gene encodes the enzyme, galactoside 2-L-fucosyltransferase, which mediates the transfer of fucose to the terminal galactose on glycan chains of cell surface glycoproteins and glycolipids [96]. FUT2 creates a soluble precursor oligosaccharide FuC-alpha ((1,2)Galbeta-) called the H antigen which is an essential substrate for the final step in the soluble ABO blood group antigen synthesis pathway [96]. The expression of these histo-blood group antigens on mucosal glycans can serve as nutritional sources, receptors and attachment sites for microorganisms, parasites and viruses thereby playing a key role in host-microbe interactions [97,98] and in microbiome composition [99–103]. Twenty per cent of Europeans carry a null allele of a SNP (W134X; rs601338 G>A) in the FUT2 gene, and its presence therefore causes deficiencies in expression of these antigens and microbiota composition and function. The null allele is associated with increased risk of T1D [104] and IBD [105], mumps and several other autoimmune and infectious diseases [16] [106] but is protective against norovirus infection[107] and childhood ear infections[16].

The content and function of the microbiota can alter the low molecular weight metabolites present in human blood [108]. Some bacterial metabolites, for example short chain fatty acids (SCFAs), have widespread and important effects on many aspects of host biology, including immune function, inflammation and risk of autoimmunity, e.g. increased levels of SCFAs protect from autoimmune diabetes in the NOD mouse model [109] [110] [111] [112] [95]. Decreased microbial diversity has been observed in individuals with T1D [113]

[114] and alterations in the gut microbiota were observed in a small cohort of seropositive children who progressed to overt T1D [115]. In line with this, changes in the gut microbiome composition, known as dysbiosis, accompanied by a decline in blood and fecal SCFAs have been reported to occur before the diagnosis of T1D [116 117 118]. Another study has linked the IL-2 pathway, which has a major role in both murine and human autoimmune diabetes, to alterations in murine and human gut microbiota [119]. Together with the loss in microbiome diversity caused by lifestyle changes associated with industrialisation, these phenomena are thought to contribute to the continuous rise in T1D incidence [116 117 118 120 121]. This area is incredibly complex with environmental factors dominating [7], but one approach will be to identify which genes and their variants, for example, *FUT2*, alter microbial-derived or -modified circulating and intracellular metabolites in cases, controls and large-scale population cohorts.

While functional validation of candidate variants and genes has typically been challenging, the toolbox for functional genomics studies has expanded greatly in recent years with the addition of CRISPR/Cas9 methodologies, which now facilitate the study of endogenous gene function in physiological contexts and, even, in vivo. Of note, CRISPR/Cas9 techniques do not only enable fine genome editing, and, thus, introducing a variant of interest in a controlled genetic background or correcting specific mutations in patient-derived cells, but also make possible increasing or downregulating the expression of any gene of interest, within physiological parameters, by targeting a dead Cas9 fused to either a transcriptional activator (CRISPRa) or a repressor (CRISPRi) to the relevant promoters or enhancers. Similar techniques have also been employed to identify and validate which disease-associated SNPs fall within bona fide, rather than predicted, regulatory regions [83].

Unchartered regions

Even after the chromosome conformation analyses, almost 20% of the 57 T1D regions in ImmunoBase have no named candidate gene, often because the genes in the region do not have known roles in the immune system, owing to the classical view that T1D is caused by the autoimmune destruction of

pancreatic islet β cells. One possibility is that the causal gene(s) in these regions are not acting in the immune system but in the target tissue, the β cells [122] or in both, as might be the case for the gene *TYK2* [123,124].The discovery that *GLIS3* is both a T1D and T2D causal gene, encodes a transcription factor that regulates many β-cell expressed genes, including the insulin gene, and influences β-cell apoptosis [125,126] support this possibility: that of the target tissue as an active driver of disease [127].

One T1D-associated region, 17q21.31, also associated with liver autoimmune disease, primary biliary cirrhosis (PBC), is an example of a region with no immediately obvious candidate causal gene, at least in the context of the immune destruction model, and illustrates the challenges facing researchers aiming to identify the causal genes [9] [128]. The T1D index SNP (rs1052553 A>G) lies within exon 9 of the *MAPT* gene – which encodes the tau protein – and is a perfect tag of the two main haplotypes in the region [129] [130] [131], H1 and H2, the latter of which is protective for Parkinson's disease and other tauopathies [107]. H2 is also protective for T1D and PBC. There are eight protein-coding genes in this region, which comprises of a megabase-long ancient inversion polymorphism and several copy number variants, and with over 3,100 SNPs in near perfect LD. Any one or more of the eight genes (or non-coding RNAs or transcripts) could be causal for T1D and PBC. However, in tauopathies, owing to the fact that rare, highly disease-penetrant mutations of *MAPT* have been identified, it is certain that *MAPT* is a causal disease gene in the region. The identities of the common causal variants in neurodegenerative disease remain uncertain not least due to the high LD and the huge number of differences in the sequences of the two haplotypes [132]. eQTL analysis is also not very informative: not only are several of the genes in the wider *MAPT* region expressed in islets, but recent analyses of human islets revealed that several SNPs in the rs1052553 LD block act as islet eQTLs to increase expression of *MAPT*, *MAPT-AS1*, *NSF*, *CRHR1* and *KANSL1*, and decrease expression of *PLEKHM1* and *ARL17A* [133] [134]. Some of these lie in putative enhancer regions (rs55649944, rs111794853, rs2732650) or predicted active transcription start sites (rs58879558) in islets. Of note, the activation of *CRHR1* (also known as *CRFR1*) has been proposed

to promote β-cell proliferation, potentiate glucose-stimulated insulin secretion and protect against cytokine-induced β-cell death [135] [136] [137].

Nevertheless, MAPT could be causal, and therefore, then as a first step, one asks if the protein it encodes, Tau, and its pathogenic form, hyperphosphorylated Tau, are expressed in pancreatic β cells. Two reports in 2010 suggest it is expressed in human islets and in a rat β-cell line [138,139], supported by evidence in the Human Protein Atlas (https://www.proteinatlas.org/), results from a recent transgenic mouse model [140] and we have obtained evidence for detectable Tau protein expression in human β cells and not islet α cells (unpublished results). Despite originating from different germ layers, many similarities exist between the cell biology and developmental programs of pancreatic β cells and neurons [141], including the GLIS3 protein [142]. In addition, both cell types are considered post-mitotic cells with a very limited to no regenerative capacity, which renders them both vulnerable to the effects of misfolded, aggregated proteins and ER stress [143,144]. We propose that common cellular mechanisms underlie the fragility of β cells in T1D and in T2D and of dopaminergic neurons in Parkinson's disease.

High-density interaction maps for the 17q21.31 region and its neighbourhood used in conjunction with information on chromatin states specifically derived from pancreatic islets, if not pure β-cell populations, will greatly assist efforts of pinpointing the causal variants and genes driving the T1D associations. However, caution should be exercised in excluding variants or genes on the basis of information derived from non-diabetic islets or from islets carrying the benign allele. Stresses, similar to those encountered in the pre-diabetic and diabetic pancreas can alter chromatin states dramatically, such that SNPs located within quiescent regions in the absence of disease might map to active regions in tissues from diseased individuals.

Concluding remarks

Understanding the genetic basis of the biological pathways and processes underlying the etiology of autoimmune diseases has the potential to improve the success rate and safety of drugs [145] and has raised the possibility of repositioning and repurposing of approved drugs [146] for example, recombinant IL-2 (aldesleukin) [147 148 149 150]. As data from an increasing number of sources becomes widely available, so the opportunity to integrate these data together becomes possible. We are now able to examine each autoimmune disease related region in the genome and fine map the region to obtain a small number of SNPs most likely causing the increase in disease susceptibility; to identify the most likely cell type driving this association by examining which of these SNPs lie in functional regions of which cell type; and to highlight the most likely genes the SNPs are regulating through eQTL and chromatin contact experiments. This will not only help our understanding of autoimmune diseases but can point to potential therapeutic pathways for drug development. In parallel very significant efforts in defining the functions of the human immune system [151,152] and pancreatic islets [153] in health and disease will converge and greatly accelerate our understanding not only of the primary causes of common diseases but also our ability to stratify patients recruited into drug trials and explain why some patients respond and others do not or relapse [154,155].

Acknowledgements

We thank the JDRF (grant codes 9-2011-253 and 5-SRA-2015-130-A-N) and Wellcome (grant codes 091157 and 107212). O.S.B. is funded by Wellcome (grant code WT107881).

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Box 1: Winner's Curse

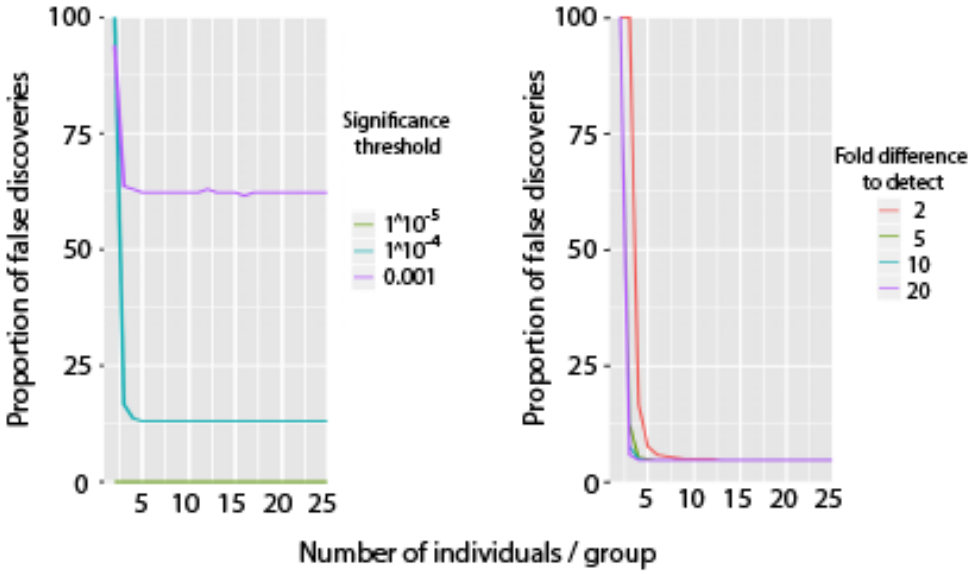
There have been many examples of genetic associations that have failed to replicate in subsequent studies. The reason for this is usually due to a lack of statistical power, which leads to inflated estimated effect sizes of truly associated SNPs as well as false positive associations. Inflation in effect sizes in underpowered analyses of SNPs occurs since studies will typically report SNPs associated below a particular p-value threshold. If the analysis is underpowered then the effect size at the declared associated SNP will need to be higher by chance than its true value in order to compensate for the lack of power. This distortion in effect sizes is termed the winners curse [156], since the first time an association is detected it is likely to have a larger estimated effect than the actual value. However, if replicated in an appropriately powered study, the estimated effect size is likely to better reflect the real effect of the variant.

Box 2: Mendelian Randomisation

Mendelian randomisation (MR) aims to deduce the causality of an intermediate phenotype (e.g. vitamin D levels) on an outcome (e.g. MS) [39,40] by using the SNPs associated with the intermediate phenotype, referred to as instrumental variables, to obtain predicted levels by genotype. These predicted levels of the intermediate phenotype should be independent of any confounding factors, since confounding effects should, by chance, be evenly distributed between the genotype groups, given the random allocation of genotypes from parents to offspring. These predicted levels of the intermediate phenotype are then regressed against the outcome of interest in order to assess the association between the genetically controlled variation in the intermediate phenotype and the outcome. The procedure is analogous to a randomised control trial: for example, if two groups were randomised to either high doses of vitamin D or placebo. In a simple MR setting this could be the major and minor homozygotes at a SNP that is known to affect vitamin D levels. In the clinical trial setting, randomisation would occur in order to minimise the chances of observing spurious associations due to confounding factors; in the MR setting, this randomisation has already occurred at birth.

618 Outlined in the main text are the fundamental assumptions in MR that are
619 required and need to be considered carefully. Another example of a
620 successful MR study is linking branched chain amino acids to T2D [157].

621 **Box 3: Sample size calculations and statistical power**



622 Sample size calculations are essential in answering almost all scientific
623 questions in a robust manner. We illustrate this with an example for a
624 microarray sample size calculation method [158], using data from ArrayExpress
625 (access number E-MTAB-4852). The figure left hand panel shows the median
626 proportion of false discoveries after 1,000 simulations at each sample size
627 and each p-value threshold for significance. The method assigns groups
628 randomly to each individual and sets the effect size for the truly associated
629 transcripts; from here we can work out the proportion of truly associated
630 transcripts called significant and the proportion of false positives since we
631 know which are the true associations. It can be observed that choosing a
632 significance threshold of p=0.001 will result in >60% of 'significantly'
633 associated transcripts being false positives as a result of there being a large
634 number of transcripts analysed (>30,000). Decreasing the p value threshold to
635 1^10^{-4} is more stringent and the proportion of false positive associations
636 decreases. Having a more stringent p-value threshold increases the
637 probability of missing truly associated transcripts, so sample size calculations
638 should be based around minimizing the probability of both detecting false
639 positives and also not missing true associations. The importance of sample

size calculations comes from the question you are trying to answer: attempting to detect 10-fold differences in transcript expression will require fewer individuals than attempting to detect 2-fold differences since the effect size is much larger and thus less likely to be observed by chance. The figure right hand panel illustrates this point. It is also true that the false negative rate is higher if the effect size is 2-fold compared to 10-fold if the study is underpowered. It is therefore critical when designing your experiment to have an estimate of the effect size you are expecting to observe. It is then possible to include enough individuals/samples in the experiment to detect this difference in a robust and reproducible manner, as illustrated in a transcriptome profiling study to distinguish different subsets of human naive T cells [34]. For a detailed exploration of sample size in eQTL studies, see [156].

Figure 1. Refining complex disease associations in different cellular activation states using chromatin annotation and chromatin conformation capture. Disease-associated regions with low recombination rate often harbour hundreds of credible SNPs, any one of which could be causal. Here we show the *IL2-IL21* region as an exemplar to highlight the utility of integrating regRNA (non-coding RNA associated with transcription as defined in Burren et al. [78]) annotations with ATAC-seq (open, potentially transcriptionally active chromatin) and PCHi-C (chromosome conformation

anlaysis using promoter capture Hi-C) profiles to reduce the number of credible SNPs and refine disease associations. T1D and celiac disease credible SNP sets extend over a 522 kb region on chromosome 4q27. Integrating the credible SNPs with regRNA (grey blocks), with areas of open chromatin using ATAC-seq (non-activated $CD4^+$ T cells; dark blue bars, activated $CD4^+$ T cells; light green bars, unpublished data) and with the published chromatin conformation capture PCHi-C data [78] [green (activated $CD4^+$ T cells) and blue blocks (non-activated $CD4^+$ T cells) representing promoter interacting regions (PIRs)] reduces the number of SNPs (n=208) to numbers that can be explored experimentally (n=14). The importance of investigating different $CD4^+$ T cell activation states is highlighted by the ATAC-seq and PCHi-C datasets. No PIRs were detected above threshold for the protein-coding genes *KIAA1109, IL21* and *ADAD1* in $CD4^+$ T cells in the region, the latter not being expressed, therefore they are not considered as candidate causal genes in this cell type. $CD4^+$ T cells were chosen for this analysis because we observed the greatest enrichment of T1D SNPs in enhancers in these cells, along with B cells, $CD8^+$ T cells and $CD34^+$ stem cells in a previous analysis [9], indicating that these cells are major T1D-associated cell types in blood. Two activation-induced interactions extend from the *IL21-ASI* (anti-sense non-coding transcript) promoter into the gene body of *KIAA1109* and overlap disease-associated SNPs but not open chromatin that is dynamically changed following $CD4^+$ T cell activation. The intergenic region between the IL-2 and IL-21 genes is enriched in activation-dependent PIR contacts (thicker and darker lines represent higher CHICAGO scores and therefore more frequent interactions [77]) that intersect with activation-induced open chromatin and link to the *IL2* promoter. *IL2* is therefore considered as the strongest candidate causal gene in this region, in this cell type under these conditions. These results justify further detailed investigations of chromosome conformation using higher resolution methods e.g. Capture-C [159], gene and regulatory region transcript expression [78] and mutagenesis of specific variants coupled to detailed haplotype mapping [23].

Region	GRCh37 coordinates	Candidate Genes	PCHi-C prioritised protein coding genes	PCHi-C prioritised non-protein coding transcripts
1p13.2	chr1:113830745-114551845	***PTPN22 PHTF1***	***ST7L DCLRE1B AP4B1***	**-**
1q32.1	chr1:206882358-207040938	***IL10***	***FCAMR IL20 FAIM3 PIGR CD55 IL24 IL19***	**-**
2q24.2	chr2:162960873-163360803	*IFIH1*	***FAP PSMD14 GCG***	**-**
2q33.2	chr2:204613986-204816575	*CTLA4*	***CYP20A1***	**-**
5p13.2	chr5:35798682-36036182	*IL7R*	***SPEF2*** ***DNAJC21***	***CTD-2113L7.1*** ***RNU7-130P***
6q15	chr6:90806835-91030155	***BACH2***	***MDN1***	***AL391559.1*** ***ENSG00000238747*** ***RP11-63K6.7 RP3-512E2.2***
7p15.2	chr7:26657962-27202289	-	***HOXA3 HOXA1***	***HOXA-AS2 HOTAIRM1 HOXA-AS3***
7p12.1	chr7:50900900-51134029	*COBL*	***GRB10***	**-**
7p12.2	chr7:50366637-50691711	***IKZF1***	***FIGNL1 HUS1 ZPBP C7orf72***	***RNU6-1091P*** ***AC020743.2***
10p15.1	chr10:6030243-6188338	***RBM17*** ***IL2RA***	***GDI2 PRKCQ ANKRD16 FAM208B*** ***FBX018***	***RP11-536K7.3*** ***PRKCQ-AS1***
10q23.31	chr10:90005048-90271019	*RNLS*	***PTEN KLLN***	**-**
11p15.5	chr11:2113931-2281231	***INS***	***TRPM5 TSSC4***	***AC124057.5***
12q13.2	chr12:56351346-56798435	***ERBB3*** *DGKA IKZF4*	***SMARCC2 TSPAN31 AGAP2 ZC3H10 SLC26A10 DTX3 PIP4K2C ARHGEF25 SUOX RPS26 CTDSP2 ESYT1***	***AC025165.8*** ***RP11-603J24.9***
12q24.13	chr12:111716376 -113030487	***SH2B3*** *NAA25*	***CUX2 MYL2***	***AC002978.1***
12p13.31	chr12:9519172-9972763	*CD69*	***CLEC7A CLEC9A*** ***TMEM52B GABARAPL1*** ***CLEC12B CLEC12A CLEC1B***	***RP11-656E20.5 RP11-133L14.5*** ***RNU6-700P***
13q32.3	chr13:99892888-100186578	***GPR183***	***GPR18 UBAC2***	***MIR623***
14q32.2	chr14:98361346-98604701	-	***ZFYVEL6***	**-**
14q24.1	chr14:69163455-69318062	-	***RAD51B***	**-**
14q32.2	chr14:101283661 -101328739	*DLK1*	***DIO3***	***MIR770 MEG3***
15q25.1	chr15:79001699-79261136	***CTSH***	***BCL2A1***	**-**
15q14	chr15:38814377-38994113	***RASGRP1***	***FAM98B C15orf53***	**-**
16p11.2	chr16:28295306-29025978	*IL27*	***SBK1 GSG1L***	**-**
16p13.13	chr16:11017058-11466511	***DEXI*** ***CLEC16A***	***RMI2 SOCS1 HNRNPCP4 GSPT1***	***RP11-485G7.6*** ***RP11-485G7.5 AC007216.1 AC009121.1***

16q23.1	chr16:75216240-75521030	*BCAR1*	***CTRB1 CTRB2 CHST6 GABARAPL2 SYCE1L WWOX WDR59 ZNRF1 MON1B***	***ENSG00000252122*** ***RNU6-758P***
17q12	chr17:37382674-38240761	*ORMDL3 GSDMB*	***ZPBP2***	*-*
19p13.2	chr19:10390709-10628548	***TYK2***	***ICAM3 ICAM4 ICAM1 OLFM2 MRPL4 ICAM5 PPAN EIF3G ANGPTL6 S1PR5 ZGLP1 PPAN-P2RY11 P2RY11 DNMT1 RAVER1 FDX1L***	***CTD-2369P2.4*** ***CTD-2369P2.12*** ***CTD-2369P2.8*** ***SNORD105B*** ***SNORD105***
21q22.3	chr21:43809176-43878660	*UBASH3A ICOSLG*	*-*	***RNU6-1149P AP001057.1***
22q12.2	chr22:30066344-30669187		***LIF***	***RP1-102K2.8***
22q12.3	chr22:37567843-37658804	***C1QTNF6*** *RAC2*	***TMPRSS6*** ***IL2RB***	***RP5-1170K4.7*** ***RP1-151B14.6***

Table 1. Promoter capture Hi-C (PCHi-C) increases the number of candidate genes and non-coding RNAs for T1D.

Integration of PCHi-C data from 17 different blood cell types [77] with ImmunoChip [9] and imputed GWAS data [8], using COGS [73] prioritised 97 novel T1D candidate protein-coding genes and 39 non-coding transcripts. We limit, for clarity, the regions in the table to 29 T1D-associated regions where new or previously nominated candidates (in bold text) were prioritised by PCHi-C and to gene biotypes that were captured with sufficient coverage (90% of protein-coding transcripts down to 50% of microRNAs but excludes non-coding biotypes such as lincRNAs with only 14% coverage) in the PCHi-C experimental design. We used a COGS gene score threshold of 0.5 for the prioritising of genes based on an integrative analysis of genetic association data and promoter-promoter interacting region (PIR) contacts. However, if a region has strong LD and many SNPs associated with the disease, the posterior probability for each SNP to be causal is lowered, therefore attenuating the overall COGS gene score. For this reason, genes in such regions, for example the *IL2-IL-21* region on 4q27, do not appear in this table, even though there are clear PIRs that overlap GWAS significant associations (Fig. 1). Currently, in total there are 57 T1D regions including the MHC listed in ImmunoBase (https://www.immunobase.org/disease/T1D/) plus one from an age-at-diagnosis analysis [33] making 58 regions in total as of February

2018. Gene annotations were derived from the Ensembl 75 (GRCh37) gene build [160]. We divided the table into separate columns denoting T1D-associated regions, GRCh37 coordinates, genes previously assigned candidacy for T1D, novel protein-coding genes and novel non-coding RNAs prioritised by PCHi-C.

Bibliography

1. Wang, W. Y. S., Barratt, B. J., Clayton, D. G. & Todd, J. A. Genome-wide association studies: Theoretical and practical concerns. *Nat. Rev. Genet.* **6,** 109–118 (2005).
2. Kong, A. *et al.* The nature of nurture: Effects of parental genotypes. *Science* **359,** 424–428 (2018).
3. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **101,** 5–22 (2017).
4. Timpson, N. J., Greenwood, C. M. T., Soranzo, N., Lawson, D. J. & Richards, J. B. Genetic architecture: The shape of the genetic contribution to human traits and disease. *Nat. Rev. Genet.* **19,** 110–124 (2018).
5. Liu, J. Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47,** 979–986 (2015).
6. Lange, K. M. De *et al.* Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat. Genet.* **49,** 256–261 (2017).
7. Rothschild, D. *et al.* Environment dominates over host genetics in shaping human gut microbiota. *Nature* (2018). doi:10.1038/nature25973
8. Cooper, N. J. *et al.* Type 1 diabetes genome-wide association analysis with imputation identifies five new disease regions Authors. *BiorXiv* (2017). doi:https://doi.org/10.1101/120022
9. Onengut-Gumuscu, S. *et al.* Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. *Nat. Genet.* **47,** 381–6 (2015).
10. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447,** 661–78 (2007).
11. Mahajan, A. *et al.* Fine-mapping of an expanded set of type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *bioRxiv* 245506 (2018). doi:10.1101/245506

12. Ziegler, A. G. *et al.* Primary prevention of beta-cell autoimmunity and type 1 diabetes - The Global Platform for the Prevention of Autoimmune Diabetes (GPPAD) perspectives. *Mol. Metab.* **5,** 255–262 (2016).
13. Thomas, N. J. *et al.* Frequency and phenotype of type 1 diabetes in the first six decades of life: A cross-sectional, genetically stratified survival analysis from UK Biobank. *Lancet Diabetes Endocrinol.* **6,** 122–129 (2018).
14. Brorsson, C. A. *et al.* Genetic Risk Score Modelling for Disease Progression in New-Onset Type 1 Diabetes Patients : Increased Genetic Load of Islet-Expressed and Cytokine-Regulated Candidate Genes Predicts Poorer Glycemic Control. *J. Diabetes Res.* **2016,** Article ID 9570424 (2016).
15. Sudlow, C. *et al.* UK Biobank : An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Med.* **12,** 1–10 (2015).
16. Tian, C. *et al.* Genome-wide association and HLA region fine-mapping studies identify susceptibility loci for multiple common infections. *Nat. Commun.* **8,** 599 (2017).
17. Ferreira, M. A. *et al.* Shared genetic origin of asthma, hay fever and eczema elucidates allergic disease biology. *Nat. Genet.* **49,** 1752–1757 (2017).
18. Wijmenga, C. & Zhernakova, A. The importance of cohort studies in the post-GWAS era. *Nat. Genet.* (2018). doi:10.1038/s41588-018-0066-3
19. Cortes, A. *et al.* Bayesian analysis of genetic association across tree-structured routine healthcare data in the UK Biobank. *Nat. Genet.* **49,** 1311–1318 (2017).
20. Bycroft, C. *et al.* Genome-wide genetic data on ~ 500, 000 UK Biobank participants. *BioRxiv* (2017). doi:https://doi.org/10.1101/166298
21. Zhou, X. & Stephens, M. Genome-wide efficient mixed model analysis for association studies. *Nat. Genet.* **44,** 821–824 (2012).
22. Ma, C., Blackwell, T., Boehnke, M. & Scott, L. J. Recommended Joint and Meta-Analysis Strategies for Case-Control Association Testing of Single Low-Count Variants Genetic Epidemiology. *Genet. Epidemiol.* **37,** 539–550 (2013).
23. Rainbow, D. B., Pekalski, M., Cutler, A. J. & Burren, O. A rare IL2RA haplotype identifies SNP rs61839660 as causal for autoimmunity. *bioRxiv* (2017). doi:https://doi.org/10.1101/108126

24. Morris Andrew. Transethnic Meta-Analysis of Genomewide Association Studies. *Genet. Epidemiol.* **35,** 809–822 (2011).
25. Kichaev, G. & Pasaniuc, B. Leveraging Functional-Annotation Data in Trans-ethnic Fine-Mapping Studies. *Am. J. Hum. Genet.* **97,** 260–271 (2015).
26. Farh, K. K.-H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518,** 337–43 (2015).
27. Wallace, C. *et al.* Dissection of a Complex Disease Susceptibility Region Using a Bayesian Stochastic Search Approach to Fine Mapping. *PLOS Genet.* **11,** e1005272 (2015).
28. Huang, H. *et al.* Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature* **547,** 173–178 (2017).
29. Chen, W. *et al.* Fine mapping causal variants with an approximate bayesian method using marginal test statistics. *Genetics* **200,** 719–736 (2015).
30. Benner, C. *et al.* FINEMAP: Efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32,** 1493–1501 (2016).
31. Newcombe, P. J., Conti, D. V. & Richardson, S. JAM: A Scalable Bayesian Framework for Joint Analysis of Marginal SNP Effects. *Genet. Epidemiol.* **40,** 188–201 (2016).
32. Benner, C. *et al.* Prospects of Fine-Mapping Trait-Associated Genomic Regions by Using Summary Statistics from Genome-wide Association Studies. *Am. J. Hum. Genet.* **101,** 539–551 (2017).
33. Inshaw, J. R. J., Walker, N. M., Wallace, C., Bottolo, L. & Todd, J. A. The chromosome 6q22.33 region is associated with age at diagnosis of type 1 diabetes and disease risk in those diagnosed under 5 years of age. *Diabetologia* **61,** 147–157 (2018).
34. Pekalski, M. L. *et al.* Neonatal and adult recent thymic emigrants produce IL-8 and express complement receptors CR1 and CR2. *JCI insight* **2,** e93739 (2017).
35. Davies, J. L. *et al.* Increased THEMIS first exon usage in CD4+ T-cells is associated with a genotype that is protective against multiple sclerosis. *PLoS One* **11,** 1–11 (2016).
36. Ferreira, R. C. *et al.* Functional IL6R 358Ala Allele Impairs Classical IL-6 Receptor Signaling and Influences Risk of Diverse Inflammatory Diseases. *PLoS*

814 *Genet.* **9,** e1003444 (2013).
815 37. Al-Mossawi, H. *et al.* The autoimmune disease risk allele rs6897932 modulates
816 monocyte IL7R surface and soluble receptor levels in a context-specific
817 manner. *BiorXiv* (2018). doi:10.1101/262410
818 38. Lappalainen, T. & Greally, J. M. Associating cellular epigenetic models with
819 human phenotypes. *Nat. Rev. Genet.* **18,** 441–451 (2017).
820 39. Mokry, L. E. *et al.* Vitamin D and Risk of Multiple Sclerosis : A Mendelian
821 Randomization Study. *PLOS Med.* **12,** 1–20 (2015).
822 40. Rhead, B. *et al.* Mendelian randomization shows a causal effect of low vitamin
823 D on multiple sclerosis risk. *Am. Acad. Neurol.* **2,** e97 (2016).
824 41. Burgess, S., Timpson, N. J., Ebrahim, S. & Smith, G. D. Mendelian
825 randomization: Where are we now and where are we going? *Int. J. Epidemiol.*
826 **44,** 379–388 (2015).
827 42. Koellinger, P. D. & Harden, K. P. Using nature to understand nurture. *Science*
828 **359,** 657–658 (2018).
829 43. Kindt, A. S. D. *et al.* Allele-specific methylation of type 1 diabetes susceptibility
830 genes. *J. Autoimmun.* doi:https://doi.org/10.1016/j.jaut.2017.11.008
831 44. Hemani, G., Tilling, K. & Davey Smith, G. Orienting the causal relationship
832 between imprecisely measured traits using GWAS summary data. *PLoS Genet.*
833 **13,** e1007081 (2017).
834 45. Hartwig, F. P., Davies, N. M., Hemani, G. & Smith, G. D. Counterfactual
835 causation: Avoiding the downsides of a powerful, widely applicable but
836 potentially fallible technique. *Int. J. Epidemiol.* **45,** 1717–1726 (2016).
837 46. Inoshita, M. *et al.* Retraction: A significant causal association between C-
838 reactive protein levels and schizophrenia. *Sci. Rep.* **8,** 46947 (2018).
839 47. Feingold, E. A. *et al.* The ENCODE (ENCyclopedia Of DNA Elements) Project.
840 *Science* **306,** 636–640 (2004).
841 48. Adams, D. *et al.* BLUEPRINT to decode the epigenetic signature written in
842 blood. *Nat. Biotechnol.* **30,** 224–226 (2012).
843 49. Bernstein, B. E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium
844 complex. *Nat. Biotechnol.* **28,** 1045–1048 (2010).
845 50. Andersson, R. *et al.* An atlas of active enhancers across human cell types and

tissues. *Nature* **507,** 455–461 (2014).
51. Elkon, R. & Agami, R. Characterization of noncoding regulatory DNA in the human genome. *Nat. Biotechnol.* **35,** 732–746 (2017).
52. Hnisz, D., Day, D. S. & Young, R. A. Insulated Neighborhoods: Structural and Functional Units of Mammalian Gene Control. *Cell* **167,** 1188–1200 (2016).
53. Degner, J. F. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482,** 390 (2012).
54. Maurano, M. T. *et al.* Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science* **337,** 1190–1195 (2012).
55. Hnisz, D. *et al.* Super-Enhancers in the Control of Cell Identity and Disease. *Cell* **155,** 934–947 (2013).
56. Vahedi, G. *et al.* Super-enhancers delineate disease-associated regulatory nodes in T cells. *Nature* **520,** 558–562 (2015).
57. Kichaev, G. *et al.* Integrating Functional Data to Prioritize Causal Variants in Statistical Fine-Mapping Studies. *PLoS Genet.* **10,** e1004722 (2014).
58. Pickrell, J. K. Joint Analysis of Functional Genomic Data and Genome-wide Association Studies of 18 Human Traits. *Am. J. Hum. Genet.* **94,** 559–573 (2014).
59. Yang, J., Fritsche, L. G., Zhou, X. & Abecasis, G. R. A Scalable Bayesian Method for Integrating Functional Information in Genome-wide Association Studies. *Am. J. Hum. Genet.* **101,** 404–416 (2017).
60. Marson, A., Housley, W. J. & Hafler, D. A. Genetic basis of autoimmunity. *J. Clin. Invest.* **125,** 2234–2241 (2015).
61. Kasela, S. *et al.* Pathogenic implications for autoimmune mechanisms derived by comparative eQTL analysis of CD4+versus CD8+T cells. *PLoS Genet.* **13,** e1006643 (2017).
62. De Jager, P. L. *et al.* ImmVar project: Insights and design considerations for future studies of 'healthy' immune variation. *Semin. Immunol.* **27,** 51–57 (2015).
63. Todd, J. A. Evidence that UBASH3 is a causal gene for type 1 diabetes. *Eur. J. Hum. Genet.*
64. Ongen, H. *et al.* Estimating the causal tissues for complex traits and diseases.

878 *Nat. Genet.* **49,** 1676–1683 (2017).
879 65. Reinius, B. *et al.* Analysis of allelic expression patterns in clonal somatic cells
880 by single-cell RNA-seq. *Nat. Genet.* **48,** 1430–1435 (2016).
881 66. Wijst, M. G. P. van der, Brugge, H., Vries, D. H. de & Franke, L. H. Single-cell
882 RNA sequencing reveals cell-type specific cis-eQTLs in peripheral blood
883 mononuclear cells. *bioRxiv* 177568 (2017). doi:10.1101/177568
884 67. Sun, B. B. *et al.* Genomic atlas of the human plasma proteome. *Nature* In press
885 (2018).
886 68. Keshishian, H. *et al.* Multiplexed, Quantitative Workflow for Sensitive
887 Biomarker Discovery in Plasma Yields Novel Candidates for Early Myocardial
888 Injury. *Mol. Cell. Proteomics* **14,** 2375–2393 (2015).
889 69. Dendrou, C. A. *et al.* Cell-specific protein phenotypes for the autoimmune
890 locus IL2RA using a genotype-selectable human bioresource. *Nat. Genet.* **41,**
891 1011–5 (2009).
892 70. Corbin, L. J. *et al.* Formalising recall by genotype as an efficient approach to
893 detailed phenotyping and causal inference. *Nat. Commun.* **9,** 711 (2018).
894 71. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing Chromosome
895 Conformation. *Science* **295,** 1306–1312 (2002).
896 72. Smemo, S. *et al.* Obesity-associated variants within FTO form long-range
897 functional connections with IRX3. *Nature* **507,** 371–375 (2014).
898 73. Davison, L. J. *et al.* Long-range DNA looping and gene expression analyses
899 identify DEXI as an autoimmune disease candidate gene. *Hum. Mol. Genet.* **21,**
900 322–333 (2012).
901 74. Lieberman-aiden, E. *et al.* Comprehensive Mapping of Long-Range
902 Interactions Reveals Folding Principles of the Human Genome. *Science* **326,**
903 289–294 (2009).
904 75. Rao, S. S. P. *et al.* A 3D Map of the Human Genome at Kilobase Resolution
905 Reveals Principles of Chromatin Looping. *Cell* **159,** 1665–1680 (2014).
906 76. Bonev, B. *et al.* Multiscale 3D Genome Rewiring during Mouse Article
907 Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell* **171,**
908 557–572 (2017).
909 77. Javierre, B. M. *et al.* Lineage-Specific Genome Architecture Links Enhancers

and Non-coding Disease Variants to Target Gene Promoters. *Cell* **167,** 1369–1384 (2016).

78. Burren, O. S. *et al.* Chromosome contacts in activated T cells identify autoimmune disease candidate genes. *Genome Biol.* **18,** 165 (2017).

79. Hughes, J. R. *et al.* Analysis of hundreds of cis -regulatory landscapes at high resolution in a single , high-throughput experiment. *Nat. Genet.* **46,** 205–212 (2014).

80. Davies, J. O. J. *et al.* Multiplexed analysis of chromosome conformation at vastly improved sensitivity. *Nat. Methods* **13,** 74–80 (2016).

81. Li, G. *et al.* ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. *Genome Biol.* **11,** 1–13 (2010).

82. Mumbach, M. R. *et al.* Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. *Nat. Genet.* **49,** 1602–1612 (2017).

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

84. Alasoo, K. *et al.* Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. *Nat. Genet.* **50,** 424–431 (2018).

85. Kumasaka, N., Knights, A. & Gaffney, D. High resolution genetic mapping of causal regulatory interactions in the human genome. *bioRxiv* (2017).

86. Ercolini, A. M. & Miller, S. D. The role of infections in autoimmune disease. *Clin. Exp. Immunol.* **155,** 1–15 (2008).

87. Matzaraki, V., Kumar, V., Wijmenga, C. & Zhernakova, A. The MHC locus and genetic susceptibility to autoimmune and infectious diseases. *Genome Biol.* **18,** 76 (2017).

88. Rodriguez-calvo, T., Sabouri, S., Anquetil, F. & Herrath, M. G. Von. The viral paradigm in type 1 diabetes : Who are the main suspects? *Autoimmun. Rev.* **15,** 964–969 (2016).

89. Ferreira, R. C. *et al.* A Type 1 Interferon Transcriptional Signature Precedes Autoimmunity in Children Genetically at Risk for Type 1 Diabetes. *Diabetes* **63,** 2538–2550 (2014).

90. Beyerlein, A., Donnachie, E., Jergens, S. & Ziegler, A. Infections in Early Life and Development of Type 1 Diabetes. *JAMA* **315,** 1899–1901 (2016).
91. Trowsdale, J. & Knight, J. C. Major Histocompatibility Complex Genomics and Human Disease. *Annu. Rev. Genomics Hum. Genet.* **14,** 301–323 (2013).
92. Todd, J. A., Bell, J. I. & McDevitt, H. O. HLA-DQB gene contributes to susceptibility and resistance to insulin-dependent diabetes mellitus. *Nature* **327,** 599–604 (1987).
93. Howson, J. M. M., Walker, N. M., Clayton, D. & Todd, J. A. Confirmation of HLA class II independent type 1 diabetes associations in the major histocompatibility complex including HLA-B and HLA-A. *Diabetes, Obes. Metab.* **11,** 31–45 (2009).
94. Eriksson, N. *et al.* Web-Based , Participant-Driven Studies Yield Novel Genetic Associations for Common Traits. *PLoS Genet.* **6,** e1000993 (2010).
95. Bush, W. S., Oetjens, M. T. & Crawford, D. C. Unravelling the human genome-phenome-wide association studies. *Nat. Rev. Genet.* **17,** 129–145 (2016).
96. Kelly, R. J., Rouquier, S., Giorgi, D., Lennon, G. G. & Lowe, J. B. Sequence and expression of a candidate for the human Secretor blood group alpha(1,2)Fucosyltransferase gene (FUT2). Homozygosity for an enzyme-inactivating nonsense mutation commonly correlates with the non-secretor phenotype. *J. Biol. Chem.* **270,** 4640–4649 (1995).
97. Indesmith, L. I. S. a L. *et al.* Human susceptibility and resistance to Norwalk virus infection. *Nat Med* **9,** 548–553 (2003).
98. Boren, T., Falk, P., Roth, K., Larson, G. & Normark, S. Attachment of Helicobacter pylori to human gastric epithelium mediated by blood group antigens. *Science* **262,** 1892–1895 (1993).
99. Rausch, P. *et al.* Colonic mucosa-associated microbiota is influenced by an interaction of Crohn disease and FUT2 (Secretor) genotype. *PNAS* **108,** 19030–19035 (2011).
100. Wacklin, P. *et al.* Faecal Microbiota Composition in Adults Is Associated with the FUT2 Gene Determining the Secretor Status. *PLoS One* **9,** e94863 (2014).
101. Wacklin, P. *et al.* Secretor genotype (FUT2 gene) is strongly associated with the composition of bifidobacteria in the human intestine. *PLoS One* **6,** e20113

(2011).
102. Tong, M. *et al.* Reprograming of gut microbiome energy metabolism by the FUT2 Crohn ' s disease risk polymorphism. *ISME J.* **8,** 2193–2206 (2014).
103. Jacobs, J. P. & Braun, J. Immune and genetic gardening of the intestinal microbiome. *FEBS Lett.* **588,** 4102–4111 (2014).
104. Smyth, D. J. *et al.* FUT2 Nonsecretor Status Links Type 1 Diabetes Susceptibility and Resistance to Infection. *Diabetes* **60,** 3081–3084 (2011).
105. Mcgovern, D. P. B. *et al.* Fucosyltransferase 2 (FUT2) non-secretor status is associated with Crohn ' s disease. *Hum. Mol. Genet.* **19,** 3468–3476 (2010).
106. Hall, A. B., Tolonen, A. C. & Xavier, R. J. Human genetic variation and the gut microbiome in disease. *Nat. Rev. Genet.* **18,** 690–699 (2017).
107. Thorven, M. *et al.* A Homozygous Nonsense Mutation (428G -> A) in the Human Secretor (FUT2) Gene Provides Resistance to Symptomatic Norovirus (GGII) Infections. *J. Virol.* **79,** 15351–15355 (2005).
108. Dodd, D. *et al.* A gut bacterial pathway metabolizes aromatic amino acids into nine circulating metabolites. *Nature* **551,** 648–652 (2017).
109. Kamada, N., Seo, S. U., Chen, G. Y. & Núñez, G. Role of the gut microbiota in immunity and inflammatory disease. *Nat. Rev. Immunol.* **13,** 321–335 (2013).
110. Mclean, J. S. Advancements toward a systems level understanding of the human oral microbiome. *Front. Cell. Infect. Microbiol.* **4,** 98 (2014).
111. Köhling, H. L., Plummer, S. F., Marchesi, J. R., Davidge, K. S. & Ludgate, M. The microbiota and autoimmunity : Their role in thyroid autoimmune diseases. *Clin. Immunol.* **183,** 63–74 (2017).
112. Yurkovetskiy, L. A., Pickard, J. M. & Chervonsky, A. V. Microbiota and Autoimmunity : Exploring New Avenues Minireview. *Cell Host Microbe* **17,** 548–552 (2015).
113. Brown, C. T. *et al.* Gut Microbiome Metagenomics Analysis Suggests a Functional Model for the Development of Autoimmunity for Type 1 Diabetes. *PLoS One* **6,** 1–9 (2011).
114. de Goffau, M. C. *et al.* Fecal Microbiota Composition Differs Between Children With b -Cell Autoimmunity and Those Without. *Diabetes* **62,** 1238–1244 (2013).

115. Kostic, A. D. *et al.* The Dynamics of the Human Infant Gut Microbiome in Development and in Progression toward Type 1. *Cell Host Microbe* **17,** 260–273 (2015).
116. Needell, J. C. & Zipris, D. The Role of the Intestinal Microbiome in Type 1 Diabetes Pathogenesis. *Curr. Diab. Rep.* **16,** 89 (2016).
117. Paun, A., Yau, C. & Danska, J. S. The Influence of the Microbiome on Type 1 Diabetes. *J. Immunol.* **198,** 590–595 (2017).
118. Dunne, J. L. *et al.* The intestinal microbiome in type 1 diabetes. *Clin. Exp. Immunol.* **177,** 30–37 (2014).
119. Mullaney, J. A. *et al.* Type 1 diabetes susceptibility alleles are associated with distinct alterations in the gut microbiota. *Microbiome* **6,** 35 (2018).
120. Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K. & Knight, R. Diversity , stability and resilience of the human gut microbiota. *Nature* **489,** 220–230 (2012).
121. Mosca, A., Leclerc, M. & Hugot, J. P. Gut Microbiota Diversity and Human Diseases : Should We Reintroduce Key Predators in Our Ecosystem? *Front. Microbiol.* **7,** 455 (2016).
122. Santin, I., Dos Santos, R. S. & Eizirik, D. L. Pancreatic Beta Cell Survival and Signaling Pathways: Effects of Type 1 Diabetes-Associated Genetic Variants. *Methods Mol. Biol.* **1433,** 21–54 (2016).
123. Marroqui, L. *et al.* TYK2 , a Candidate Gene for Type 1 Diabetes , Modulates Apoptosis and the Innate Immune Response in Human Pancreatic β -Cells. *Diabetes* **64,** 3808–3817 (2015).
124. Dendrou, C. A. *et al.* Resolving TYK2 locus genotype-to-phenotype differences in autoimmunity. *Sci. Transl. Med.* **8,** 363ra149 (2016).
125. Dooley, J. *et al.* Genetic predisposition for beta cell fragility underlies type 1 and type 2 diabetes. *Nat. Genet.* **48,** 519–527 (2016).
126. Nogueira, T. C. *et al.* GLIS3 , a Susceptibility Gene for Type 1 and Type 2 Diabetes , Modulates Pancreatic Beta Cell Apoptosis via Regulation of a Splice Variant of the BH3-Only Protein Bim. *PLoS Genet.* **9,** e1003532 (2013).
127. Graham, K. L. *et al.* Pathogenic Mechanisms in Type 1 Diabetes : The Islet is Both Target and Driver of Disease. *Rev. Diabet. Stud.* **9,** 148–168 (2012).

128. Liu, J. Z. *et al.* Dense fine-mapping study identifies new susceptibility loci for primary biliary cirrhosis. *Nat. Genet.* **44,** 1137–1141 (2012).
129. Boettger, L. M., Handsaker, R. E., Zody, M. C. & McCarroll, S. A. Structural haplotypes and recent evolution of the human 17q21.31 region. *Nat. Genet.* **44,** 881–885 (2012).
130. Bekpen, C., Tastekin, I., Siswara, P., Akdis, C. A. & Eichler, E. E. Primate segmental duplication creates novel promoters for the LRRC37 gene family within the 17q21.31 inversion polymorphism region. *Genome Res.* **22,** 1050–1058 (2012).
131. Zody, M. C. *et al.* Evolutionary toggling of the MAPT 17q21.31 inversion region. *Nat. Genet.* **40,** 1076–1083 (2008).
132. Lai, M. C. *et al.* Haplotype-specific MAPT exon 3 expression regulated by common intronic polymorphisms associated with Parkinsonian disorders. *Mol. Neurodegener.* **12,** 79 (2017).
133. Bunt, M. Van De *et al.* Transcript Expression Data from Human Islets Links Regulatory Signals from Genome- Wide Association Studies for Type 2 Diabetes and Glycemic Traits to Their Downstream Effectors. *PLoS Genet.* **11,** e1005694 (2015).
134. Varshney, A. *et al.* Genetic regulatory signatures underlying islet gene expression and type 2 diabetes. *PNAS* **114,** 2301–2306 (2017).
135. Huising, M. O. *et al.* CRFR1 is expressed on pancreatic β cells, promotes β cell proliferation, and potentiates insulin secretion in a glucose-dependent manner. *PNAS* **107,** 912–917 (2010).
136. Blaabjerg, L. *et al.* CRFR1 activation protects against cytokine-induced beta cell death. *J. Mol. Endocrinol.* **53,** 417–427 (2015).
137. Schmid, J. *et al.* Modulation of pancreatic islets-stress axis by hypothalamic releasing hormones and 11β-hydroxysteroid dehydrogenase. *PNAS* **108,** 13722–13727 (2011).
138. Miklossy, J. *et al.* Beta amyloid and hyperphosphorylated tau deposits in the pancreas in type 2 diabetes. *Neurobiol. Aging* **31,** 1503–1515 (2010).
139. Maj, M. *et al.* Expression of TAU in insulin-secreting cells and its interaction with the calcium-binding protein secretagogin. *J. Endocrinol.* **205,** 25–36

(2010).
140. Wijesekara, N. *et al.* Amyloid-β and islet amyloid pathologies link Alzheimer disease and type 2 diabetes in a transgenic model. *FASEB J.* **31,** 5409–5418 (2017).
141. Eberhard, D. Neuron and beta-cell evolution : Learning about neurons is learning about beta-cells. *BioEssays* **35,** 584 (2013).
142. Calderari, S. *et al.* Molecular genetics of the transcription factor GLIS3 identifies its dual function in beta cells and neurons. *Genomics* **110,** 98–111 (2018).
143. Marroqui, L. *et al.* Interferon-alpha mediates human beta cell HLA class I overexpression, endoplasmic reticulum stress and apoptosis, three hallmarks of early human type 1 diabetes. *Diabetologia* **60,** 656–667 (2017).
144. Perri, E. R., Thomas, C. J., Parakh, S., Spencer, D. M. & Atkin, J. D. The Unfolded Protein Response and the Role of Protein Disulfide Isomerase in Neurodegeneration. *Front. cell Dev. Biol.* **3,** 80 (2015).
145. Nelson, M. R. *et al.* The support of human genetic evidence for approved drug indications. *Nat. Genet.* **47,** 856–860 (2015).
146. Sanseau, P. *et al.* Use of genome-wide association studies for drug repositioning. *Nat. Biotechnol.* **30,** 317–320 (2012).
147. Koreth, J. *et al.* Interleukin-2 and Regulatory T Cells in Graft-versus-Host Disease. *N. Engl. J. Med.* **365,** 2055–2066 (2017).
148. He, J. *et al.* Low-dose interleukin-2 treatment selectively modulates CD4 + T cell subsets in patients with systemic lupus erythematosus. *Nat. Med.* **22,** 991–993 (2016).
149. Saadoun, D. *et al.* Regulatory T-Cell Responses to Low-Dose Interleukin-2 in HCV-Induced Vasculitis. *N. Engl. J. Med.* **365,** 2067–2077 (2017).
150. Todd, J. A. *et al.* Regulatory T Cell Responses in Participants with Type 1 Diabetes after a Single Dose of Interleukin-2: A Non-Randomised, Open Label, Adaptive Dose-Finding Trial. *PLOS Med.* **13,** Article ID 27727279 (2016).
151. Vodovotz, Y. *et al.* Solving Immunology? *Trends Immunol.* **38,** 116–127 (2017).
152. Pappalardo, J. L. & Hafler, D. A. The Human Functional Genomics Project: Understanding Generation of Diversity. *Cell* **167,** 894–896 (2017).

153. Segerstolpe, A. *et al.* Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. *Cell Metab.* **24,** 593–607 (2016).
154. Ivison, S., Des Rosiers, C., Lesage, S., Rioux, J. D. & Levings, M. K. Biomarker-guided stratification of autoimmune patients for biologic therapy. *Curr. Opin. Immunol.* **49,** 56–63 (2017).
155. West, N. R. *et al.* Oncostatin M drives intestinal inflammation and predicts response to tumor necrosis factor-neutralizing therapy in patients with inflammatory bowel disease. *Nat. Med.* **23,** 579–589 (2017).
156. Huang, Q. Q., Ritchie, S. C., Brozynska, M. & Inouye, M. Power, false discovery rate and Winner's Curse in eQTL studies. *bioRxiv* (2017). doi:https://doi.org/10.1101/209171
157. Lotta, L. A. *et al.* Genetic Predisposition to an Impaired Metabolism of the Branched-Chain Amino Acids and Risk of Type 2 Diabetes: A Mendelian Randomisation Analysis. *PLOS Med.* **13,** e1002179 (2016).
158. Tibshirani, R. A simple method for assessing sample sizes in microarray experiments. *BMC Bioinformatics* **7,** 106 (2006).
159. Oudelaar, A. M., Davies, J. O. J., Downes, D. J., Higgs, D. R. & Hughes, J. R. Robust detection of chromosomal interactions from small numbers of cells using low-input Capture-C. *Nucleic Acids Res.* **45,** e184–e184 (2017).
160. Flicek, P. *et al.* Ensembl 2014. *Nucleic Acids Res.* **42,** D749–D755 (2014).