A natural allelic series of complex structural variants and its influence on the risk of lupus and schizophrenia

A natural allelic series of complex structural variants and its influence on the risk of lupus and schizophrenia

A dissertation presented

by

Aswin Sekar

to

The Division of Medical Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Genetics and Genomics

Harvard University

Cambridge, Massachusetts

June 2014

Dissertation Advisor: Professor Steven A. McCarroll Aswin Sekar

A natural allelic series of complex structural variants and its influence on the risk of lupus and schizophrenia

Abstract

The human genome's strongest influences on two common diseases, systemic lupus erythematosus (SLE) and schizophrenia, arise from genetic variation in the *Human Leukocyte Antigen* (*HLA*) locus. However, the genes and functional alleles driving these genetic relationships have remained unknown. We hypothesized that a complex, multi-allelic form of structural variation in the *Complement component 4* (*C4*) gene, within the *HLA* locus, underlies these relationships.

Loci that exist in many structural forms and vary widely in copy number have been difficult to analyze molecularly. As a result, we know little about their population genetic properties or their influence on phenotypes. In this work, we developed molecular and statistical methods to characterize such loci and to evaluate their contribution to phenotypes.

Applying these methods to the *C4* locus, we found that *C4* segregates in four common and at least eleven low-frequency structural forms in human populations. Although there was only partial correlation between *C4* structural variation and individual single nucleotide polymorphisms (SNPs), we developed an imputation approach to enable statistical prediction of *C4* structural states from flanking SNP haplotypes.

C4 structural variation associated to gene expression in lymphoblastoid cell lines and human brain tissue. Applying our imputation strategy to SLE and schizophrenia case-control cohorts totaling > 75,000 individuals, we found that structural variation in *C4* contributes to risk of both phenotypes in a manner predicted by its effect on gene expression in relevant tissues, and with largely opposite directions of effect – alleles that were protective for schizophrenia increased risk for SLE, and vice versa. Leveraging a natural allelic series of *C4* structural forms, we developed a novel form of association testing and showed that the association to *C4* is unlikely to be caused by correlation with *HLA* SNPs. *C4* was expressed in human neurons, whereas other upstream complement pathway genes were expressed primarily by microglia. Mice lacking *C4* showed a deficit in synaptic pruning that was rescued by human *C4*.

The methods developed in this thesis enable analysis of complex structural variation, and our results identify a novel form of genome variation as making a strong contribution to phenotypes.

Table of Contents

Acknowledgements

My graduate school experience has been more wonderful than I could have asked for, and I have many people to thank for that.

First, I would like to thank my thesis advisor, Steve McCarroll, who has been an incredible mentor. It is an honor to be Steve's first graduating student, and I am privileged to have had the opportunity to be mentored by him within the first few years of his starting the lab. Steve has been an extremely approachable mentor and generous with his time, whether it is in answering the flood of emails that I routinely send him, meeting with me whenever I needed his advice, or giving me meticulous slide-by-slide feedback when preparing for presentations. I have learned so much from Steve, and I hope to emulate the many qualities that I admire in him.

I would like to thank members of the McCarroll lab for making the lab an incredibly fun, supportive, and engaging environment to be a part of. I am grateful for all the contributions that Vanessa Van Doren, Heather de Rivera and Sam Rose have made to the work presented in this thesis. I could not have asked for better officemates than Linda Boettger, Avery Davis, Cindy Liu, and Christina Usher. It has been so much fun working alongside them, and I thank them for their support and friendship. I thank Nolan Kamitaki for the many fun and interesting conversations and brainstorming sessions we had. I thank Elizabeth Fels for going above and beyond in providing administrative support.

I am grateful for having had the opportunity to work with and learn from a wonderful group of collaborators, who have significantly contributed to the depth and breadth of my training – Prof. Beth Stevens, Prof. Michael Carroll, Allison Bialas, Jessy Presumey, Matthew Baum, and Arnaud Frouin. Their contributions have expanded the scope of my thesis research, and the weekly meetings that I have had with them have been one of the highlights of my graduate school experience.

I am privileged to have had the opportunity to be guided by a Dissertation Advisory Committee of world-class experts in their respective fields – Professors Michael Carroll, Mark Daly, and Steve Hyman. I would also like to thank Professors Guoping Feng, Jesse Gray, Joel Hirschhorn, and Steve Hyman for kindly agreeing to be a part of my dissertation examination committee.

I appreciate the support that I have received from the MD/PhD Program leadership and staff, including Prof. Loren Walensky, Prof. Thomas Michel, and Amy Cohen, and I thank Kate Hodgins and the Biological and Biomedical Sciences Program staff, who have kept me and my fellow students on track.

I thank my friends Michael Marcus and Jennifer Kim, who have been a source of support and encouragement, even from afar. I also thank Alejandro de Los Angeles, my MD/PhD classmate and friend, with whom conversations during the first two years of medical school influenced my decision to pursue graduate training.

I am indebted to my parents, Saraswathi and Chander Sekar, who have made so many sacrifices for me. They have made it possible for me to have the ability and freedom to focus on what I want to, and have been there to make sure that everything else is taken care of. I also thank my brother, Aravind, for helping me make tough decisions along the way. I am grateful to have them as family and thank them for their unwavering support.

Chapter 1

Introduction

Overview

Genome-wide association studies have identified thousands of genetic variants associated with human traits[1]. However, the functional variants responsible for driving the vast majority of these associations are largely unknown. Association of phenotypes to genetic markers in the *Human Leukocyte Antigen* (*HLA*) locus on chromosome 6 represents a particularly challenging example of this problem. The *HLA* locus has been found to associate with dozens of autoimmune and infectious disease phenotypes, and it harbors the human genome's strongest influences on the risk of two phenotypes that are investigated in this thesis – systemic lupus erythematosus (SLE) and schizophrenia.

Genome-scale association studies of SLE, schizophrenia, and other complex phenotypes have so far ascertained only the simplest forms of genetic variation, whether rare or common: single nucleotide variation and simple deletions and duplications of genomic segments. However, a large and interesting class of genetic variation – complex and multi-allelic forms of structural variation – has yet to be rigorously investigated because of difficulties in analyzing it molecularly, and open questions remain: how do structurally complex loci vary across human populations? How can they be characterized molecularly and statistically? How can their contribution to phenotypes be investigated?

Complement component 4 (*C4*)*,* a gene within the *HLA* locus, exhibits a complex form of structural variation, and in this thesis, we

(i) developed molecular and statistical approaches to enable analysis of complex genome structural variation (Chapter 2); and

(ii) applied these methods toward evaluating whether structural variation in *C4* contributes to the risk of SLE (Chapter 3) and schizophrenia (Chapter 4).

We begin by motivating and contextualizing the studies described in the chapters that follow.

SLE and schizophrenia are complex diseases for which a better mechanistic understanding and new treatment options are needed

SLE

Systemic lupus erythematosus is a chronic autoimmune disorder that affects multiple organ systems including skin, kidneys, brain, and blood, with a tendency to affect women of childbearing age. SLE has an estimated heritability of 66%[2], a prevalence ranging from 20-150 cases per 100,000[3], and it affects at least five million people worldwide. A diagnosis of SLE is made when any four out of eleven clinical criteria set by the American College of Rheumatology are met either serially or simultaneously, reflecting the clinical heterogeneity of the disease. Genetic and environmental factors resulting in immune system dysfunction with production of auto-antibodies, immune complexes, and overactive B and T cells are thought to play a role in many of the symptoms of SLE, but the pathogenesis is not completely understood.

SLE patients are typically treated with nonsteroidal anti-inflammatory drugs (NSAIDs), antimalarial drugs, glucocorticoids, and immunosuppressive agents depending on the specific organs involved. Despite the availability of these treatment options, most patients experience a relapsing and remitting course – complete remission for more than 5 years is rare (< 2% of patients)[4] – and treatment can be associated with serious side effects such as bone marrow suppression and infections.

Schizophrenia

Schizophrenia is a disabling psychiatric disorder involving chronic or recurrent psychosis, with a typical age of onset in late adolescence to early adulthood. The symptoms of schizophrenia can be grouped into three broad categories: positive symptoms (delusions, hallucinations, and disorganized thought and speech), negative symptoms (such as flat affect and poverty of speech), and cognitive symptoms (impairment in memory, attention, and executive function). Schizophrenia has a heritability estimated to be as high as 80%[5], a lifetime prevalence of about 1%, and it affects at least 25 million people worldwide. The World Health Organization has ranked schizophrenia among the top ten illnesses contributing to the global burden of disease, based on Years Lost due to Disability (YLD)[6], and schizophrenia is associated with a higher all-cause mortality compared to the general population[7].

Multiple antipsychotic medications exist for treating schizophrenia. However, the molecular targets of approved drugs in use for schizophrenia today are the same as those from more than 40 years ago[8], reflecting the fact that the pathophysiology of schizophrenia is poorly understood. Less than 20% of schizophrenia patients remain relapse-free after experiencing their first psychotic episode[9], and antipsychotic medications can cause significant side effects. Despite the need for the development of new treatment options, there has been an exodus of pharmaceutical companies away from drug development for psychiatric disorders in the recent years[8].

Genetic studies in humans can provide novel biological insights and identify new molecular drug targets

While there is a strong need for new treatment options for SLE, schizophrenia, and many other diseases, the vast majority of candidate drugs that enter phase I clinical trials across multiple therapeutic areas are not successfully approved[10]. In most cases, the reason for the failure was not because the drugs did not act on the intended target, but they failed despite doing so; the therapeutic hypothesis regarding which targets would be efficacious and safe to perturb was incorrect in those cases[11].

Studying the genetic basis of disease in humans has the potential to provide novel insights into the pathogenesis of disease[12]. For example, the role of the complement pathway in age-related macular degeneration (AMD) was not widely appreciated until one of the first genome-wide association studies (GWAS) identified an association to a polymorphism in *complement factor H* (*CFH*)[13]. Similarly, GWAS also implicated autophagy in the pathogenesis of Crohn's disease[14].

Human genetics can also serve as a valuable tool for identifying drug targets, and there are prospective as well as retrospective examples of this[11]. For example, linkage and sequencing studies led to the identification of a gain-of-function mutation in *Proprotein convertase subtilisin/kexin type 9*, (*PCSK9*) causing hypercholesterolemia[15], a loss-of-function allele in the same gene in patients with low levels of low-density lipoprotein (LDL) cholesterol levels[16], and a spectrum of other *PCSK9* alleles contributing to LDL levels[17]. In phase I clinical trials, a monoclonal antibody to PCSK9 significantly reduced LDL cholesterol in healthy subjects as well as those with hypercholesterolemia[18]. As another example, the identification of mutations in the *sodium channel, voltage-gated, type IX, alpha subunit* (*SCN9A*) gene in patients with a rare pain disorder[19] and loss-of-function mutations in the same gene causing a congenital inability to feel pain[20] has led investigators to develop drugs that block the gene product as a treatment for pain[21]. There are also retrospective examples in which human genetics has identified molecular targets for drugs that were already in use, as being important in disease. For example, a recent genome-wide association study of schizophrenia[22] identified associations to *dopamine receptor D2* (*DRD2*), which is the target of effective antipsychotic medications for schizophrenia. In addition, common variants in the *3-hydroxy-3-methylglutaryl-CoA reductase* (*HMGCR*) gene, the product of which is the target of statins, have been found to associate to blood lipid levels[23].

The presence of a series of alleles within the same gene that contributes to a phenotype – as is the case with *PCSK9* and *SCN9A*, described above – can be an extremely valuable tool that human genetic studies can deliver. Such an allelic series strengthens confidence in a gene's involvement in a disease, and can generate a genotype-phenotype dose-response curve[11] that predicts the effect of modulating the gene's function on the phenotype. Along these lines, the following chapters in this thesis will describe the characterization of a natural allelic series of structural variants of the *C4* gene (Chapter 2), identification of its functional correlates, and evaluation of its role in clinical phenotypes (Chapters 3 and 4).

Challenges in realizing the potential of genome-wide association studies to provide actionable insights

While genome-wide association studies (GWAS) have identified thousands of genetic variants associating to human traits, there are analytical as well as experimental challenges in extracting actionable biological and therapeutic insights from these findings.

Linkage disequilibrium (LD) blocks – regions of the genome containing SNPs whose genotypes are substantially correlated with each other – identified by the International HapMap Consortium[24], allowed for 'tag SNPs' to be used for genotyping in GWAS, whereby a subset of common genetic variants being directly genotyped in an association study provides information about a larger set of common variants because of correlation between their genotypes. This very property that is useful in reducing the number of variants that needs to be experimentally genotyped has also made fine-mapping challenging. GWAS-identified loci typically contain multiple genes, depending on the size of the LD blocks – and in the case of the *HLA* locus described below, it can span hundreds of genes. Within a locus, the variant that is most strongly associated to a phenotype need not be the one that is causing the association (i.e., the 'causal variant'), for several reasons. First, because GWASs do not interrogate every genetic variant (common and rare), the observed association to variants that are evaluated can be driven by variants that are not analyzed in the study. Second, because of statistical fluctuation, the variant that is most strongly associated within a locus can change across studies and with increasing sample size. Third, in regions where multiple, independent effects on the phenotype exist, variants that underlie none of the effects can appear to associate strongly due to their correlation with those causally driving the association.

Even after the genes and causal alleles implicated in a phenotype are identified, utilizing these findings toward gaining biological insights can be challenging. Most variants identified in GWAS are in non-coding regions of the genome[25] (and many are not in LD with a protein-coding variant), in contrast to the highly penetrant, often protein-altering mutations causing Mendelian diseases, and understanding their mechanistic and functional implications can be challenging. In addition, for diseases like schizophrenia, for which there are neither counterparts in model organisms nor robust molecular endophenotypes, it is also not clear which assays, cell types, or model systems to use toward studying the functional consequences of disease-associated genetic variants.

Dozens of genetic risk loci have been identified for SLE and schizophrenia, but the functional alleles and their mechanistic implications remain largely unknown

SLE

Genome-wide association studies of SLE have identified more than 30 loci, which account for less than 10% of the genetic heritability. Among the variants that are most strongly associated within each locus, eleven are intergenic and ten are intronic[26].

While the functional alleles and their mechanisms remain largely unknown, the likely causal variant and its functional consequences have been identified for at least one of these loci. A non-synonymous variant in the *integrin, alpha M* (*ITGAM*) was found to associate with SLE in European-Americans, Hispanic-Americans, and African-Americans and to best explain the association of SLE to the *ITGAM* locus[27]. *ITGAM* is expressed mainly on myeloid cells, and one of its multiple ligands is iC3b, a proteolytic cleavage fragment of the central protein of the complement pathway. The SLE-associated non-synonymous variant in *ITGAM*, rs1143679, was shown to cause impaired cell adhesion to iC3b and reduced iC3b-dependent phagocytosis[28].

Genes within SLE-associated loci fall into multiple pathways[3]: immune complex processing and innate immunity (to which genes of the complement pathway belong); dendritic cell function and interferon signaling; T-cell function and signaling; B-cell function and signaling; cell cycle, apoptosis, and cellular metabolism; and transcriptional regulation. Some of the loci associated with SLE have also been implicated in other autoimmune diseases, whereas other loci appear to be unique to SLE. For example, the *STAT4* (*signal transducer and activator of transcription 4*) locus has also been implicated in inflammatory bowel disease[29], and Sjogren's syndrome[30]. Likewise, the *TNFAIP3* (*tumor necrosis factor alpha induced protein 3*) locus is also associated with rheumatoid arthritis[31], celiac disease[32], and psoriasis[33], in addition to SLE. However, other loci, including *ITGAM* and *SLC15A4* (*Solute carrier family 15, member 4*), do not have known associations to other autoimmune phenotypes from genome-wide association studies[26]. Overall, a comprehensive analysis of shared loci between SLE and seventeen autoimmune phenotypes found limited genetic overlap outside of the *HLA* locus[34]. This relatively distinct genetic susceptibility for SLE compared to several other autoimmune diseases raises the question of to what extent SLE would be considered an autoimmune phenotype based on its genetic risk factors if it did not associate to the *HLA* locus.

Schizophrenia

Over the past few years, various types of genetic analyses – studies of rare copy number variation, genome-wide association studies, and exome sequencing studies – have provided insights into the genetic architecture of schizophrenia and have shown that genetic variants of all sizes and frequencies contribute to the risk of schizophrenia.

Results from common variant as well as rare variant studies have taught us that schizophrenia is a highly polygenic disorder – its risk is likely influenced by a very large number of genes, each contributing a small effect. In 2009, Purcell and colleagues[35] developed the concept of polygenic risk score, whereby they calculated an aggregate score by quantifying variation across nominally associated loci and provided evidence that the risk of schizophrenia is influenced by thousands of common alleles of small effect. A recent exome sequencing study of schizophrenia[36] provided additional support for the polygenic model by demonstrating that rare, disruptive mutations were distributed across many genes. Earlier findings of an increased genome-wide burden of rare copy number variants in schizophrenia[37] are also consistent with the polygenic nature of the disorder.

Genetic risk loci associated to schizophrenia have also been implicated in other neuropsychiatric disorders, both adult- and childhood-onset. The polygenic component contributing to schizophrenia also influences risk of bipolar disorder, but not other, non-psychiatric diseases[35]. In an analysis of genome-wide pleiotropy between schizophrenia and four other psychiatric disorders (bipolar disorder, major depression, autism spectrum disorder, and attention deficit hyperactivity disorder) based on genetic correlation quantified using common variants, schizophrenia was shown to have the highest co-heritability with bipolar disorder among the phenotypes studied[38]. A joint analysis of these five psychiatric disorders identified associations to four loci, two of which were to voltage-gated calcium channel subunits[39]. The sharing of risk loci across neuropsychiatric disorders has also been appreciated from studies of rare copy number variation. For example, duplication of the 16p11.2 locus is associated with schizophrenia, bipolar disorder, and autism, and deletions of the same locus are associated with autism and developmental disorders[40]. In addition, rare, recurrent deletions of the 1q21.1 locus are associated with schizophrenia[37] and duplications of the same locus have also been described in patients with autism[41].

A recent, large GWAS of schizophrenia[22] identified 108 associated loci and 128 statistically independent effects within these loci. The associated regions included multiple subunits of the L-type voltage-gated calcium channels and genes involved in glutamatergic neurotransmission, both of which have also been implicated by studies of rare variants[36, 42] . Among the 108 loci, the physical boundaries of which were delimited by SNPs whose correlation to the most strongly associated variant in the region was an $r^2 \geq 0.6$, 17 had no known genes within the boundaries, 45 had a single gene, 11 had 2 genes, 26 had between 3 and 10 genes, and 9 had more than 10 genes. Of these, the locus that was least well mapped was the *HLA* locus, where the association signal spanned dozens to hundreds of genes.

The *HLA* locus harbors the genome's strongest influences on the risk of SLE and schizophrenia, but the association is complex and unsolved

The *HLA* locus stands out among the loci associated with SLE and schizophrenia from genome-wide association studies, both in terms of strength of association and effect size (**Figure 1.1**), and this association is well replicated in both phenotypes[43-49]. However, the gene(s) and functional allele(s) within the *HLA* locus driving the association to these phenotypes have not been identified, nor is the reason for the seemingly outsized strength of association to the *HLA* locus known.

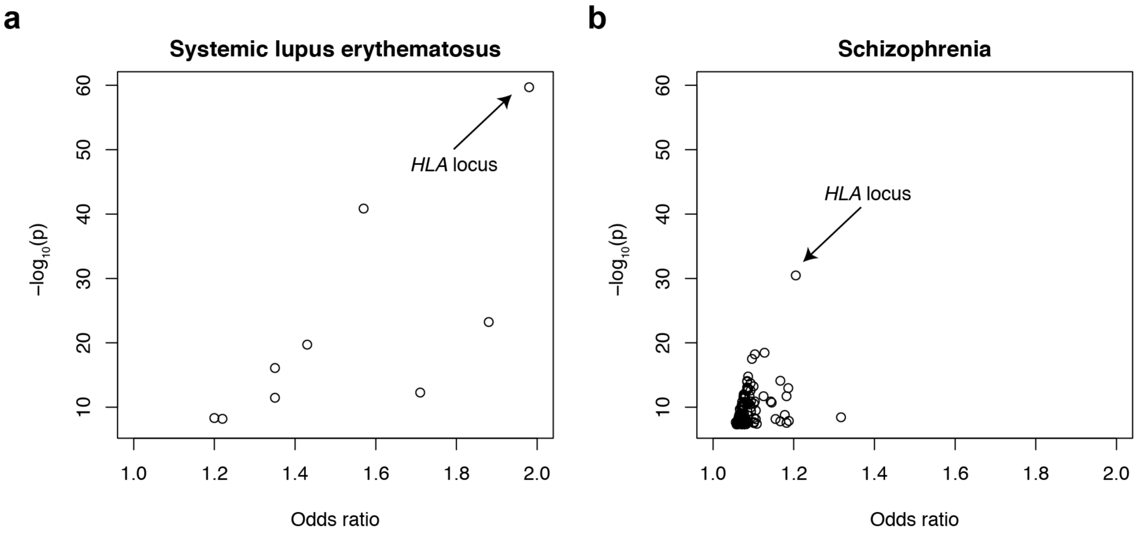


Figure 1.1. The *HLA* locus harbors the variant(s) most strongly associated to SLE and schizophrenia. -log$_{10}$(p) (strength of association) is plotted against odds ratio (effect size) for the risk allele of the most strongly associated variant from each locus that surpasses the threshold for genome-wide significance, 5x10^{-8} in **(a)** SLE[43] and (**b**) schizophrenia[22].

The *HLA* locus is an extensively investigated region of the human genome, owing to its interesting and complex population genetic properties, and its influence on dozens of autoimmune and infectious disease phenotypes. The extended *HLA* locus spans approximately 8 megabases on human chromosome 6, from 26 Mb to 34 Mb. Although the region is referred to as the '*HLA* locus', it contains hundreds of genes, only a small fraction of which encode for classical *HLA* molecules that are involved in presenting antigens to cells of the immune system[50]. Indeed, members of diverse gene families, including 3 complement genes – *C4, complement factor B* (*CFB*), and *C2* – and genes encoding tRNAs, histones, and olfactory receptors, make up the majority of genes in this locus.

The *HLA* locus is characterized by strong and long-range LD, whereby the genotypes of variants at distances of even 1-2 Mb or greater are strongly correlated with each other[51]. As a result, markers spanning a large number of genes can have similar strengths of association to phenotypes. The *HLA* locus is also highly polymorphic, with at least eight common haplotypes[52, 53]. The extensive, long-range LD, high gene- and polymorphism-density have all contributed to the difficulty in fine-mapping associations of this locus to phenotypes.

The association of the *HLA* locus to SLE spans a few dozen genes, over approximately a 2 Mb distance (**Figure 1.2a**). The locus appears to harbor multiple independent effects on SLE[44, 54] and classical alleles of the *HLA-DRB1* gene, *DRB1*03:01* and *DRB1*15:01* have been consistently implicated in SLE[55, 56]. In a joint analysis of classical *HLA* alleles and SNPs in non-*HLA* genes, the association to the *HLA* locus was proposed to be best explained by a combination of multiple alleles in *HLA* genes as well as in non-*HLA* genes[54]. The *C4* locus, in particular, has also been extensively investigated as a risk factor for SLE, although the contribution of the full extent of structural variation in *C4* has not been studied in the manner described in Chapter 3. While complete deficiency of *C4* (rare) is one of the strongest risk factors for SLE[57], the results with respect to the contribution of common forms of copy number variation in *C4* have been conflicting. Some studies have reported an association to SLE[58, 59], whereas others have concluded that *C4* does not contribute to SLE independently of *HLA* classical alleles[60] and it remains unclear whether the apparent association of *C4* to SLE is due to LD with variation in *HLA* classical genes.

The association of the *HLA* locus to schizophrenia (**Figure 1.2b**), as with SLE, is complex and unsolved, and is characterized by the following properties. First, the strength of the association to this locus is much greater than to other loci identified from

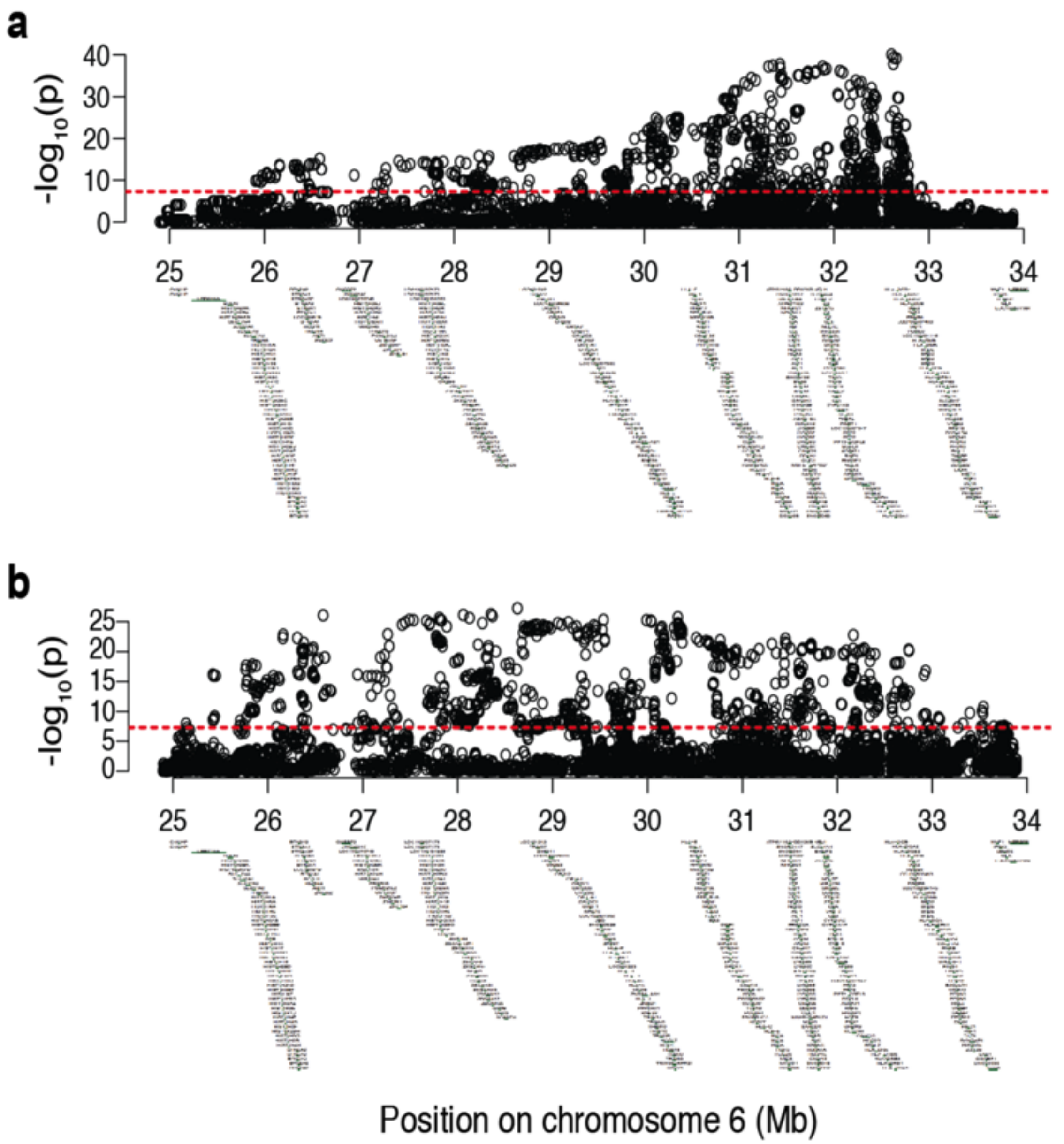


Figure 1.2. Association of the extended *HLA* locus to (**a**) SLE and (**b**) schizophrenia. The SLE association data are from 2,390 cases and 8,707 controls of European ancestry, and the schizophrenia data are from 28,799 schizophrenia cases and 35,986 controls (described in further detail in Chapters 3 and 4). The dashed red line represents the genome-wide significance threshold (p = 5x10^{-8}). The annotations below the x-axis indicate the genes within the extended *HLA* locus, with at least dozens of genes within the span of the association signals in each phenotype.

GWAS (Figure 1.1b). Second, the association signal spans several megabases and dozens to hundreds of genes. Third, no single variant within the locus is sufficient to explain the entire signal – there appear to be multiple, independent effects, as in SLE. Fourth, the location of the most highly associated variant within the *HLA* locus across multiple studies has spanned more than 3 Mb[35, 47, 48, 61].

The association of the *HLA* locus to schizophrenia has been a topic of investigation and discussion since 1974[62]. Dozens of studies have tested for association of classical alleles of *HLA* genes to schizophrenia[63], with some reporting associations to specific alleles. However, such associations could be the result of LD to variants in non-*HLA* genes and even genes that do not function in the immune system. Furthermore, many of these studies were plagued by the same problems that characterized many pre-GWAS candidate gene association studies – small sample sizes, lack of sufficiently stringent statistical thresholds, spurious associations due to population stratification, and questionable genotyping accuracy[64]. Nevertheless, consistent associations to the *HLA* locus, if not to specific alleles within *HLA* genes, have been found by genome-wide association studies as well[35, 47, 48, 61].

Some of the hypotheses for why a neuropsychiatric phenotype associates to a locus that is enriched for immune genes have focused on the possible involvement of the classical *HLA* molecules. The higher risk of schizophrenia in individuals born during late winter or early spring months[65], when there is an increased occurrence of certain infections, has led to the notion that schizophrenia may have an infectious trigger. Prenatal infections have been associated with an increased risk of schizophrenia[66], and it has been hypothesized that the immune response to an infection during susceptible periods of gestation, modulated by genetic variation in *HLA* genes, may influence neurodevelopment and predispose to later developing schizophrenia. The hypothesis that schizophrenia may be an immune-mediated disease has been supported by the increased risk of schizophrenia in individuals with a history of autoimmune diseases[67], by reports of auto-antibodies against neurotransmitter receptors in some patients[68, 69], and by studies investigating levels of inflammatory cytokines in patients with schizophrenia[70]. More recently, a role for classical *HLA* molecules in the developing nervous system has been appreciated[71-74], suggesting that their potential involvement in schizophrenia may be independent of their function in the immune system.

A few studies have specifically investigated the *C4* locus for its contribution to schizophrenia[75-77]. However, these studies were performed without consideration of variation in the *HLA* locus outside of *C4* (which, because of the extensive LD in the region, is critically necessary to evaluate), their samples sizes were considerably (1-2 orders of magnitude) smaller than current genetic studies of schizophrenia, and the results were conflicting.

Complex forms of genome structural variation are not well understood

As recently as a decade ago, the extent to which genome structural variation – variation in copy number, length, orientation, genomic location of DNA segments – at sub-microscopic scales contributes to germline genetic variation among humans was unknown. Early studies of copy number variation (CNV)[78, 79] suggested that human genome structure is more dynamic than previously appreciated, and since then, we have learned that copy number variation affects as many nucleotides in the human genome as single nucleotide variation[80], it is heritable[81], and can contribute to disease risk[37, 40, 82, 83].

Simple forms of structural variation, such as bi-allelic deletions (**Figure 1.3a)** and duplications, in which a locus segregates in only two different forms, have been well characterized. Existing approaches for measuring copy number such as quantitative

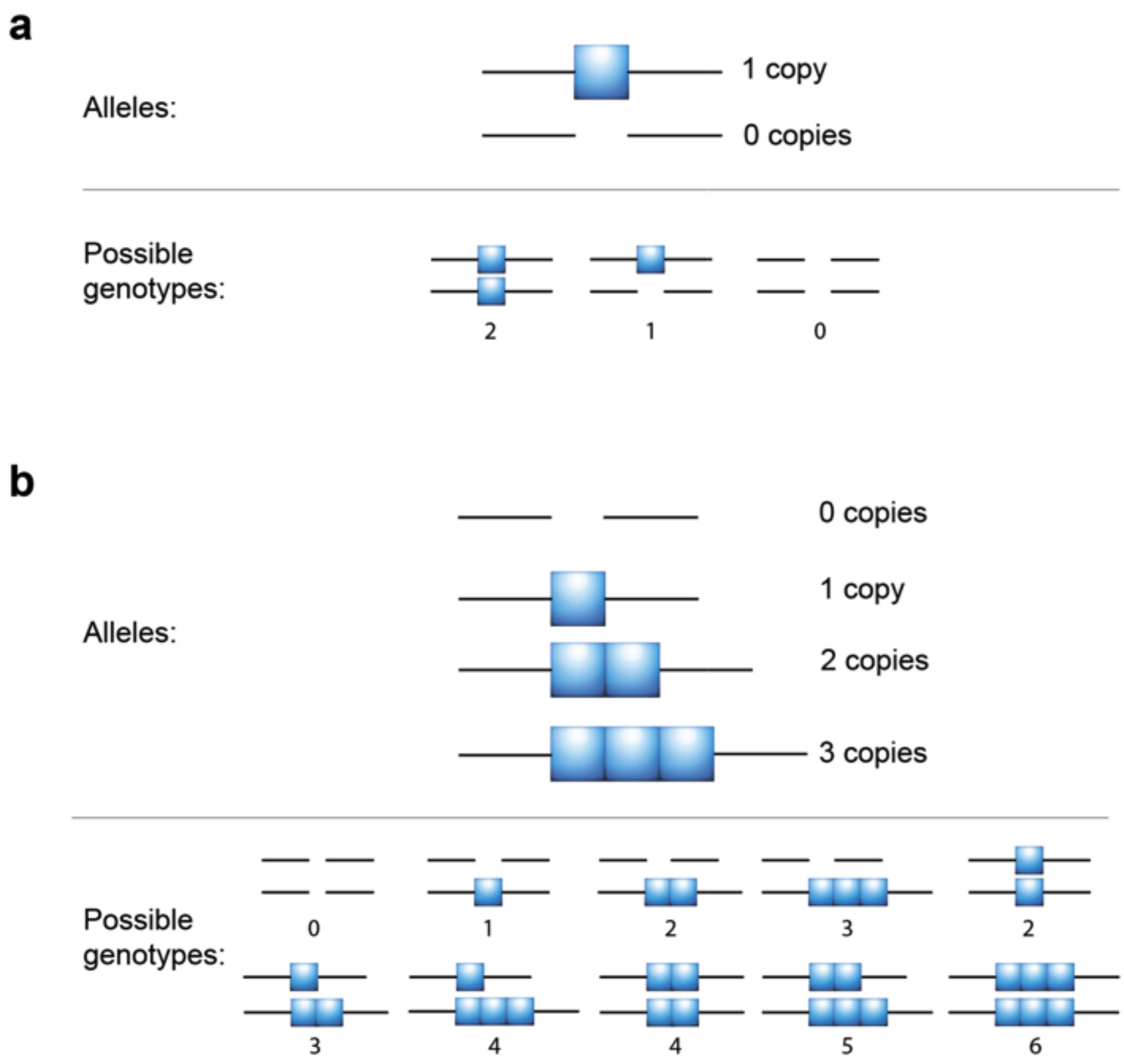


Figure 1.3. A schematic example of (**a**) simple copy number variant (CNV) and (**b**) a complex CNV. In this example, the simple CNV has two alleles and three possible genotypes; the complex CNV has four alleles and ten possible genotypes. The number below the genotypes indicates the copy number that would be measured per diploid genome, given each genotype. Note that, for a complex CNV, multiple combinations of alleles can give rise to the same diploid copy number.

PCR (qPCR) and oligonucleotide arrays have been successfully used for molecular analysis of simple forms of copy number variation[84]. Simple CNVs are part of catalogs of human genetic variation[85, 86], and their population genetic properties have been described – most common deletions are ancestral mutations that segregate on specific SNP haplotypes[84, 87]. In addition, the contribution of both rare and common forms of simple CNVs to clinical phenotypes has been investigated[37, 40, 82, 83].

In contrast, structurally complex loci are not well understood. We define structurally complex loci as those segregating in more than two (and often many more than two) structural forms. A locus can be structurally complex if multiple structural features such as copy number variation, variation in length of the locus, and multiple paralogous forms are present at the same locus (as is the case with *C4,* described in Chapter 2). Even when the only structurally variable feature of a locus is its copy number, multi-allelic variation in copy number can give rise to structural complexity (**Figure 1.3b**). Without molecular or statistical methods for characterizing such loci accurately, they have not been characterized in catalogs of genome variation and their contribution to clinical phenotypes has not been extensively investigated. Even in cases where association between complex CNVs and clinical phenotypes has been evaluated, lack of reproducibility has raised doubts about the accuracy with which these CNVs were measured[88-91].

The challenge in molecular ascertainment of multi-allelic CNVs is acknowledged in current literature[92, 93]. Distinguishing between a wide range of copy numbers such as 0-10 is a difficult signal-to-noise problem that approaches such as qPCR are often inadequate for. For example, while qPCR may be able to readily distinguish between 0, 1, 2, and 3 copies, distinguishing between 4 and 5 copies is more challenging as it represents a smaller fold-change in copy number that requires greater precision and sensitivity. In addition, determining allelic states of multi-allelic CNVs can be complex because multiple combinations of alleles can give rise to the same diploid copy number that is measured in a molecular assay (**Figure 1.3b**). To this end, we describe, in Chapter 2, molecular and statistical approaches that enable accurate measurement of copy number of multi-allelic CNVs and determination of their allelic states.

Several factors motivate the study of complex forms of genome structural variation. First, multi-allelic CNVs affect the dosage of hundreds of genes[94], and have been estimated to be five times more likely to contain protein-coding genes than simple deletion CNVs[95], suggesting that they may be a functionally important class of genetic variation. Second, because of the wider range of variation that affects loci segregating in multiple structural forms, the effect that such loci can have on phenotypes (and the power to detect association to them) has the potential to be greater. Third, by definition of structurally complex loci segregating in more than two alleles, their correlation to individual SNPs, which are bi-allelic, can be only partial. This has two implications: (i) to the extent that it is poorly tagged by individual SNPs that are tested for association in GWASs, such variation is interrogated neither directly nor indirectly in current association studies – accurately genotyping complex structural variants in GWAS data sets therefore has the potential to generate novel associations; and (ii) the partial correlation between complex CNVs and SNPs yields greater power in analyses of conditional association to evaluate the primary source of the association signal, which for regions like the *HLA* locus that are characterized by extensive LD, is especially valuable. Fourth, an association of a phenotype to a series of multiple structural alleles in the same gene provides confidence in the involvement of that gene in a phenotype (as we will demonstrate in Chapters 3 and 4).

C4 **functions in the complement pathway, which has dual roles in the immune and nervous systems**

The complement system derives its name from its ability to complement other components of the immune system in mounting an immune response. It consists of more than thirty proteins that protect against damage by pathogens, immune complexes, and cellular debris. There are three pathways to activation of the complement cascade, each involving different types of triggers – classical, alternative, and lectin pathways. The classical pathway is triggered when antibodies bind to microbes or other antigens, and the first component of the classical pathway, C1q, recognizes this immune complex. The alternative pathway is triggered when complement proteins are directly activated by antigens on the microbial surface. The lectin pathway is activated when mannose-binding lectin (a plasma protein) binds to mannose residues on microbes. While the early steps of activation differ across the three pathways, they all converge on the proteolytic activation of the central component of the pathway, C3, into its major fragment, C3b.

The complement cascade has three primary functions in host defense. First, C3b serves as an opsonin – it coats microbes and antigens and promotes their binding to C3b receptors, such as complement receptor 1 (CR1), on phagocytic cells. Second, activated fragments of complement such as C3a and C5a serve as chemo-attractants, which recruit inflammatory cells to the site of complement activation. Third, if the activation of complement progresses to its terminal stage, a membrane attack complex (MAC) is formed, which forms pores on cell membranes, resulting in osmotic lysis and cell death.

C4 plays an important role in the classical pathway of the complement system (**Figure 1.4**). Upon activation by a complex of C1 proteins, it forms an enzymatic complex, together with the activation product of C2. This complex, called the C3 convertase, produces C3b from C3. There are two primary activation fragments of C4, C4a and C4b (not to be confused with *C4A* and *C4B,* which refer to different paralogous

Figure 1.4. Classical complement pathway. The classical complement pathway is triggered by the binding of C1q to antibodies, which are in turn bound to antigens. The C1 complex activates C2 and C4, and their activated fragments form the C3 convertase, which binds covalently to the antigen. The C3 convertase activates the central component of complement, C3. C3b covalently binds the antigen and serves as an opsonin to enhance phagocytosis by cells that express receptors for C3b. In addition, C3b also forms part of the C5 convertase, which activates C5. The late steps of the complement pathway lead to formation of the Membrane Attack Complex (MAC), which inserts into cell membranes, causing osmotic lysis and cell death. In studies detailed in Chapter 4, we investigated the expression, in *postmortem* human brains, of many of the genes shown in this figure – *C1q*, *C2*, *C3*, *C4*, and *CSMD1*.

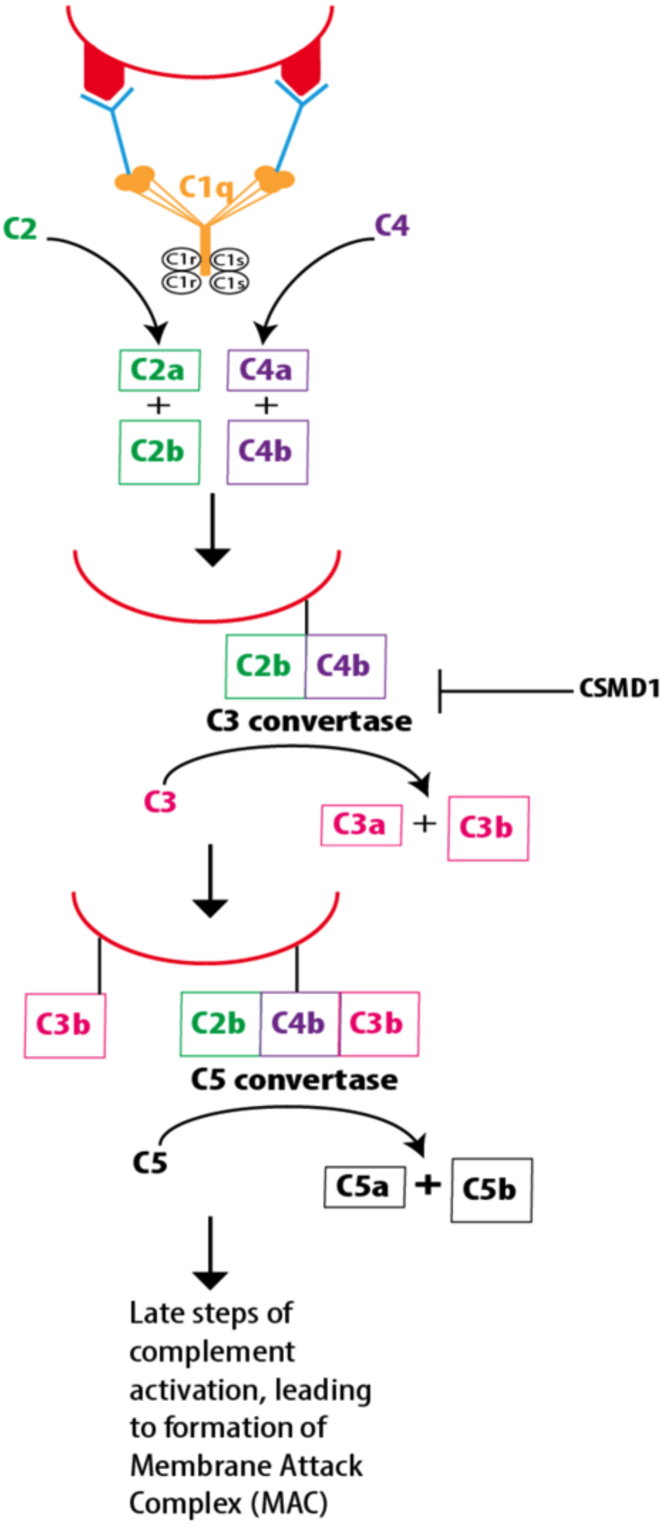


Figure 1.4 (continued)

forms of the *C4* gene, described in Chapter 2). C4a serves to stimulate inflammation and C4b, in addition to being a component of the C3 convertase complex, may also serve as an opsonin[96]. C4 also plays a role in maintaining immune tolerance by regulating elimination of self-reactive B cells[97].

Membrane-bound as well as plasma proteins of the complement system serve important regulatory roles in preventing damage to the host organism that may otherwise result from the activation of complement. One of these regulatory complement proteins is CSMD1 (CUB and Sushi multiple domains 1), an inhibitor of the classical complement pathway[98] that promotes degradation of C4b and C3b[99]. C4 binding protein is another regulatory protein, and it binds to C4[100] and interferes with the assembly of the classical pathway C3 convertase.

The complement pathway may be involved in the pathogenesis of SLE at multiple levels, with both beneficial and detrimental effects[101]. Deficiency of the early components of the complement pathway such as *C1q* and *C4* have been causally implicated in the development of SLE[102], consistent with the fact that these complement components play a role in the clearance of necrotic debris[101] (thereby preventing it from acting as a source of auto-antigens) and in maintenance of self-tolerance[103]. However, complement may also be involved in mediating tissue damage – auto-antibodies can trigger activation of the classical complement pathway, leading to inflammation and cell injury. Levels of complement proteins in the serum are used clinically to track disease activity and severity in SLE, and patients typically have low levels of complement proteins. However, this observation alone does not clarify the role of complement in SLE as low levels could be a result of a genetic deficiency and/or consumption of complement during disease progression.

While functions of the complement pathway in the immune system have been appreciated for decades, an important role for the pathway in the nervous system has come to light in just the past few years. In 2007, Stevens and colleagues showed that *C1q* and *C3* are necessary for synaptic pruning, a process by which excessive and under-utilized synapses are eliminated during development of the central nervous system[104]. More recently, this process was shown to be mediated by microglia in a complement- and neuronal activity-dependent manner[105]. Genetic studies have also pointed to a role for complement in the nervous system – genome-wide association studies of Alzheimer's disease have identified associations to *CR1*, which encodes a receptor for C3b and C4b, and *CLU*, which serves as an inhibitor of the complement pathway[106] (among its other functions).

The dual roles of the complement pathway in the immune and nervous systems raise the question of what the contributions of genes in this pathway are to disorders of the respective systems. More specifically, the strong, yet unexplained nature of the association of the *HLA* locus to SLE and schizophrenia – among other factors, discussed in Chapters 3 and 4 – motivates an investigation of whether *C4*, within the *HLA* locus, contributes to these phenotypes. Such an investigation, however, would require the development of new approaches to characterize complex structural variation and to make such variation accessible to association analysis. We proceed in the following chapter by describing these approaches.

Chapter 2

Molecular and population genetics of the human *complement component 4* (*C4*) locus

The human *C4* gene exhibits a complex form of structural variation

Complement component 4 (*C4*), in the *HLA* locus, is one of the most polymorphic genes in the human genome, harboring multiple forms of genetic variation (**Figure 2.1**). First, there is variation in the number of copies of the *C4* gene (**Figure 2.1a**), with a diploid genome typically containing 2-6 copies. Second, a length polymorphism gives rise to a long (*C4L*) and short (*C4S*) form of the gene (**Figure 2.1b**). This polymorphism is due to an insertion of a ~6.4 kb Human Endogenous Retroviral (HERV) sequence that is present in intron 9 of *C4L* but absent in *C4S*[107]. The HERV element is present in an opposite orientation to that of the *C4* gene, and because it has accrued many protein-disrupting mutations since the time of insertion, it is not thought to encode functional proteins. However, the HERV element contains sequences that may be transcriptionally active, such as TATA boxes and an SV40-type enhancer sequence. Third, sequence variants clustered in exon 26 define two paralogous forms of the gene, *C4A* and *C4B* (**Figure 2.1c**). Despite the high sequence similarity between the two paralogs, they have different biochemical properties[108]. C4A has a higher affinity for amino groups (such as those found in proteins, including immune complexes), whereas C4B has higher affinity for hydroxyl groups (such as those found in bacterial carbohydrate antigens). All four combinations of *C4* structural features exist – *AL*, *AS*, *BL*, and *BS* – and each of these forms varies in copy number.

Variation of this nature has not been investigated in genome-wide association studies (GWAS) or exome sequencing studies, so its relationship to phenotypes is not well characterized. Such variation is also absent from catalogs of human genetic variation such as the HapMap[24] and 1000 Genomes Projects[86], and we do not understand its population genetic properties.

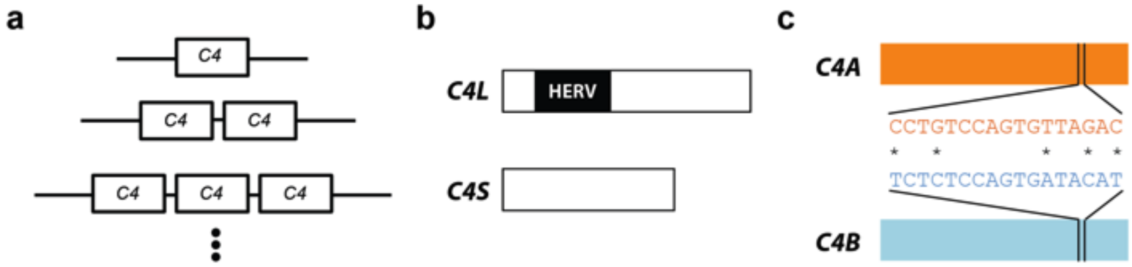


Figure 2.1. Structurally variable features of the *C4* locus. (**a**) The number of *C4* genes is variable, with typically 1-3 per chromosome. (**b**) A Human Endogenous Retroviral (HERV) sequence is present in the long form of the *C4* gene (*C4L*) and is absent in the short form (*C4S*). (**c**) There are two paralogous forms of *C4*, *C4A* and *C4B*, which are defined by nucleotide differences in exon 26 of the gene.

Strategy for characterizing complex genome structural variation

Our approach for characterizing structural variation in *C4* is summarized in **Figure 2.2** and is described in detail in the remaining sections of this chapter**.** We first determined the copy number of each *C4* structural form (*AL, AS, BL,* and *BS*) per diploid genome, in a set of father-mother-offpsring trios of European ancestry (HapMap CEU sample). Next, we inferred the copy number of each *C4* form *per chromosome*, based on diploid copy number, allele frequencies, and inheritance in trios. We then defined *C4* structural haplotypes – combinations of *C4* structural forms that segregate together on the same chromosome – based on co-transmission of these forms in trios. We jointly phased these *C4* structural haplotypes with haplotypes of SNPs in the *HLA* locus in order to evaluate the relationship between the two. By creating these integrated haplotypes of *C4* structural variation and SNPs, we generated a reference panel, which we then tested for its ability to impute (statistically predict) *C4* structural states using existing SNP genotype data.

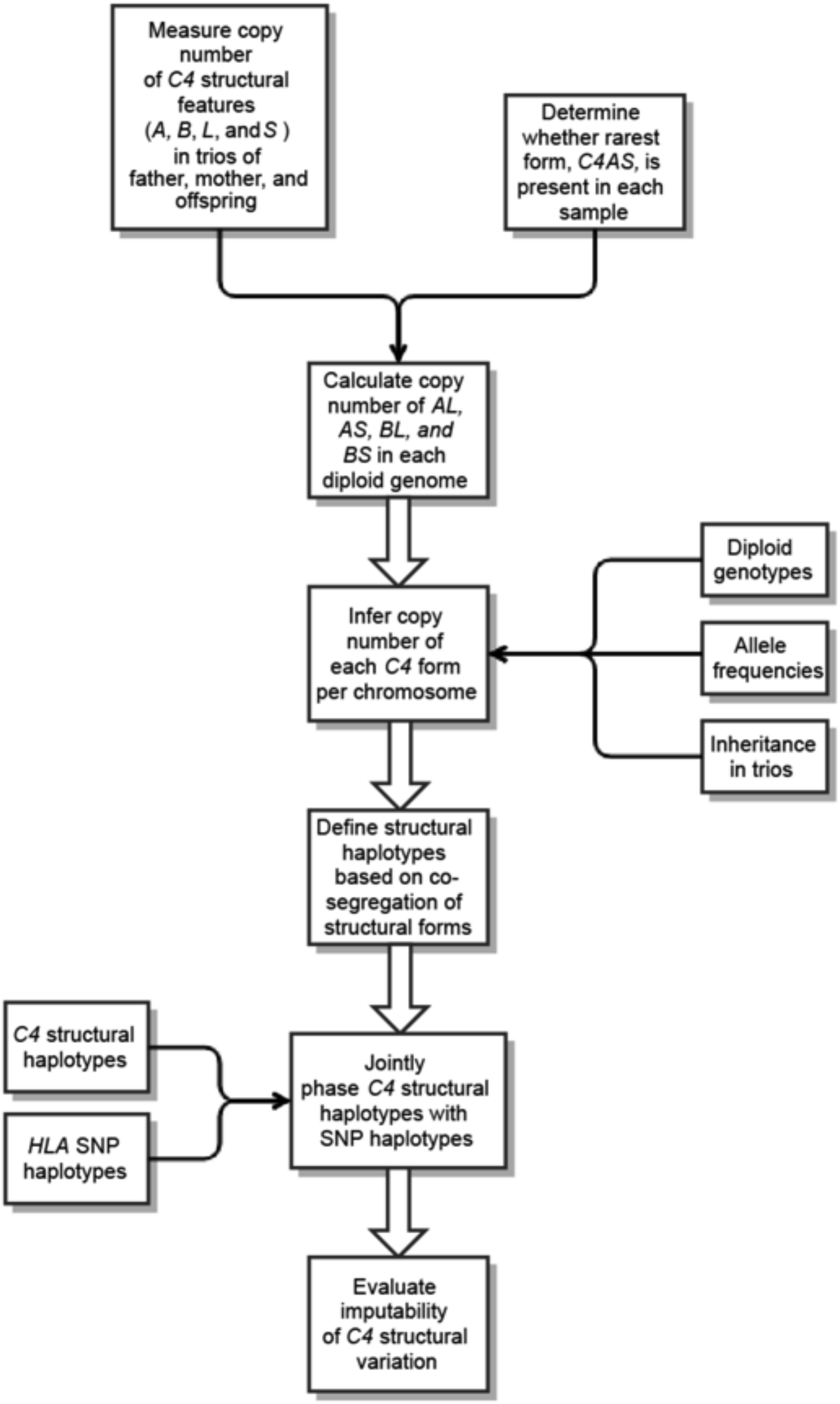

Measure copy number of *C4* structural features (*A*, *B*, *L*, and *S*) in trios of father, mother, and offspring
Determine whether rarest form, *C4AS*, is present in each sample
Calculate copy number of *AL*, *AS*, *BL*, *and* *BS* in each diploid genome
Infer copy number of each *C4* form per chromosome
Diploid genotypes
Allele frequencies
Inheritance in trios
Define structural haplotypes based on co-segregation of structural forms
C4 structural haplotypes
HLA SNP haplotypes
Jointly phase *C4* structural haplotypes with SNP haplotypes
Evaluate imputability of *C4* structural variation

Figure 2.2. Strategy for characterizing complex structural variation in *C4.* Each of these steps is discussed in detail in the text.

A droplet-based PCR approach enables precise molecular analysis of the *C4* locus

Existing methods for measuring copy number variation are inadequate for molecular analysis of C4 *structural variation*

Oligonucleotide microarrays have been used successfully to measure copy number of simple CNVs[80, 84], in which the amount of signal generated by genomic DNA segments binding to probes on a microarray is used to infer copy number. We considered the ability of these arrays to be used for measuring copy number of *C4* structural features. However, commonly used genotyping arrays lack probes in informative locations to detect specific *C4* features (**Figure 2.3a**). We then evaluated whether quantitative PCR (qPCR), another approach commonly used to measure copy number variation, can yield integer copy numbers for the *C4* CNVs. We used reaction conditions and assays that were based on qPCR assays previously reported to yield accurate copy number measurements of the *C4* CNVs[109], and we compared the measurements to results from Southern blot analysis (considered to be the gold standard for determining *C4* copy number), of the same samples performed by another study[110]. Although the measurements from qPCR were correlated with the expected copy number, they did not resolve into clusters, and confident assignment of copy number genotypes was not possible for most of the samples (**Figure 2.3b**).

While Southern blot analysis could be used to determine copy number, it would not provide the throughput needed to genotype *C4* CNVs in a large number of samples. Therefore, we turned to a new approach, droplet-digital PCR, for measuring copy number of *C4* structural features.

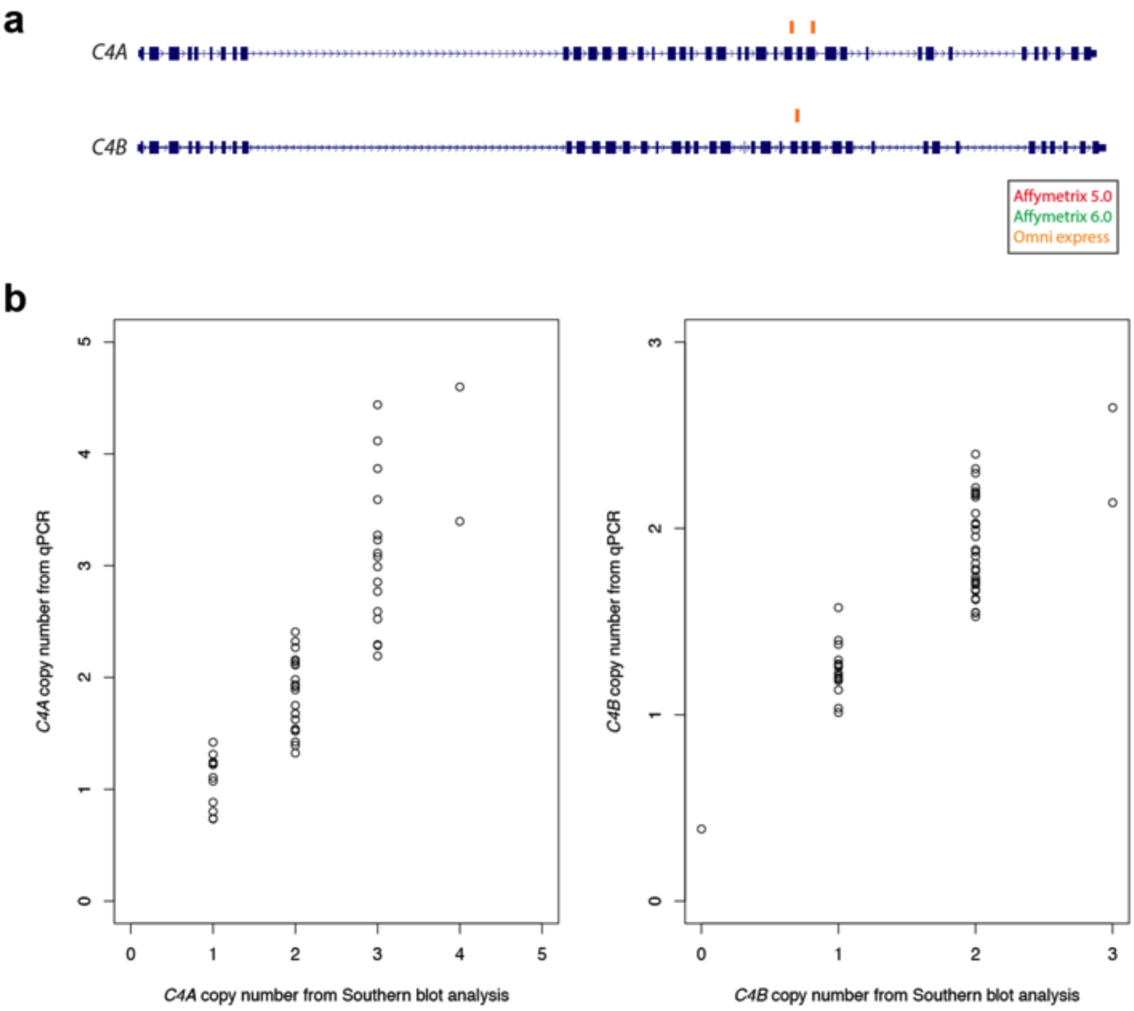


Figure 2.3. (**a**) Location of probes within *C4A* and *C4B* genes on commonly used oligonucleotide microarrays, Affymetrix 5.0, Affymetrix 6.0, and Illumina Human OmniExpress (the former two do not contain any probes within the *C4* gene). (**b**) Copy number estimates for *C4A* (left) and *C4B* (right) from qPCR based on triplicate measurements plotted against copy number determined by Southern blot analysis for the same samples by Fernando et al[110]**.**

Droplet-digital PCR (ddPCR) *yields integer copy number measurements of the* C4 *CNVs*

We measured copy number of each *C4* structural feature – *A, B, L*, and *S* – using a new molecular approach, droplet-digital PCR (ddPCR)[111] (**Figure 2.4**) in 180 individuals from the HapMap CEU sample of European ancestry[24]. ddPCR involves partitioning of a standard real-time PCR reaction mixture consisting of primers and fluorescent probes that are directed to the target (copy number variable) locus as well as to a copy number-invariant reference locus (**Figure 2.4a**), into thousands of nanoliter-sized droplets. A restriction digest of the DNA prior to the generation of droplets is performed to separate multiple copies of the same locus that are in tandem (as is the case with the *C4* locus) and allow independent segregation of the template into droplets. Droplet generation produces micro-droplets that are uniform in size (**Figure 2.4b**) and that compartmentalize the PCR reaction, such that each droplet contains zero, one, or very few copies of the template from each locus. The higher the copy number of a locus, the greater the number of droplets that contain a PCR template for it. PCR amplification occurs within the droplets and an end-point fluorescence measurement is taken for each of the two fluorophores from each droplet. This resolves the droplets into four classes: (i) FAM-VIC-, (ii) FAM-VIC+, (iii) FAM+VIC-, and (iv) FAM+VIC+ (**Figure 2.4c**). Variation in copy number is reflected in the fraction of droplets that are positive for the template, as opposed to differential kinetics of amplification as in qPCR. Also in contrast to qPCR, ddPCR provides absolute quantification because applying Poisson correction to the fraction of positive droplets (to account for droplets that contain more than one copy of the template) yields an estimate of the number of copies of the template per droplet. Once the absolute number of copies of the target (CNV) locus and reference (two-copy) locus are calculated, the ratio between the two is used to determine absolute copy number of the CNV locus per diploid genome.

Applying this approach to the *C4* locus, we were able to resolve integer copy numbers of each of the *C4* structural features, across a range of 0-5 copies per diploid

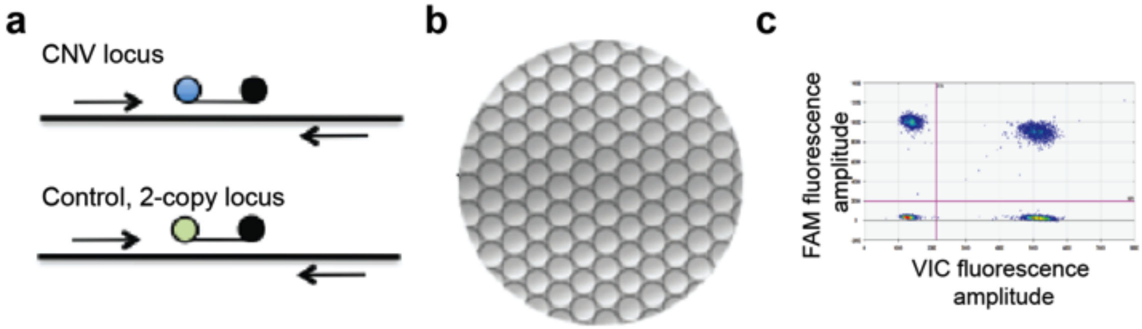


Figure 2.4. Droplet-digital PCR (ddPCR)[111]. (**a**) Each assay involves simultaneous PCR amplification of two loci – the CNV locus and a control, two-copy locus. (**b**) The reaction mixture is partitioned into thousands of nanoliter-sized droplets (Image source: Bio-Rad Laboratories). (**c**) Following PCR, FAM and VIC fluorescence are analyzed in each of the microdroplets by a droplet reader, and four classes of droplets are observed: (i) FAM-VIC-, (ii) FAM-VIC+, (iii) FAM+VIC-, (iv) FAM+VIC+. The fraction of droplets that are positive for the template is Poisson-corrected to estimate the absolute number of copies of the template in the reaction. The ratio between the number of template copies from the CNV locus and the reference locus is used to determine absolute copy number of the CNV locus per diploid genome.

genome (**Figure 2.5**). In a set of 180 HapMap CEU samples, each *C4* feature had multi-allelic variation in copy number, with a range of 0-5 copies per diploid genome of *C4A*, 0-3 copies of *C4B*, 1-5 copies of *C4L*, and 0-3 copies of *C4S*. Given that ddPCR is a new approach, we sought to evaluate the accuracy of the copy number measurements generated using this method. First, we checked for internal consistency across the four copy number measurements for each sample – if one of the measurements was incorrect, then the expectation that the number of copies of *C4A* + *C4B* = *C4L* + *C4S,* would not be met. Indeed, across all of the samples we typed, this relationship was confirmed. However, if a sample had more than one incorrect measurement (e.g., if the

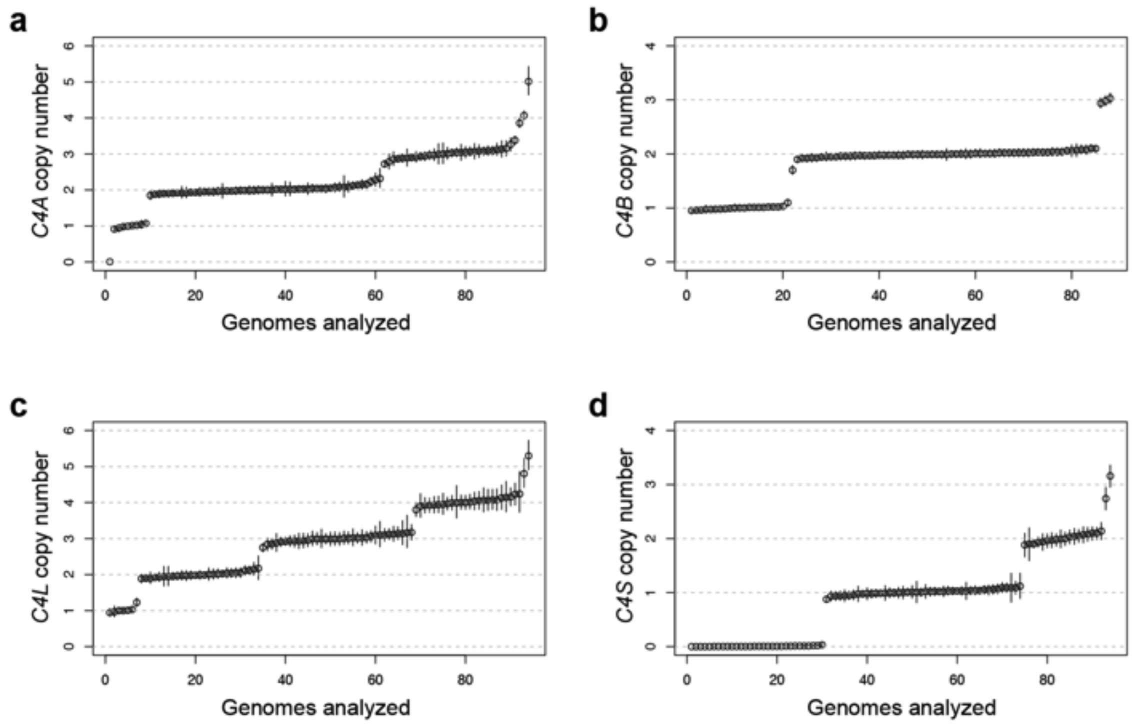


Figure 2.5. Copy number of *C4* structural features per diploid genome as determined by ddPCR. Representative results from measuring copy number of (**a**) *C4A*, (**b**) *C4B*, (**c**) *C4L*, and (**d**), *C4S* in HapMap CEU samples are shown. The genomes analyzed (x-axis values) are sorted from left to right by their copy number measurement, and the absolute copy number, with 95% Poisson confidence intervals, is on the y-axis. Note not only that there is separation between samples with different copy numbers, but also that the point estimates fall at an integer value for the vast majority of the samples.

copy number of both *C4A* and *C4L* in a given sample were incorrectly measured to be one less than the actual copy number), then it is possible for the data to still appear internally consistent. Therefore, we evaluated the accuracy of our ddPCR results by also comparing them to data from orthogonal approaches. First, we compared our copy number measurements for *C4A* and *C4B* to those generated by Fernando *et al*. using Southern blot analysis of the same samples[110]. Across 89 samples that had measurements from both approaches, there was 100% concordance for both *C4A* and *C4B* copy number. We next compared the measurement of *C4L* copy number using ddPCR to that calculated from an approach based on read-depth analysis of whole-genome sequence (WGS) data[85]. This computational approach, Genome STRucture in Populations (Genome STRiP), has been used successfully to generate high-confidence measurements of copy number in another structurally complex locus[112]. Genome STRiP exploits the fact that read depth across a genomic segment is a measure of its copy number, once technical influences, uniqueness, and alignability of the local sequence are accounted for. Across 64 samples for which we generated copy number measurements using ddPCR as well as Genome STRiP (using low-coverage WGS data from the 1000 Genomes Project) 61 (95.3%) were concordant.

Long-range relationships between C4 *structural features can be inferred using ddPCR*

While the experiments described above allowed determination of the number of copies of *C4A*, *C4B, C4L*, and *C4S* per diploid genome, they did not reveal the copy number of the compound *C4* structural forms – *AL, AS, BL,* and *BS.* In other words, for a given *C4* gene, long-range relationships between the *L/S* and *A/B* defining molecular features (separated by ~5 kb in a short *C4* gene and ~11.5 kb in a long *C4* gene) would need to be determined. Our strategy was to determine whether the rarest form of the *C4* gene, *C4AS*, was present in each genome, and then use the following relationships to calculate the diploid copy number of the other three forms of the *C4* gene (*AL, BL,* and *BS)*:

Copy number (CN) of *C4AL* = (CN of *C4A*) - (CN of *C4AS*)

CN of *C4BL* = (CN of *C4L*) - (CN of *C4AL*)

CN of *C4BS* = (CN of *C4B*) - (CN of *C4BL*)

The above relationships between the copy number of different *C4* structural features hold true because any given *C4* gene is defined by its length (long or short) and its paralogous form (*A* or *B*).

We developed a ddPCR-based approach in order to determine whether a given genome contained a *C4AS* gene. This approach exploits the fact that, when assaying for two different segments of DNA (i.e., the sequences that define the *C4A* and *C4S* features), the fraction of droplets that are positive for DNA templates containing both segments (double-positive droplets) should be a function of the fractions that are positive for only one of the two, at low concentrations of input DNA. This expectation would be met if the two segments of DNA being assayed are unlinked – i.e., able to segregate into droplets independently of each other. However, under a scenario in which the *C4A* and *C4S* features were present on the same DNA template (indicating the presence of a *C4AS* gene), this would result in an enrichment of double-positive droplets.

We first performed long-range PCR to enrich for DNA templates that spanned the *L/S* and *A/B* assay sites. The forward primer was placed upstream of the HERV sequence and the reverse primer, downstream of the *A/B* defining sequence so that every form of the *C4* gene (*AL*, *AS, BL,* and *BS*) would be amplified. We then used this long-range PCR product as input into ddPCR, with one assay specifically interrogating the presence of *C4A* and the other, *C4S*. Representative results from samples without *C4AS* and with *C4AS* are shown in **Figure 2.6a** and **Figure 2.6b**, respectively. We quantified the enrichment of droplets that were double positive using a Fisher's exact test, which readily distinguished samples with a *C4AS* gene from those without (**Figure 2.6c**). In order to evaluate the accuracy of this ddPCR approach, we designed a long-range PCR assay in which the forward primer was placed upstream of the HERV

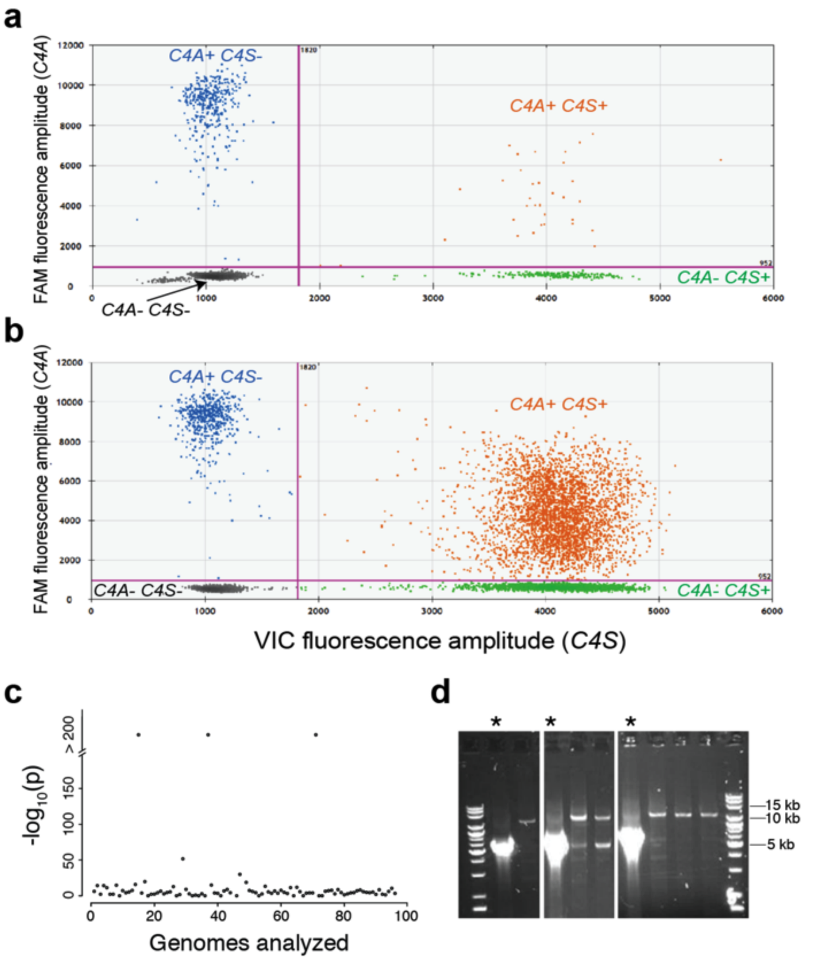


Figure 2.6. Determining whether the *AS* form of the *C4* gene is present in each genome. (**a**) A plot of fluorescence amplitudes from ddPCR for a sample without an *AS* gene, in which the two assays were *C4A* (FAM, y-axis) and *C4S* (VIC, x-axis). This sample shows no enrichment of double-positive droplets. (**b**) Plot of fluorescence amplitudes for a sample with an *AS* gene. Note the enrichment of double-positive droplets that contain templates with both *C4A* and *C4S* features. (**c**) P-values from a Fisher's exact test of non-independence of the number of droplets in each of the four quadrants. (**d**) Confirmation of the ddPCR-based results using a long-range PCR assay. The size of the expected product from a *C4AS* gene is ~5.1 kb and from a *C4AL* gene, ~11.5 kb. *, the same three samples identified using our ddPCR-based assay as containing a *C4AS* gene in (**c**).

sequence and the reverse primer was specific to the *C4A* sequence. This assay identified the same samples as containing *C4AS* as our ddPCR-based approach (**Figure 2.6d**). We used these approaches to infer the diploid copy number of each *C4* structural form in the HapMap CEU sample, and found that the number of *AL* genes varied from 0-5 in a diploid genome; *AS,* 0-1; *BL*, 0-3; and *BS*, 0-3.

The *C4* locus exists in four common and several lower-frequency structural haplotypes

In order to answer several questions about the population genetics of the *C4* locus, we needed to infer haplotype-specific copy number of *C4* structural forms (and not just their diploid copy number as determined above): What is the structure of the *C4* locus on each human chromosome 6? What are the frequencies of *C4* structural variants? What is their relationship to SNP haplotypes in the *HLA* locus? Can structural variation in *C4* be made accessible to genotype imputation, using existing SNP genotype data?

Haplotype-specific copy number of C4 *structural forms can be inferred using genotypes, allele frequencies, and inheritance patterns*

Multiple combinations of alleles can give rise to the same diploid copy number for a multi-allelic CNV. For example, if a sample has 4 copies of the *C4AL* gene in a diploid genome, this could be a result of different allelic combinations: 0+4, 1+3, or 2+2. In order to distinguish between these possibilities, we made use of allele frequency information that is implicit in the diploid genotype data and additional constraints placed by inheritance in trios (**Figure 2.7a**). We used an expectation-maximization (EM) algorithm that incorporates this information, and we applied it to each *C4* structural form (*AL, AS,*

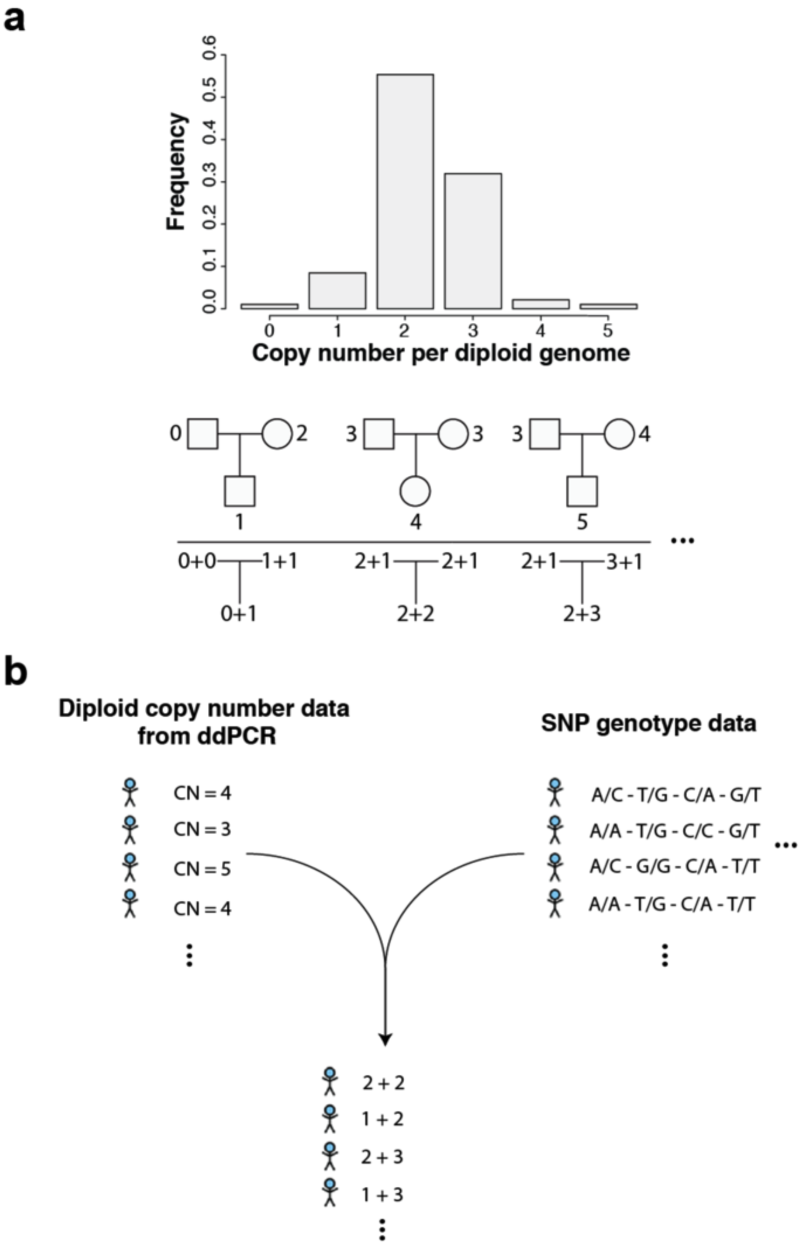


Figure 2.7. Strategy for inferring haplotype-specific copy number of *C4* structural forms. (**a**) Identifying transmitted and un-transmitted alleles using diploid copy number, estimation of allele frequencies, and inheritance patterns in father-mother-offspring trios. (**b**) An alternative approach that uses SNP backbone data to discriminate between multiple allelic combinations that can give rise to the same diploid copy number in each sample.

BL, and *BS*) separately*.* In this approach, we first enumerated the different allelic configurations that can give rise to each diploid copy number – in certain trios (e.g., a trio in which father, mother, and offspring had a copy number of 0, 2, and 1, respectively), only one configuration was possible under Mendelian inheritance, and in the rest of the trios, we made probabilistic inferences of haploid copy number. We used these inferences to estimate frequencies of each haploid allele, and then re-calculated the likelihood of each allelic combination in each trio given these allele frequency estimates. This allowed a new estimate of allele frequencies, which we then used to refine likelihoods of each allelic combination in each trio. We repeated this EM loop until the allele frequency estimates converged. Using this approach, we identified 4 alleles of *AL* copy number (0-3); 2 of *AS* copy number (0-1); 3 of *BL* copy number (0-2); and *3* of *BS* copy number (0-2).

Haplotype-specific copy number can also be inferred by integrating diploid copy number data together with SNP genotype data

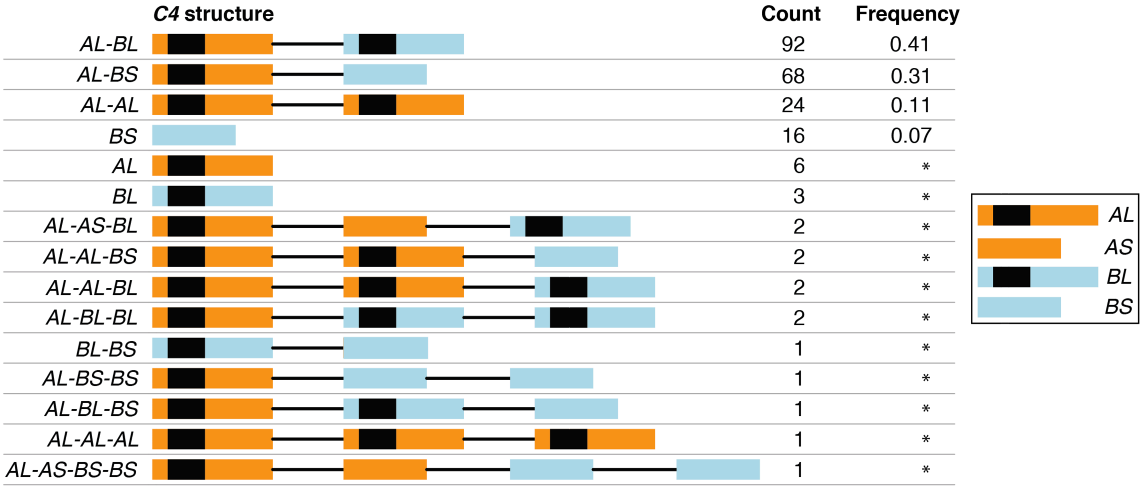
The EM approach described above uses Mendelian transmission of copy number alleles in trios as a constraint in the inference of haploid copy number. If the *C4* locus were highly mutable, it is possible for Mendelian inheritance to be violated and, in theory, not be detected when tracking inheritance in trios. We therefore developed an alternate approach to resolving haploid-copy number – one that does not assume Mendelian inheritance, but instead makes use of genotypes of SNPs surrounding the *C4* locus. We reasoned that statistical relationships between SNPs and alleles of *C4* could provide additional information that can be used to calculate the likelihood of observing each allelic combination given a diploid copy number (**Figure 2.7b**). In this approach, we assigned equal likelihoods to each genotype that could theoretically give rise to a copy number (e.g., 0+2 and 1+1 for a diploid copy number of 2). We then statistically phased these probabilistic genotypes together with SNP genotypes, using Beagle[113] (version 4), from which we obtained posterior genotype probabilities for each possible genotype. We applied this approach to the HapMap CEU sample (parents and offspring separately) and to each multi-allelic *C4* structural form (*AL*, *BL,* and *BS*). In 98.7% (163/ 165) of the samples, the maximum posterior genotype probability was > 95%. We compared the results from this approach to those from the EM algorithm and found that the results were 97.2% concordant (the alleles inferred for 108/111 unrelated individuals using the two approaches were the same across the four structural forms).

The C4 *locus segregates in multiple structural haplotypes*

By tracking inheritance of each *C4* structural form in trios, we were able to identify transmitted and un-transmitted alleles of each form. This allowed us to determine which alleles of *C4* segregated together on the same chromosome and to define *C4* structural haplotypes. Although 72 haplotypes of *C4* structural forms were theoretically possible (with 4 alleles of *AL* copy number, 2 of *AS,* 3 of *BL*, and 3 of *BS*), these structural forms segregated in only 15 haplotypes in the HapMap CEU sample (**Figure 2.8**). Of these, four (*AL-BL*, *AL-BS, AL-AL,* and *BS*) were common (5%) and accounted for more than 90% of the haplotypes, and eleven lower-frequency haplotypes accounted for the rest.

C4 **copy number is partially correlated to individual SNPs in the *HLA* locus**

Having determined both diploid copy number as well as structural haplotypes of the *C4* locus, we next evaluated the relationship between structural variation in *C4* and surrounding SNPs in the *HLA* locus (genotypes of which were available through the HapMap Project). We calculated the correlation between the dosage of the minor allele



C4 structure	Count	Frequency
AL-BL	92	0.41
AL-BS	68	0.31
AL-AL	24	0.11
BS	16	0.07
AL	6	*
BL	3	*
AL-AS-BL	2	*
AL-AL-BS	2	*
AL-AL-BL	2	*
AL-BL-BL	2	*
BL-BS	1	*
AL-BS-BS	1	*
AL-BL-BS	1	*
AL-AL-AL	1	*
AL-AS-BS-BS	1	*

Figure 2.8. *C4* structural haplotypes and their frequencies. Four common and eleven lower-frequency structural haplotypes of *C4* were identified in the HapMap CEU sample of European ancestry. Counts and frequencies of the haplotypes are shown based on 111 unrelated individuals (222 haplotypes). The order of the *C4* genes within each haplotype is not meant to represent the physical position of the genes on the chromosome; instead, the schematic only represents the structural content of each haplotype. *, frequency not well established.

of each SNP (from phase III of the HapMap Project) per diploid genome with copy number. Given the multi-allelic variation in *C4*, we expected the correlation with individual SNPs (with only two alleles) to be partial. Indeed, the maximum correlation of the genotype of any given *HLA* SNP to the copy number of *C4A*, *C4B*, *C4L*, and *C4S* was an r^2 of 0.44, 0.33, 0.44, and 0.38, respectively (**Figure 2.9**), and no SNP more than 200 kb away from the *C4* locus had an r^2 greater than 0.35 to the copy number of any *C4* structural feature. These results are in contrast to the relationships among *HLA* SNPs as well as between *HLA* SNPs and classical alleles, which are stronger and extend over a longer range[51]. The partial correlation between individual *HLA* SNPs and copy number of *C4* suggested that if *C4* contributed to phenotypes, GWASs, which test for association to single SNPs, may be underpowered to detect associations arising from copy number variation in this locus. This motivated an evaluation of whether structural variation in *C4* relates more strongly to haplotypes of multiple SNPs.

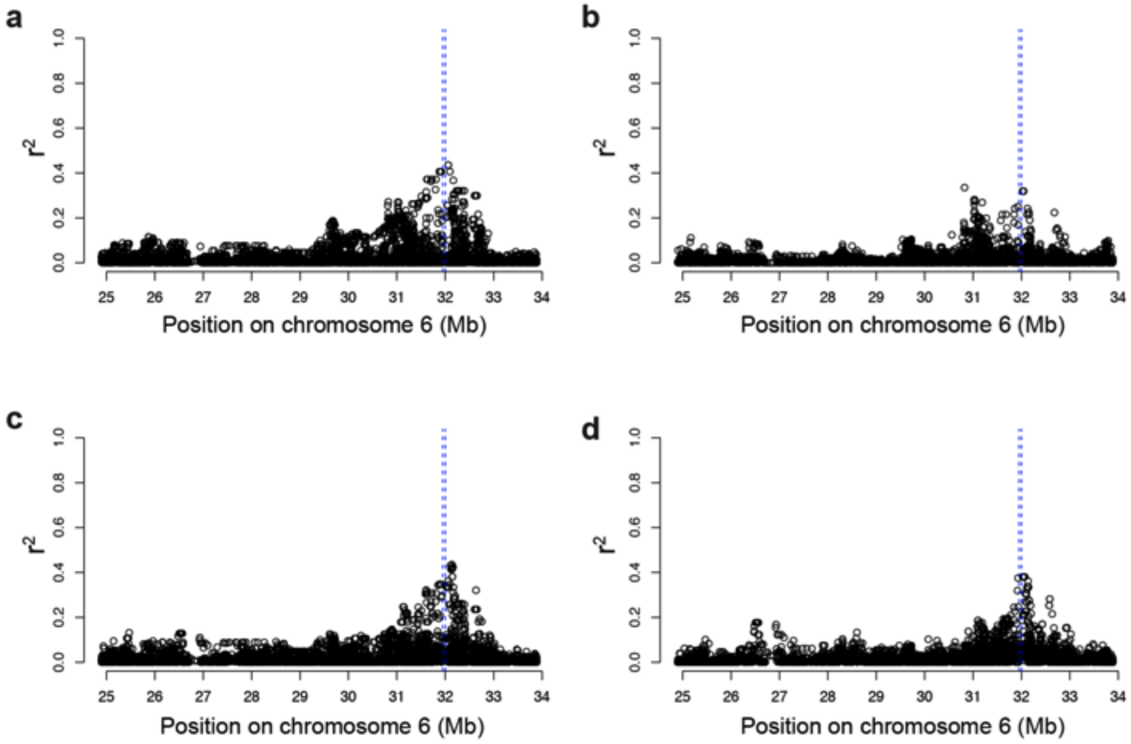


Figure 2.9. Correlation (r^2) between genotypes of individual SNPs in the extended *HLA* locus and diploid copy number of (**a**) *C4A*, (**b**) *C4B*, (**c**) *C4L*, and (**d**), *C4S* in unrelated individuals from the HapMap CEU sample. Dashed, vertical blue lines indicate the position of the *C4* locus.

Common *C4* structures segregate on multiple *HLA* haplotypes

We assembled integrated haplotypes of *HLA* SNPs and *C4* structures by jointly phasing *C4* structural haplotypes with pre-phased haplotypes of *HLA* SNPs (from 25 Mb to 34 Mb on chromosome 6) from the HapMap Project. We thereby defined the *HLA* SNP haplotype(s) on which each *C4* structure segregates. As expected, we found that no individual SNP captured this complex form of structural variation. However, we observed a strong relationship between *C4* structures and haplotypes of multiple SNPs, with a given *C4* structure segregating on 1-5 common *HLA* haplotypes and a few less-characteristic ones (**Figure 2.10**). The three most common *C4* structures each segregated on 3-5 distinct SNP haplotypes, indicating that they may have arisen multiple times among human ancestors.

We also related *C4* structures to classical alleles of *HLA-A, HLA-C, HLA-B, HLA-DRB1, HLA-DQA1*, and *HLA-DQB1,* using publicly available data for the classical *HLA* alleles[114]. The strongest relationship between any of the four common *C4* structural haplotypes and an *HLA* classical allele was between the *BS* haplotype (which was long and uniform, **Figure 2.10**) and *HLA-DRB1*03:01*. 90% of the *BS* haplotypes contained an *HLA-DRB1*03:01* allele and 69% of the *HLA* haplotypes with the *HLA-DRB1*03:01* allele also contained the *C4 BS* structure, with an r^2 of 0.59. None of the other four common *C4* structural haplotypes correlated to an HLA classical allele with an r^2 of > 0.3.

C4 structural variation can be analyzed by imputation

Based on the relationships between *C4* structures and *HLA* SNP haplotypes, we hypothesized that it is possible to analyze *C4* structural variation by imputation (**Figure 2.11**). We used the integrated haplotypes of *HLA* SNPs and *C4* structures from 111 unrelated individuals from the HapMap CEU sample that we created above as a reference panel for imputation. In order to evaluate the accuracy of inferences using this imputation approach, we performed a series of leave-one-out trials. In these trials, a different individual was removed from the reference panel in each trial, and the rest of the reference haplotypes were used to impute either copy number of *C4* structural

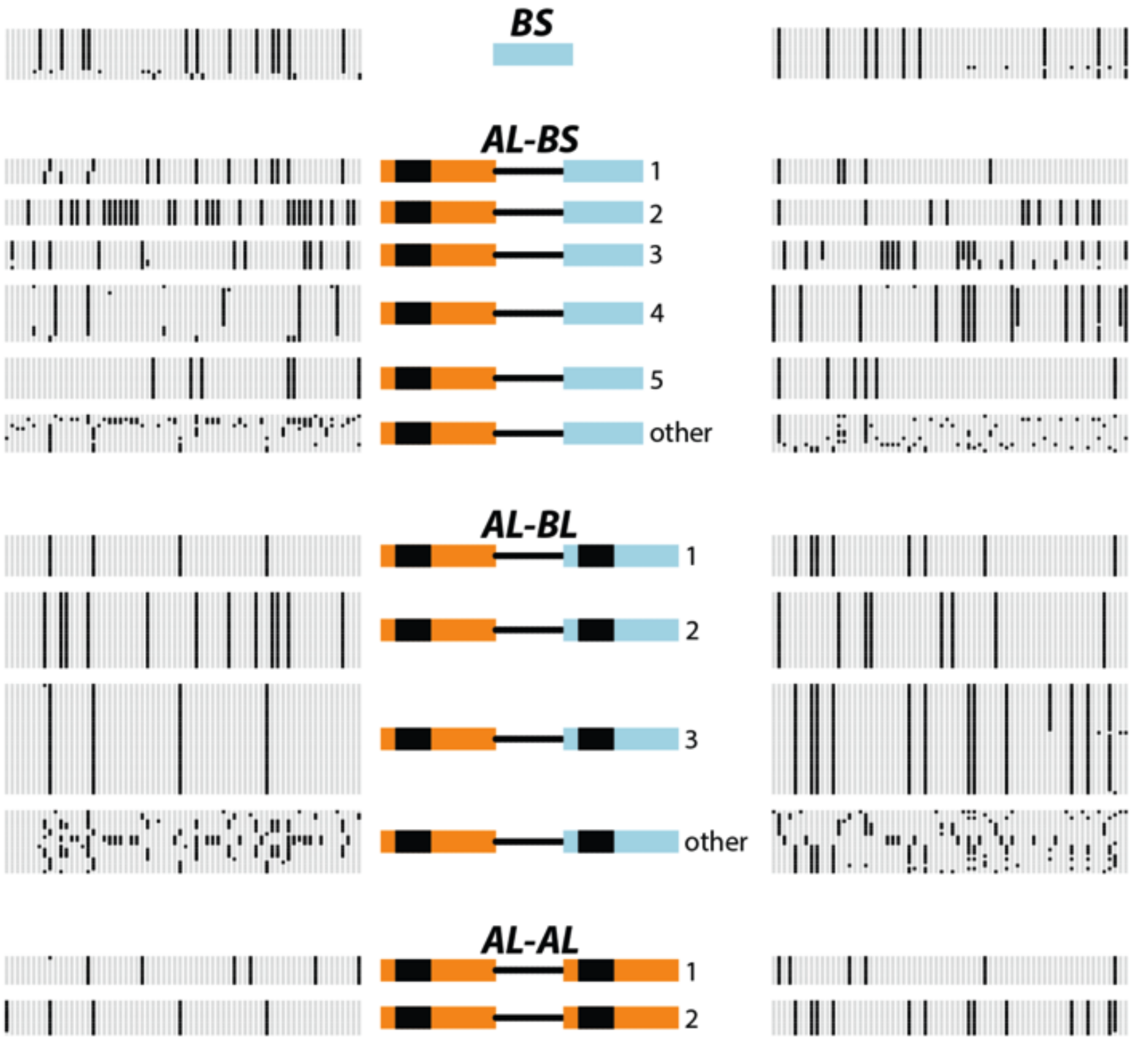


Figure 2.10. *HLA* SNP haplotypes on which the four most common *C4* structures segregate in the HapMap CEU sample. Each vertical line is a SNP, with gray indicating the major allele and black, the minor allele. The figure spans ~250 kb around the *C4* locus. Note that the three most common *C4* structures each segregate on multiple *HLA* SNP haplotypes – *AL-BS* on 5 distinct haplotypes and 1 poorly-defined set of ("other") haplotypes; *AL-BL* on 3 well-defined and 1 other set of haplotypes and *AL-AL* on 2 haplotypes. Each set of haplotypes shown in this figure is present at a frequency of at least 3.5% (at least 8 occurrences) in the CEU population sample.

features or *C4* structures based on the test individual's genotypes for different sets of SNPs: HapMap Phase III, Illumina Immunochip, Illumina OmniExpress, and Affymetrix 6.0. We calculated the correlation (r^2) between the probabilistic dosage from imputation (using Beagle[115]) and the experimentally-determined genotypes.

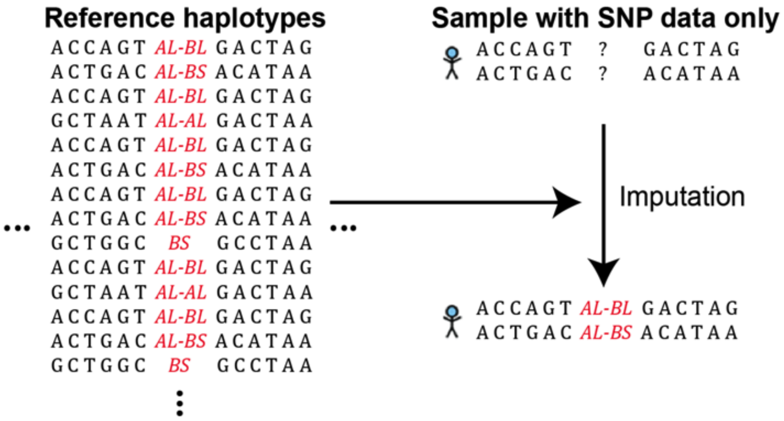


Figure 2.11. Simplified schematic of imputation of *C4* structures from SNP genotype data. A panel of reference haplotypes constructed by jointly phasing *C4* structures with *HLA* SNP haplotypes was used to impute *C4* structural variation from surrounding SNP genotype data in individuals (e.g., such as those in a GWAS) for whom SNP genotypes (but not *C4* structure) is known.

Through these leave-one-out trials, we found that *C4* structures and copy number can be imputed from SNP haplotypes imperfectly, but with greater accuracy than was possible using the best individual tag SNP that we identified (**Figure 2.12**). The prediction from imputation correlated with the copy number measurements from ddPCR more strongly for copy number of *C4L, C4S*, and *C4A* (r^2 of 0.80-0.85, 0.68-0.80, and 0.6-0.75, respectively) than for copy number of *C4B* (r^2 of 0.36-0.45). Imputation was also able to predict genotypes of the four common *C4* structures – with greater accuracy for *BS* and *AL-BS* (r^2 of 0.75-0.92 and 0.81-0.84, respectively) than for *AL-BL* and *AL-AL*

Figure 2.12. Imputability of *C4* structural variation, as estimated from a series of leave-one-out trials within our reference panel. (**a**) r^2 from Pearson correlation between the copy number of each structural feature, as determined by ddPCR, and the genotype of a single SNP that we identified as correlating most strongly with copy number (gray) or imputation-based probabilistic dosages (colored bars) using different sets of SNPs as indicated in the legend. (**b**) r^2 between the dosage per diploid genome (0, 1, or 2) of each of the four common *C4* structures that we identified using our approach described earlier, and the genotype of the best tag SNP we identified or the dosage from imputation. Error bars represent 95% confidence intervals around the Pearson r^2 value.

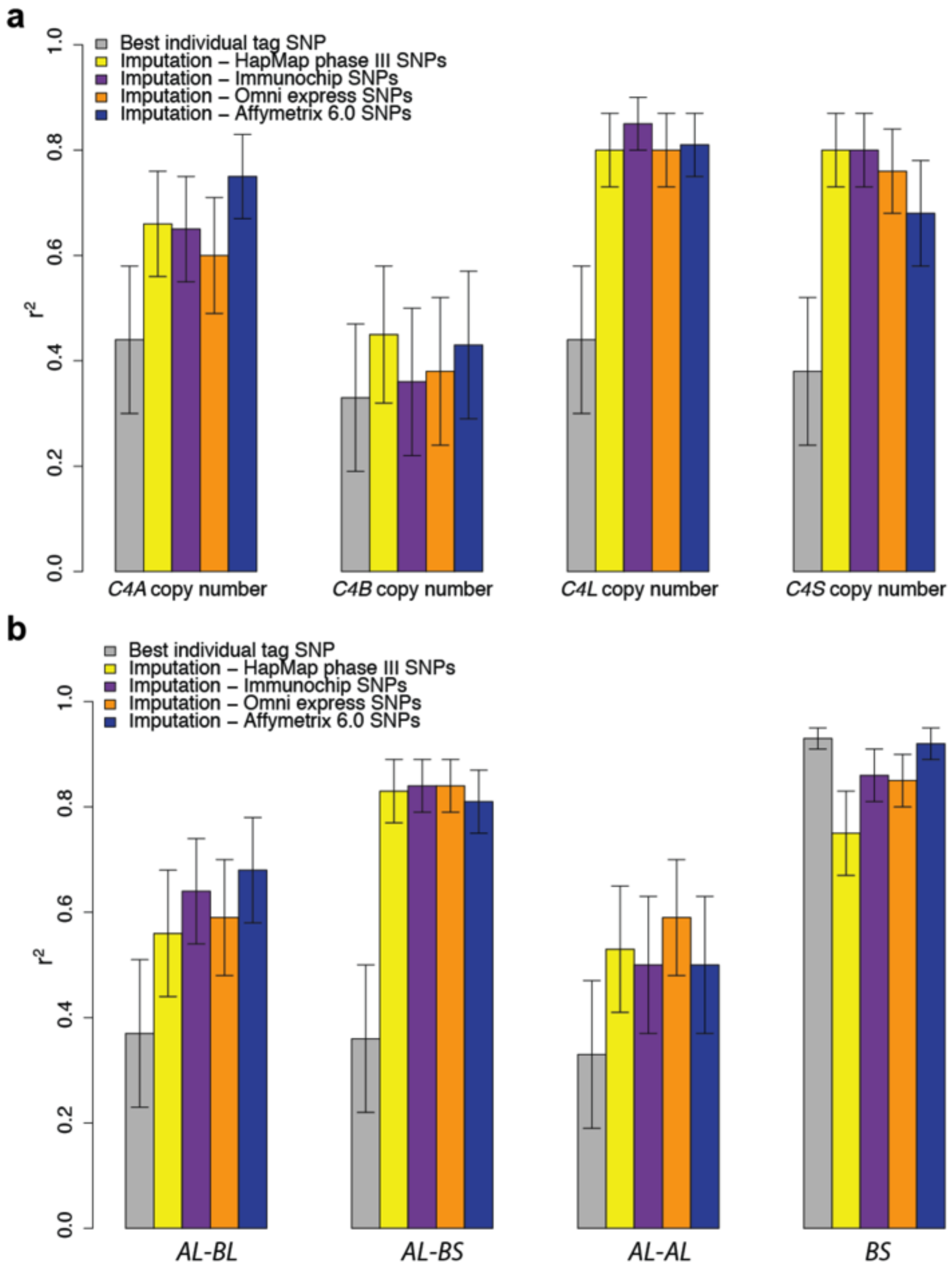


Figure 2.12 (continued)

(r^2 of 0.56-0.68 and 0.50-0.59, respectively). The accuracy of imputation using sets of *HLA* SNPs on different genotyping array platforms was comparable.

Discussion

We have described a novel combination of molecular and statistical approaches for characterizing complex genome structural variation and have applied it to the *C4* locus. The extension of this approach to other loci in the human genome that are structurally complex will provide a better understanding of how such loci vary across individuals and populations and reveal their relationship to single nucleotide variation.

The high level of concordance between our ddPCR approach and orthogonal molecular and computational approaches provides confidence in our copy number measurements of the *C4* locus. Although these measurements of diploid copy number revealed a high degree of polymorphism of the individual structural features of *C4,* our analysis of *C4* structure on a per-chromosome level indicated that four common haplotypes of *C4* structure were sufficient to account for much of this apparent structural complexity in the locus.

Prior studies of the *C4* locus[109, 110, 116] have also described the extent of copy number variation in *C4* and revealed its various structural forms. Based on evaluating the relationship of structural variation in *C4* to a limited number of SNPs in the *HLA* locus, these studies concluded that SNP genotypes cannot serve as proxies for *C4* copy number[110] nor can *C4* structural variation be used for prediction of *HLA* haplotypes[116]. Our work is the first to critically evaluate the extent to which *C4* structural variation can be captured through imputation. The results from our leave-one-out analyses suggested that *C4* structural variation can be predicted from SNP genotypes using imputation, with multiple features of *C4* variation predicted with strong correlation ($r^2 > 0.7$) between imputed and experimentally-derived genotypes. Although imputation did not capture the full extent of structural variation in *C4*, our approach is enabling in powerful ways. First, it allows inexpensive and immediate investigation of the contribution of *C4* structural variation to phenotypes, making use of already existing, vast SNP genotype data from GWASs. The power lost as a result of imperfectly predicting genotypes through imputation is offset by the ability gained to evaluate association to phenotypes in exceedingly large GWAS data sets. This ability to conduct well-powered studies is critical for complex phenotypes like SLE and schizophrenia in which the effect that individual loci have on the phenotype is relatively small. Furthermore, because ddPCR is highly scalable (i.e., compared to other approaches for analyzing the *C4* locus, such as Southern blot analysis or whole-genome sequencing), we can critically evaluate association results from imputation by direct molecular analysis in large, clinical cohorts. Finally, the ability to relate phenotypes to a series of specific structures of *C4* (even if they are not determined with full confidence) enables testing association to phenotypes in novel and powerful ways, which we will demonstrate in Chapters 3 and 4.

Contributions

The strategy presented in this chapter for characterizing complex structural variation was developed with Steve McCarroll's guidance and input. Robert Handsaker performed read-depth analysis of whole-genome sequence data to calculate *C4* copy number using Genome STRiP.

Methods

Determining copy number of C4 *CNVs using qPCR*

The qPCR assays and reaction conditions for typing the *C4* CNVs were based on Wu et al[109]. The oligonucleotide sequences for the primers and probes are listed in Supplemental Table 1. The reverse primer of the assays used for typing *C4A* and *C4B* copy number were designed to the paraolog-defining sequence on exon 26. The reactions were performed in 10 µl, with 15 ng of input genomic DNA, 2x TaqMan universal PCR master mix (Applied Biosystems), forward and reverse primers each at a final concentration 500 nM, VIC-labeled TaqMan MGB probe for *C4* and a FAM-labeled TaqMan MGB probe for *RP1* (reference locus), each at a final concentration of 100 nM. PCR was performed on a CFX384 Real-Time PCR Detection System (Bio-Rad Laboratories) with the following cycling conditions: 95°C for 10 minutes, followed by 40 cycles of 95°C for 15 seconds and 60°C (for *C4A*) or 59°C (for *C4B*) for 1 minute.

Determining copy number of C4 *CNVs using ddPCR*

50 ng of genomic DNA was digested in 10 µl, using 1 unit of AluI (New England Biolabs) at 37°C for 1 hour. The digested DNA was diluted two-fold with water. The oligonucleotide sequences for the primers and probes for assaying copy number of *C4A*, *C4B*, *C4L*, and *C4S* were from Wu et al.[109] and are listed in Supplemental Table 1. Reaction mixtures consisted of 6.25 µl of the digested, diluted DNA, 1 µl of 20x primer-probe mix (containing 18 µM of forward and reverse primers each and 5 µM of fluorescent probe) for *C4* and a reference locus (*RPP30*) each, and 2x ddPCR Supermix for Probes (Bio-Rad Laboratories). For each sample, this reaction mixture was each emulsified into approximately 20,000 droplets in an oil/aqueous emulsion, using a microfluidic droplet generator (Bio-Rad Laboratories). PCR was performed within the droplets using the following cycling conditions: 95°C for 10 minutes, 40 cycles of 94°C for 30 seconds and 60°C (for *C4A* and *C4L*) or 59°C (for *C4B* and *C4S*) for 1 minute, followed by 98°C for 10 minutes. After PCR, the fluorescence from each of the two fluorophores in each droplet was read by a QX100 droplet reader (Bio-Rad Laboratories). Data were analyzed using the QuantaSoft software (Bio-Rad Laboratories), which estimates absolute concentration of DNA templates by Poisson-

correcting the fraction of droplets that are positive for each amplicon (*C4* or *RPP30*). The ratio of the concentration of the *C4* amplicon to that of the reference (*RPP30*) amplicon multiplied by two (the number of copies of the reference locus in a diploid genome) yields the absolute number of copies of the *C4* CNV in a diploid genome.

Detecting the presence of the C4AS *structural form using ddPCR*

A long-range PCR was first performed to enrich for DNA templates that spanned the *L/S* and *A/B*-defining molecular features (in intron 9 and exon 26, respectively). The oligonucleotide sequences for this long-range PCR assay are provided in Supplemental Table 1. The reaction was performed in 50 µl and consisted of 50 ng of input genomic DNA, 5X Long Range Buffer (Mg2+ free) (Kapa Biosystems), 25 mM $MgCl_2$, 10 mM of each dNTPs and 10 uM each of forward and reverse primers and 12.5 units of Kapa LongRange DNA Polymerase. Cycling conditions were as follows: 94°C for 2 minutes; 35 cycles of 94°C for 25 seconds, 61.4°C for 15 seconds, and 68°C for 11 minutes; and 72°C for 11 minutes.

The PCR product from the long-range PCR was used as input into the ddPCR assay. In order to add an appropriate amount of DNA into the ddPCR assay, the long-range PCR products were quantified using a Qubit Fluorometer (Life Technologies). Based on the quantification, PCR products were diluted to contain approximately 4,000 copies per µl and 8 µl of this diluted DNA was added to a ddPCR mixture containing 1 µl of a 20x primer-probe mixture of the *C4A* assay (FAM), 1 µl of a 20x primer-probe mixture of *the C4S* assay (VIC), and 10 µl of 2x ddPCR Supermix for Probes (Bio-Rad Laboratories). The generation of droplets and the PCR cycling conditions were as described above for the ddPCR assays of *C4* copy number, with an annealing temperature of 59°C. After droplets were read, a Fisher's exact test was performed to test independence of the number of droplets falling into each of the four possible classes: (i) FAM-VIC-, (ii) FAM-VIC+, (iii) FAM+VIC-, and (iv) FAM+VIC+.

Detecting the presence of the C4AS *structural form using long-range PCR*

In order to evaluate the accuracy of the results from the ddPCR-based approach for detecting the presence of the *AS* form of the *C4* gene, we designed a second long-range PCR assay. In this assay, we used a forward primer upstream of the HERV sequence and the reverse primer against the sequence specific to *C4A* in exon 26 (sequences in Supplemental Table 1), in order to amplify both *AS* and *AL* genes, and distinguish them based on the size of the product (~5.1 kb from a *C4AS* gene and ~11.5 kb from a *C4AL* gene). Reaction mixtures and cycling conditions were as described for the long-range PCR above, with an annealing temperature of 59°C.

Inferring phase of haploid-specific copy number alleles in ambiguous trios

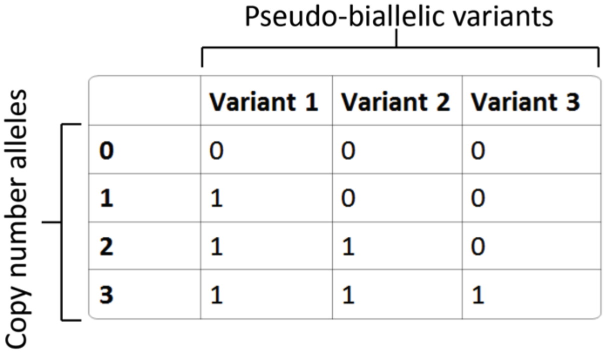
In some trios, it was not possible to unambiguously identify which alleles were transmitted and un-transmitted using our EM algorithm. For example, if all three individuals in a trio had a copy number of 1, each parent could have transmitted either the 0- or the 1-copy allele to the offspring. In order to resolve these ambiguities, we made use of the SNP backbone data available for these individuals from the HapMap project. We created a reference set of haplotypes containing both SNPs and copy number alleles from trios in which haploid-specific copy number and the transmission status of each allele were inferred confidently. We then used this core set of haplotypes as a reference to phase the copy number alleles onto SNP haplotypes using Beagle[113] for trios in which transmission status of the copy number alleles was ambiguous.

Inferring haploid-specific copy number using diploid genotypes of copy number and SNP genotypes

We used genotypes of SNPs within ~1 Mb on either side of the *C4* locus (from 31 Mb to 33 Mb on chromosome 6) in order to distinguish between the multiple allelic combinations that could theoretically give rise to each diploid copy number. For each given diploid copy number, we provided a vector of genotype likelihoods as input for phasing in Beagle[117] (version 4). For example, for an individual with a diploid copy number of 4, two allelic combinations are possible – 2+2 and 1+3 – if the alleles segregating in the population were 0, 1, 2, and 3. In the corresponding vector of likelihoods, these two genotypes were encoded as equally likely ($\log_{10}$ likelihood of -0.3), and the remaining eight genotypes that do not give rise to a diploid copy number of 4 were assigned a $\log_{10}$ likelihood of -1000 (i.e., to indicate that they are extremely unlikely). We phased these likelihoods together with SNP genotypes and obtained posterior genotype probabilities for each possible genotype, for each individual. These probability estimates readily identified the most likely genotype for each individual (the maximum posterior genotype probability was > 95% for ~99% of the samples).

Imputation of C4 *structural variants using Beagle*[115]

In order to use existing imputation algorithms (which are well-equipped to handle biallelic variants such as SNPs) for analysis of multi-allelic variants, we encoded structural variants in *C4* as a series of pseudo-biallelic variants. For *C4* CNVs, we created n-1 pseudo-biallelic variants for a variant with n alleles and encoded them such that the sum of the alleles across these variants would equal the copy number, as illustrated below (**Figure 2.13**):


Pseudo-biallelic variants
Copy number alleles

	Variant 1	Variant 2	Variant 3
0	0	0	0
1	1	0	0
2	1	1	0
3	1	1	1

Figure 2.13. Encoding *C4* copy number variants (CNVs) as a series of pseudo-biallelic variants. For a *C4* CNV with n alleles, n-1 biallelic variants were created such that the sum of the alleles across these variants would equal the copy number.

For *C4* structures, we created a pseudo-biallelic variant for each of the 15 structures. If a haplotype contained a given *C4* structure, the corresponding variant would be encoded with a '1' allele and, if absent, a '0' allele.

This encoding scheme allowed Beagle to yield probabilistic dosages from imputation for each variant. For imputation of *C4* CNVs, we summed these dosages across the constituent pseudo-biallelic variants to obtain the imputed estimate of copy number. For imputation of *C4* structures, we used the imputed dosage from the corresponding pseudo-biallelic variant directly as an estimate of the dosage of that given structure in a diploid genome. We then evaluated correlation between these imputed dosages and experimentally derived genotypes to calculate an r^2 value as a metric of imputation accuracy.

Chapter 3

Structural variation in *C4* associates to risk of systemic lupus erythematosus independently of *HLA-DRB1* risk alleles

Introduction

The association of genetic variation in the *HLA* locus to systemic lupus erythematosus (SLE) is strong, complex, and unsolved, as described in Chapter 1. The *C4* gene, within the *HLA* locus, has been investigated in SLE since 1974[118], but the question of whether structural variation in *C4* contributes to the *HLA* association signal in SLE is unsettled due to various limitations of these studies. Early studies of *C4* in SLE[119, 120] profiled C4 protein levels in the serum and used this as a marker of genetic variation in the *C4* locus. However, such studies were confounded by the fact C4 and other complement components are consumed during the course of active disease in SLE. Other studies have investigated copy number variation at the *C4* locus and have reported association to low copy number[58, 59] but without accounting for variation in the *HLA* locus outside of *C4*. As a result, it is unclear whether the apparent association of *C4* to SLE is due to linkage disequilibrium (LD) with variation in *HLA* classical genes. A more recent study investigated copy number variation in *C4* in the context of variation elsewhere in the *HLA* locus and concluded that *C4* CNVs are not an independent risk factor for SLE[60]. However, this report did not examine the full range of copy number variation in *C4*, nor did it investigate association to the length polymorphism in *C4.*

Having characterized structural variation in *C4* and its relationship to surrounding variation in the *HLA* locus, we sought to apply the methods developed in Chapter 2 toward evaluating the role of the *C4* locus in SLE. We hypothesized that structural variation in *C4* contributes to SLE (and to the *HLA* association signal) based on the following lines of reasoning. First, complete deficiency of *C4,* though rare, is one of the strongest risk factors for SLE[57, 121], and we reasoned that common forms of copy number variation in the *C4* locus could also contribute to SLE. Second, other genes that function in the complement pathway have also been implicated in SLE. Complete deficiencies of *C1q*, which functions upstream of *C4* in the complement pathway also contribute to

SLE[122, 123]. In addition, common variants in the *ITGAM* locus, which encodes a receptor for iC3b (a proteolytic cleavage fragment of the central protein of the complement pathway), have been found to associate to SLE from genome-wide association studies[44, 46]. Third, studies of *C4* knockout mice support a role for *C4* in SLE[124] – whereas association studies of the *C4* locus in humans are complicated by LD to *HLA* classical alleles, the presence of a lupus-like phenotype in *C4*-deficient mice raises the suspicion that *C4* may independently contribute to SLE pathogenesis. Fourth, a role for *C4* in SLE is biologically plausible and is consistent with the current understanding of SLE pathogenesis. Complement proteins may play a role in clearance of apoptotic cells[125], and patients with SLE have defective clearance of apoptotic debris[126], which can act as a source of auto-antigens[127] and contribute to inflammation[128].

***C4A* gene expression in lymphoblastoid cell lines is directly proportional to *C4A* gene copy number**

We evaluated the relationship between *C4* structural variation and *C4A* expression in 46 lymphoblastoid cell lines from HapMap CEU and YRI samples. We measured gene expression using reverse-transcriptase droplet digital PCR (RT-ddPCR). Analogous to the ddPCR approach described in Chapter 2, RT-ddPCR involves a reverse-transcription step that occurs within the droplets prior to PCR amplification. Primers and fluorescent probes against the target locus of interest (*C4*) as well as a control locus are included in the RT-ddPCR assay, and absolute quantification of the number of RNA templates from each locus is obtained. By normalizing the concentration of *C4* RNA to the RNA from a control locus, variation in the amount of input RNA is accounted for, allowing for comparison of gene expression levels across samples.

Across 46 unrelated samples (23 from CEU and 23 from YRI), *C4A* copy number associated with *C4A* expression ($p = 1.5\text{x}10^{-9}$), with expression levels directly proportional to gene dosage (**Figure 3.1a**). The association was driven primarily by the CEU samples (p = 3.0x10^{-7}), which showed greater variation in *C4A* copy number than YRI samples (0-4 compared to 2-3). *C4A* gene copy number explained 62% of the variance in *C4A* expression and the intercept from a linear model fitting *C4A* expression to *C4A* copy number was not statistically different from 0 (p = 0.39).

We then related *C4A* expression in LCLs to the four common *C4* structures in CEU samples. We calculated an effect size, β, associated with each of the four structures from fitting *C4A* expression to the number of chromosomes per diploid genome that carried that structure (0, 1, or 2). These four common *C4* structures associated to *C4A* expression levels in proportion to the number of *C4A* genes they contained (**Figure 3.1b**).

HLA-DRB1 **alleles and** ***C4A*** **copy number are strongly associated to SLE**

We tested for association between *C4* structural variation and SLE risk in 2,390 SLE cases and 8,707 controls of European ancestry drawn from several cohorts sampled within the US[129-131]. These samples were genotyped on the Immunochip (Illumina), a SNP microarray that contains approximately 200,000 SNPs enriched for variants associated to immune-mediated diseases, with dense coverage of the *HLA* locus. We used genotypes from the Immunochip for these samples as the SNP backbone for imputation of *C4* structural variation, using the reference panel and imputation approach described in Chapter 2. In addition, we also imputed *HLA* classical alleles, both using a reference panel enables joint imputation of *C4* structural variants and classical alleles as well as a much larger reference panel for imputation of classical alleles and *HLA* SNPs[114].

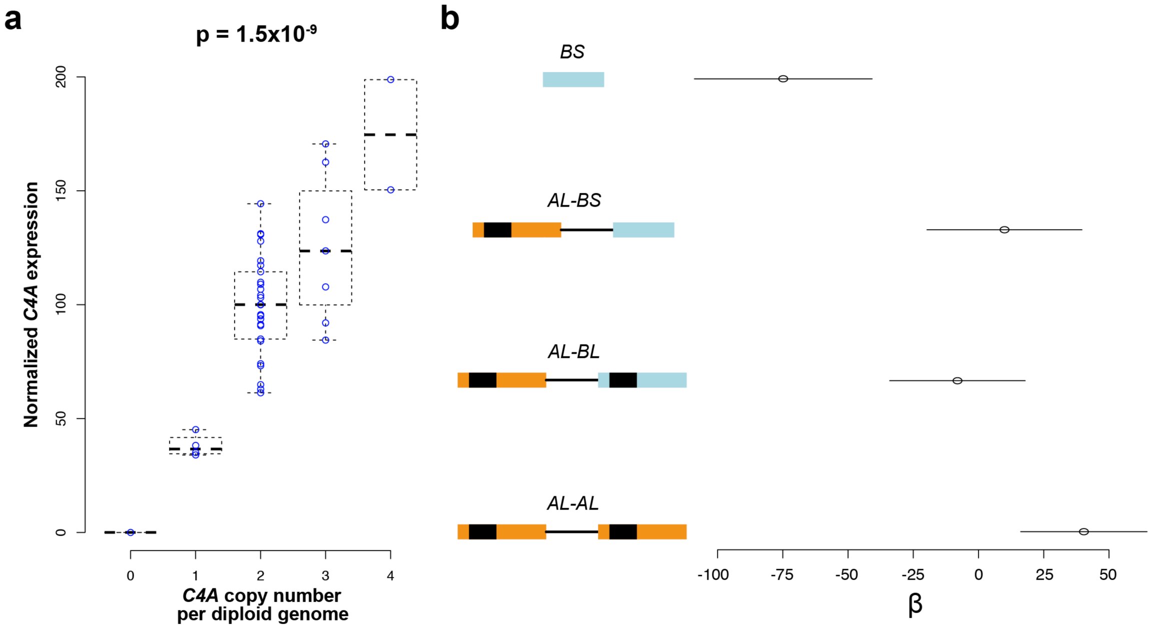


Figure 3.1 Association of *C4* structural variation to *C4A* expression in lymphoblastoid cell lines. (**a**) *C4A* expression was directly proportional to *C4A* gene copy number. (**b**) The four common structures of *C4* associated to *C4A* expression in proportion to the number of *C4A* copies (orange rectangles) they contain. β measures the effect size calculated from a linear model fitting *C4A* expression to the number of chromosomes (0, 1, or 2) carrying that *C4* structure per diploid genome.

Association analysis of *HLA* SNPs*, HLA* classical alleles, and copy number of *C4* structural features, revealed a strong association of the *HLA* locus to SLE risk (**Figure 3.2**). The variant most strongly associated to SLE was rs1059615 ($p = 2.3x10^{-41}$), a synonymous variant in *HLA-DRB1*. rs1059615 was strongly correlated with *HLA-DRB1*03:01* ($r^2 = 0.98$), which was also among the most strongly associated variants ($p = 6.2x10^{-41}$). Given that *HLA-DRB1*03:01* is a previously well-replicated variant in SLE and it was in near-perfect linkage disequilibrium with rs1059615, we investigated *HLA-DRB1*03:01* in subsequent analyses as the primary signal from the Class II region of the *HLA* locus. *HLA-DRB1**15:01, another *HLA-DRB1* allele that has been consistently implicated in SLE, showed weaker association ($p = 6.2x10^{-5}$). However, in a joint regression model with *HLA-DRB1*03:01 and HLA-DRB1*15:01*, the association to both variants strengthened ($p = 3.2x10^{-46}$ for *HLA-DRB1*03:01*; $p = 1.54x10^{-10}$ for *HLA-DRB1*15:01*). Both alleles associated to SLE risk with a dominant effect **(Figure 3.3**), and we accordingly modeled a dominant effect of these *HLA-DRB1* alleles in subsequent association analyses.

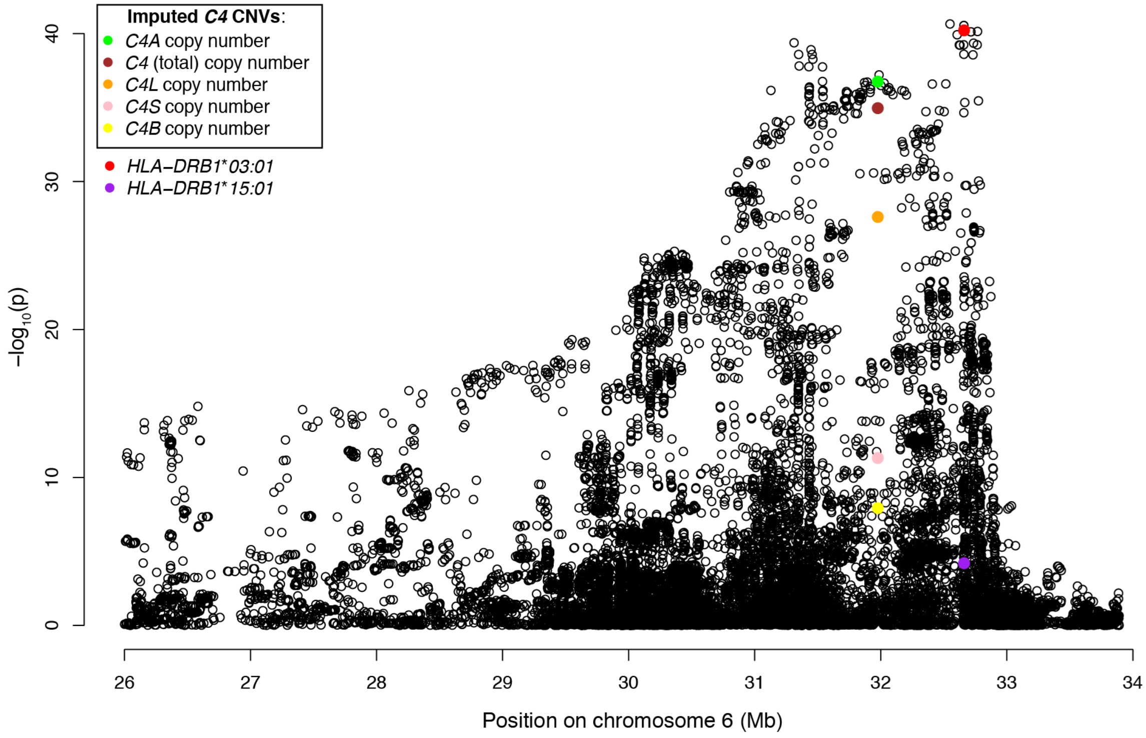


Figure 3.2. Association of the *HLA* locus to SLE. Directly genotyped variants as well as imputed variants (*C4* structural variants and *HLA* classical alleles) were evaluated for association to SLE.

Copy number of *C4A* associated more strongly to SLE than other structural features of *C4* ($p = 1.8x10^{-37}$; **Figure 3.2**). Despite an estimated r^2 of 0.65 (0.55 - 0.75) between imputed and experimentally-derived genotypes of *C4A* copy number (based on leave-one-out trials described in Chapter 2, Figure 2.12), the strength of association was comparable to the most strongly associated variants in the *HLA* locus. In joint models of association to SLE, copy number of no other *C4* structural feature explained the association to *C4A* copy number (**Table 3.1**).

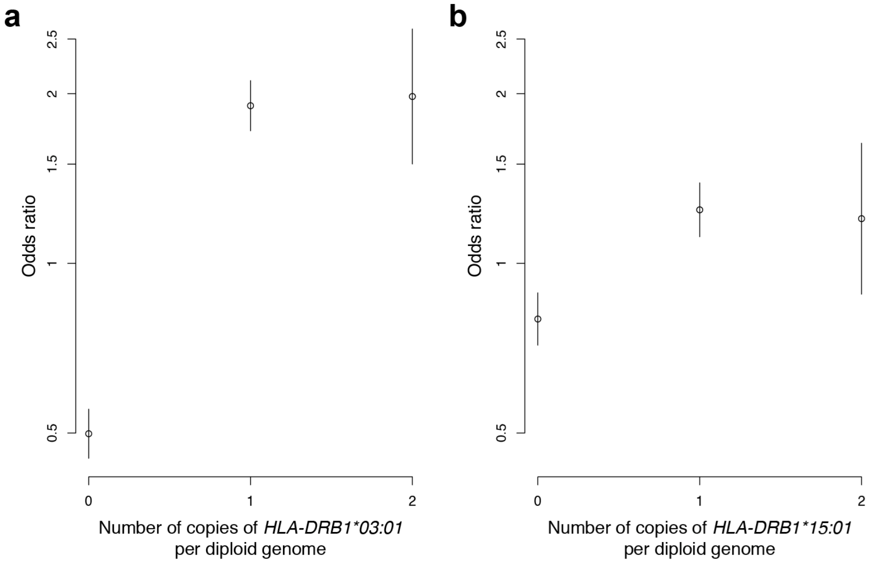


Figure 3.3. *HLA-DRB1* alleles, (**a**) **03:01* and (**b**) **15:01*, associated to SLE risk with a dominant effect.

Increased copy number of *C4A* is associated with lower risk of SLE independently of *HLA-DRB1* risk alleles

The above analyses identified a strong association of SLE to *C4A* copy number as well as *to HLA-DRB1* risk alleles, and we next sought to evaluate the independence of these effects. *C4A* copy number was partially correlated with *HLA-DRB1*03:01* (r^2 = 0.45) but not with **15:01* (r^2 = 0.005). In a joint model with these variants, *C4A* copy number remained associated to SLE (p = $9.1x10^{-7}$, **Table 3.2**).

Table 3.1. Association of *C4* CNVs to SLE. P-values for imputed copy number of *C4* structural features are shown before and after conditioning on the indicated variant.

	Conditional on:					
	-	*C4A* copy number	*C4B* copy number	*C4L* copy number	*C4S* copy number	*C4* (total) copy number
C4A copy number	$1.8\text{x}10^{-37}$	-	$1.8\text{x}10^{-32}$	$1.6\text{x}10^{-11}$	$6.4\text{x}10^{-27}$	$6.4\text{x}10^{-6}$
C4B copy number	$1.2\text{x}10^{-8}$	0.02	-	$4.3\text{x}10^{-3}$	$2.2\text{x}10^{-4}$	$6.4\text{x}10^{-6}$
C4L copy number	$2.5\text{x}10^{-28}$	0.16	$2.5\text{x}10^{-22}$	-	$5.3\text{x}10^{-25}$	0.04
C4S copy number	$4.9\text{x}10^{-12}$	0.67	$2.1\text{x}10^{-7}$	$1.2\text{x}10^{-8}$	-	0.04
C4 (total) copy number	$1.1\text{x}10^{-35}$	0.02	$1.8\text{x}10^{-32}$	$1.2\text{x}10^{-8}$	$5.3\text{x}10^{-25}$	-

Table 3.2. Association of SLE to imputed *C4A* copy number, *HLA-DRB1*03:01* and **15:01* from a joint model. The p-values for the *DRB1* risk variants are from modeling a dominant effect on SLE risk.

	Conditional on:			
	-	*HLA-DRB1*03:01* and **15:01*	*HLA-DRB1*15:01* and *C4A* copy number	*HLA-DRB1*03:01* and *C4A* copy number
C4A copy number	$1.8\text{x}10^{-37}$	$9.1\text{x}10^{-7}$	-	-
*HLA-DRB1*03:01*	$6.4\text{x}10^{-42}$	-	$6.5\text{x}10^{-12}$	-
*HLA-DRB1*15:01*	$3.2\text{x}10^{-5}$	-	-	$6.5\text{x}10^{-10}$

We then evaluated the risk associated with imputed copy number of *C4A* per diploid genome and the extent to which this was influenced by *DRB1* risk variants. Prior to conditioning on **03:01* and **15:01*, lower copy number of *C4A* (0 or 1) was associated with higher risk of SLE, and higher copy number (2 or 3) was associated with lower risk, but the effect of copy number on risk was not linear (**Figure 3.4a**). After including _*03:01_ in the model together with _C4A_ copy number, the risk associated with a diploid copy number of 0 or 1 was moderated (**Figure 3.4b**). Including _*15:01_ in the model with _C4A_ copy number had little-to-no effect on the risk associated with _C4A_ copy number (**Figure 3.4c**), consistent with the lack of correlation in the genotypes of these variants. The pattern of association to imputed copy number of _C4A_ after including both _*03:01_ and_*15:01_ in the model was consistent with a linear, dose-dependent relationship, in which increasing copy number of _C4A_ associated with lower risk of SLE (**Figure 3.4d**).

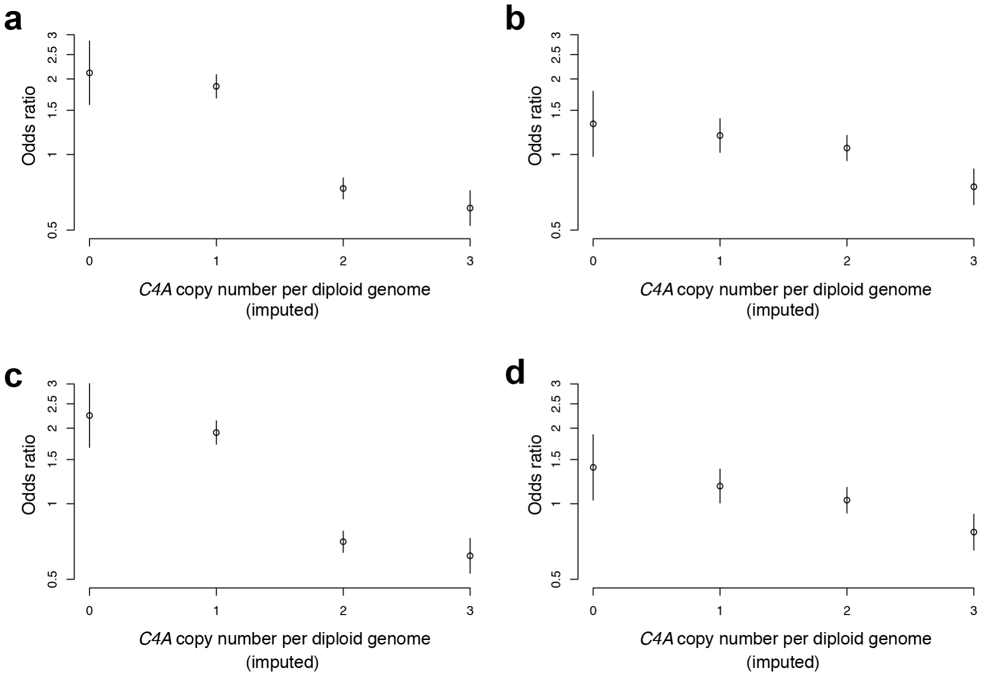


Figure 3.4. Association of SLE to (imputed) _C4A_ copy number. (**a**) Odds ratio associated with the four most common diploid copy numbers is shown prior to conditioning on _HLA-DRB1_ risk variants, (**b**) conditional on _HLA-DRB1*03:01_, (**c**) conditional on _HLA-DRB1*15:01_, and (**d**) conditional on both _HLA-DRB1*03:01_ and _HLA-DRB1*15:01_.

In order to better understand the relationship of *C4* to SLE risk, we next evaluated the association of SLE to the four most common *C4* structures (*AL-BL, AL-BS, AL-AL* and *BS*) – the substrates underlying the variation in copy number of *C4*. These structures associated to SLE with three independent levels of risk, spanning a wide range of odds ratios – *AL-AL* (two copies of *C4A*), with a lower level of risk (OR = 0.65, 95% CI: 0.57-0.75); *AL-BL* and *AL-BS*, (one copy of *C4A* each), with intermediate, similar levels of risk (OR = 0.96, 95% CI: 0.91-1.02; OR = 0.90, 95% CI: 0.83-0.97, respectively); and *BS*, (no copies of *C4A*), with a higher level of risk (OR = 1.79, 95% CI: 1.64-1.96) (**Figure 3.5a**). The effect size associated with *BS* appeared to be higher than expected based on the association of SLE to the other three *C4* structures. Consistent with the finding in Chapter 2 that 90% of *BS* haplotypes also contained an *HLA-DRB1*03:01* allele, the effect size associated with *BS* was moderated when including **03:01* in the model with *C4* (**Figure 3.5b**). Including **15:01* in the model had a minimal effect on the association of *C4* structures to SLE risk **(Figure 3.5c),** and including both **03:01* and **15:01* in the model with *C4* rationalized the pattern of association of *C4* structures to SLE risk (**Figure 3.5d**) – the effect size associated with each *C4* structure was inversely proportional to the number of copies of *C4A* the structure contains. In addition, the association of *C4* structures to SLE risk mirrored the pattern of association of the same structures to expression of *C4A* in lymphoblastoid cell lines (Figure 3.1b).

The association of *C4* structures to SLE with three distinct levels of risk suggested that this result was unlikely to be due to correlation of *C4* structural variation with a single, bi-allelic variant or SNP. However, in theory, functional variant(s) outside of the *C4* locus could load onto haplotypes of different *C4* structures with varying frequencies such that *C4* structures segregating on haplotypes with a larger fraction of such risk variants could appear to associate with higher levels of risk. In addition, variants other than the *HLA-DRB1* risk alleles that we did not explicitly model in a conditional regression analysis could be influencing association to *C4*. We therefore developed a form of association testing that exploited the population genetic properties of the *C4* locus to evaluate the extent to which the *HLA* haplotypes each *C4* structure segregates on influenced its association to SLE.

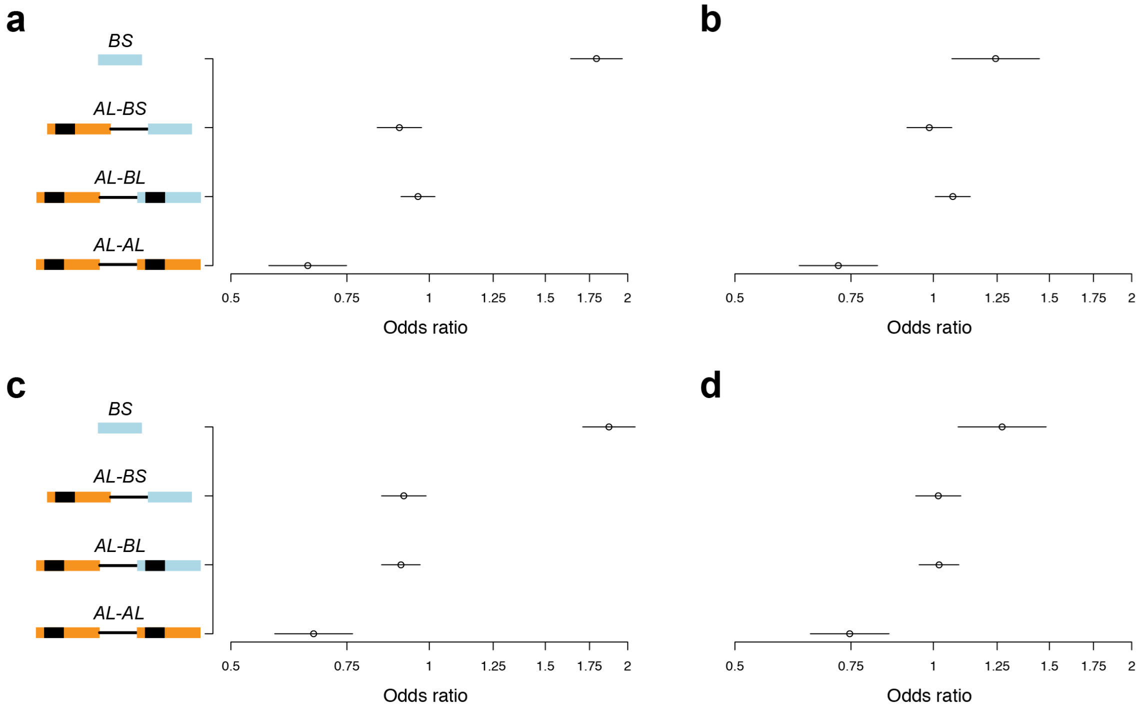


Figure 3.5. Association of SLE to the four most common *C4* structures. (**a**) Odds ratio associated with each *C4* structure is shown prior to conditioning on *HLA-DRB1* risk variants, (**b**) conditional on *HLA-DRB1*03:01*, (**c**) conditional on *HLA-DRB1*15:01*, and (**d**) conditional on both *HLA-DRB1*03:01* and *HLA-DRB1*15:01*. Note that the association of *C4* structures to SLE risk mirrors their association to *C4A* expression in lymphoblastoid cell lines (Figure 3.1b). Note also that the outsized effect associated with the *BS* structure in (**a**) was moderated upon including **03:01* in the model (**b** and **d**).

Recurrence of *C4* structures on different *HLA* haplotypes enables a novel form of association analysis

In Chapter 2, we found that the three most common *C4* structures each segregated on multiple *HLA* haplotype backgrounds (Figure 2.10). In order to distinguish between the effect of *C4* structural variation and effects elsewhere in the *HLA* locus, we exploited the fact that the same *C4* structures existed on different *HLA* haplotype backgrounds. We treated *C4* structures segregating on different haplotypes as separate alleles. The presence of this long allelic series (of 11 alleles, each at a frequency of more than 3%) enabled a novel form of association testing, in which the occurrence of a *C4* structure on each new *HLA* haplotype yielded an additional falsifiable test of its association to SLE. If *C4* structural variation had a causal influence on SLE risk, then the same *C4* structure should show similar levels of association to SLE, regardless of which haplotype it segregated on.

In order to test this prediction, we encoded *C4* structures as separate alleles if they segregated on different *HLA* haplotype backgrounds in our reference panel for imputation. We then imputed these *C4* structural alleles into the SLE case-control samples, using the surrounding SNP genotype data as the backbone for imputation, and evaluated association to each of these alleles. The ranking of effect sizes associated with each *C4* allele was consistent with the number of copies of *C4A* each allele contained (**Figure 3.6a**) – the *BS* allele, with no copies of *C4A*, was associated with a higher risk than any of the 8 alleles with 1 copy of *C4A,* which in turn were associated with a higher risk than the 2 alleles with 2 copies of *C4A* (**Figure 3.6a**). The correlation of effect sizes across these 11 alleles with their *C4A* copy number was statistically significant (p = 7.8x10^{-4}, Pearson correlation). This overall pattern of association to the 11 *C4* alleles was preserved in joint models that included *HLA-DRB1*03:01* **(Figure 3.6b**), *HLA-DRB1*15:01* (**Figure 3.6c**)***,*** or both (**Figure 3.6d**), together with *C4* alleles, with a moderation of effect size associated with *BS* in models that included **03:01*.

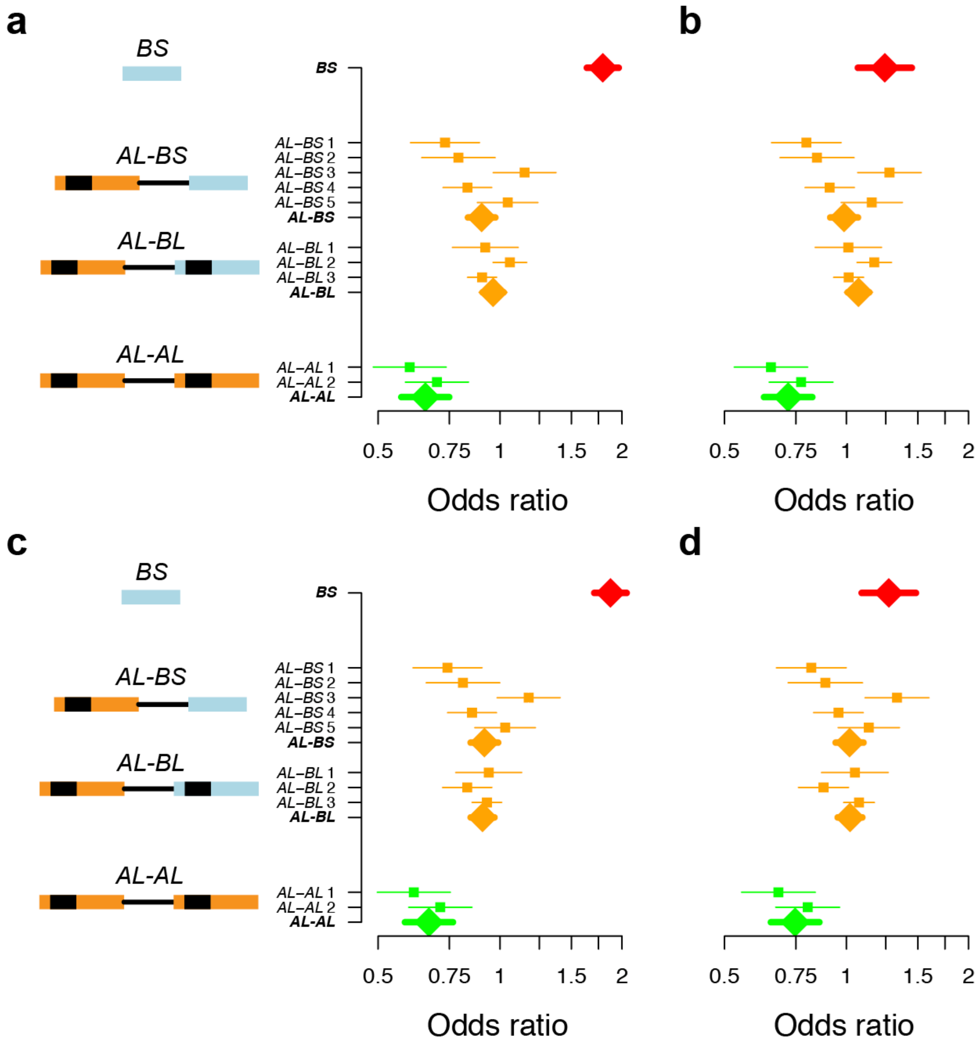


Figure 3.6. Association of SLE to *C4* structures segregating on different *HLA* haplotypes backgrounds. (**a**) Odds ratio associated with each *C4* allele is shown prior to conditioning on *HLA-DRB1* risk variants, (**b**) conditional on *HLA-DRB1*03:01*, (**c**) conditional on *HLA-DRB1*15:01*, and (**d**) conditional on both *HLA-DRB1*03:01* and *HLA-DRB1*15:01*. *C4* structures suffixed with a different number (e.g., *AL-BL* 1 and *AL-BL* 2) represent alleles with the same structural content but segregating on different *HLA* haplotype backgrounds, and they correspond to the annotation in Figure 2.10 in Chapter 2. The diamond symbols in bolder colors represent the effect size from a meta-analysis across alleles that have the same *C4* structural content.

SLE associates more strongly to *C4A* copy number measured using ddPCR than to imputed copy number

The analyses described above represent the first attempts to use imputation to assess contribution of complex structural variation to a phenotype. Therefore, we sought to evaluate the accuracy of this approach using direct molecular measurements from droplet-digital PCR (ddPCR, described in Chapter 2). We used ddPCR to measure *C4A* copy number in 698 cases and 1,226 controls available from the larger set of samples analyzed using imputation. As expected, *C4A* copy number, determined by ddPCR, associated more strongly to SLE ($p = 3.9 \times 10^{-6}$) than imputed *C4A* copy number ($p = 6.1 \times 10^{-4}$) for the same samples. In principle, this improvement in the strength of association could be a result of the ddPCR-based genotypes correlating more strongly, than imputation-based *C4A* copy number, with *HLA-DRB1* risk variants. However, the correlation was approximately the same based on ddPCR (r^2 = 0.30 to **03:01*; 0.0 to **15:01*) and imputation (r^2 = 0.35 *to *03:01*; 0.0 *to *15:01*) across these 1,924 samples.

Discussion

We have investigated the contribution of structural variation in *C4* to SLE and have found that its association to SLE was not fully explained by previously identified risk alleles, *HLA-DRB1*03:01* and **15:01*, and that it is unlikely to be a result of correlation with variation elsewhere in the *HLA* locus. To our knowledge, this is the first study to investigate the contribution of complex structural variation to a phenotype using imputation, and more specifically, of the full extent of structural variation in *C4* (including copy number of each structural feature – *C4A, C4B, C4L*, and *C4S*) to SLE.

Our finding that SLE associates to *C4* independently of *HLA-DRB1* risk alleles disagrees with that of Boteva et al.[60], who concluded that *C4* copy number is not an independent risk factor for SLE. This discrepancy can be reconciled based on the following differences between our studies. First, the sample size in Boteva et al. was 501 SLE cases and 719 controls (and an additional 464 cases and 449 controls from a second, independent cohort), whereas we analyzed a much larger set of samples (2,390 cases and 8,707 controls). Our effective sample size is reduced, however, because of imperfectly predicting genotypes using imputation. The sample size required to achieve the same level of statistical power is increased by a factor of $1/r^2$, where r^2 is the squared-correlation between the genotypes tested for association and the genotypes of the causal variant[132]. Based on this relationship and our estimates of imputation accuracy from Chapter 2 (Figure 2.12), our effective sample size is reduced by a factor of 1.3 - 1.8 (but still larger than that of Boteva et al). Second, in testing association to *C4* copy number, Boteva et al. grouped *C4A* copy numbers into high- and low-copy number classes in reference to a copy number of 2. Under a model of a linear, dose-dependent effect of *C4A* copy number on SLE risk, this approach reduces statistical power. Third, in testing conditional association, they restricted their analyses to only samples with a copy number of 0, 1, or 2 – the resulting reduction in sample size (and variance in the genotype) decreased their statistical power. Furthermore, restricting analysis to three copy number genotypes (0, 1, and 2) that can, in principle, be readily tagged by SNPs and classical alleles that are bi-allelic, could also result in a stronger correlation between the genotypes of these variants and a reduction in the strength of association from a conditional regression analysis.

Our ddPCR-based analysis of a subset of the SLE case-control samples that were analyzed by imputation revealed a stronger association of SLE to *C4A* copy number ascertained from a molecular assay, compared to the statistical prediction from imputation. Considering that imputed copy number of *C4A* was already among the most strongly associated variants in the *HLA* locus, it is possible for more accurate measurements of copy number (i.e., from ddPCR) on the entire cohort of samples to result in a stronger association of *C4A* copy number to SLE than any other variant in the *HLA* locus. In addition, our results offer a potential explanation for the observation that the *HLA* locus appears to have an outsized association signal compared to other loci that have been identified from genome-wide association studies of SLE (Chapter 1, Figure 1.1a) – the risk allele of *HLA-DRB1*03:01* and *C4 BS* are mostly present on the same, long haplotype, contributing to the strength of each other's association to SLE.

Our finding of a considerably stronger association of SLE to *C4A* copy number than to *C4B* copy number suggests a functional difference between these two paralogs (which are > 99% identical at the nucleotide level). This finding may reflect the differences in the biochemical properties of these two paralogs that have been described previously[108]. In particular, the higher affinity of C4A for amino groups (such as those in immune complexes) than C4B, which binds more strongly to hydroxyl groups (such as those in carbohydrate moieties on bacterial antigens), may underlie the differences in their contribution to SLE risk. Furthermore, the association of the structures that are protective in SLE with higher expression of *C4A* is consistent with the current understanding of the pathogenesis of the disease, in which complement is hypothesized to play a protective role by participating in clearance of apoptotic debris[101] and maintenance of self-tolerance[103].

Contributions

Steve McCarroll provided guidance and input on all aspects of this work. Robert Graham and Timothy Behrens (Genentech) provided the Immunochip genotype data from the *HLA* locus for the SLE case-controls samples, DNA for a subset of these samples, and feedback on the analysis. Additional DNA samples for typing *C4A* copy number using ddPCR were provided by Lindsey Criswell (UCSF) and Peter Gregersen

(Hofstra North Shore-LIJ School of Medicine). Vanessa Van Doren performed the ddPCR assays to type *C4A* copy number in these samples.

Methods

Analysis of gene expression in lymphoblastoid cell lines (LCLs)

We used RT-ddPCR assays to measure expression of *C4* in lymphoblastoid cell lines (LCLs). We designed the primers so that amplicons would span across splice junctions in order to minimize the detection of residual genomic DNA in the assay (sequences provided in Supplemental Table 1). Reactions were performed in 20 µl with 35 ng of total RNA as input, 10 µl of 2x one-step RT-ddPCR mix (Bio-Rad laboratories), 0.8 µl of 25 mM manganese acetate, and 1 µl of 20x primer-probe mix (containing 18 µM of forward and reverse primers each and 5 µM of fluorescent probe) for *C4* and a control locus (*NUDCD3, NudC domain containing 3*) each. For each sample, this reaction mixture was emulsified into approximately 20,000 droplets in an oil/aqueous emulsion, using a microfluidic droplet generator (Bio-Rad Laboratories). RT-ddPCR was performed within the droplets using the following cycling conditions: 60°C for 30 minutes, 95°C for 5 minutes, 40 cycles of 94°C for 30 seconds and 60°C for 1 minute, followed by 98°C for 10 minutes. After PCR, the fluorescence from each of the two fluorophores in each droplet was read by a QX100 droplet reader (Bio-Rad Laboratories). Data were analyzed using the QuantaSoft software (Bio-Rad Laboratories), which estimates absolute concentration of templates by Poisson-correcting the fraction of droplets that are positive for each amplicon (*C4* or *NUDCD3*). The concentration of the *C4* amplicon was normalized to that of the reference (*NUDCD3*) amplicon for each sample. We then converted the normalized expression in each sample to a percentage of the median expression value across all the samples (and this value is plotted on the y-axis in Figure 3.1a).

We fit a linear model to test association of *C4A* copy number (measured using ddPCR, as described in Chapter 2) to *C4A* expression in LCLs. We analyzed samples of two different ancestries, HapMap CEU and YRI. The heterogeneity in effect of copy number on expression across the two sets of samples was not statistically significant ($p = 0.10$), and we did a combined analysis of the samples, using ancestry as a binary covariate in our linear model. In order to relate *C4* structures to *C4A* expression, we fit a linear model for each of the four common *C4* structures, with the dosage (0, 1, or 2) of the *C4* structure per diploid genome (which we had determined for these samples in Chapter 2), as the explanatory variable. The coefficient form the linear regression model, β, is plotted in Figure 3.1b.

Quality control

QC was performed in PLINK[133] and the following set of SNPs and samples were removed: (i) variants within the duplicated *C4* locus (hg 19); (ii) samples missing genotypes for more than 5% of SNPs; (iii) SNPs with minor allele frequency less than 1%; (iv) SNPs missing genotypes across more than 2% of the samples; (v) SNPs that had differential rates of missing genotypes between cases and controls ($p < 10^{-5}$ from a Fisher's exact test); (vi) SNPs whose genotypes were not in Hardy-Weinberg equilibrium ($p < 10^{-5}$); (vii) SNPs whose allele minor allele frequency was more than 15% different from its frequency in the reference panel for imputation; and (viii) transversion SNPs (A/T and G/C) with minor allele frequency > 0.35 (these SNPs were removed as it can be problematic to determine whether they have the same strand assignment as SNPs in the reference panel for imputation). After applying these QC filters, 3,664 SNPs remained that were in common with our HapMap CEU reference panel for imputation of *C4* structural variation.

Determining copy number of C4A using *ddPCR*

We used ddPCR to measure copy number of *C4A* in the SLE case-control samples using reaction conditions detailed in Chapter 2. An assay designed to an ultraconserved element was used as a control and was run with the *C4A* assay for each sample (oligonucleotide sequences provided in Supplemental Table 1).

We flagged samples that met any of these criteria: (i) number of droplets generated < 4,000; (ii) width of 95% Poisson confidence interval > 1; (iii) concentration of the control locus < 10 copies per µl; and (iv) a difference of > 0.3 between the copy number estimated from ddPCR and the nearest integer, after applying a correction factor on a plate-by-plate basis. We calculated a correction factor for each plate of samples as follows. The absolute copy number estimate was divided by the value rounded to the nearest integer for each sample. A median of these measurements defined a correction factor from the first iteration, which was then used to multiply the original set of copy number estimates. A new correction factor was calculated from these corrected copy number estimates. We repeated this loop until the correction factor converged (typically within the first 5 iterations).

Excluding the samples that were flagged using the QC criteria described above could result in a differential missing rate between cases and controls, potentially skewing the association statistic. We inferred copy number for such samples with high-confidence by integrating the available ddPCR data for these samples together with their SNP genotypes, along with data for samples whose copy number calls from ddPCR was unambiguous. For samples that were flagged during QC, we considered a copy number of x and x+1 as being equally likely, where x is the truncated value (e.g., 2) of the copy number estimate from ddPCR (e.g., 2.45). Then, as described in Chapter 2, we provided a vector of genotype likelihoods as input for phasing in Beagle[113] (version 4). Using the surrounding SNP genotypes for these samples together with the deterministic inferences of copy number from ddPCR for the vast majority of the samples, we were able to obtain posterior genotype probabilities that readily identified the most likely genotype for each sample.

Imputation of C4 *structural variants*

We imputed *C4* structural variants using the reference panel we created in Chapter 2, consisting of 222 haplotypes from HapMap CEU samples and 7,751 SNPs, using Beagle[115]. For imputation of copy number of *C4* CNVs, we encoded the structural forms of *C4* as a series of pseudo-biallelic variants, as described in Chapter 2. In order to test association independently to each *C4* structure segregating on a different *HLA* haplotype background (defined based on 250 kb of SNPs around the *C4* locus), we explicitly encoded these as separate alleles in the reference panel. For example, the *AL-BS* structures segregating on the 5 different *HLA* haplotypes were encoded as *AL-BS* 1, *AL-BS 2*, *AL-BS* 3, *AL-BS* 4 and *AL-BS* 5 in the reference panel. We then used a narrow window of SNPs (100 SNPs on either side of the *C4* locus) to impute these structural alleles into the SLE case-control data.

Imputation of HLA *classical alleles*

We used two reference panels for the imputation of *HLA* classical alleles – a HapMap CEU panel consisting of both classical alleles as well as *C4* structural variants (n = 118 haplotypes) and a much larger panel (n = 9,956 haplotypes)[114] based on data collected by the Type 1 Diabetes Genetics Consortium (T1DGC), which did not include *C4* structural variants. We performed imputation using Beagle[115]. The CEU reference panel for the imputation of *HLA* classical alleles from Jia et al.[114] contained 90 individuals. We had *C4* structural variation data for 63 of these samples and were able to place *C4* structural variants onto haplotypes with *HLA* classical alleles and SNPs for 59 samples. The correlation (r^2) in the genotypes of *HLA-DRB1*03:01* and *15:01* imputed based on these two reference panels was 0.98 and 0.86, respectively. We used the imputed data from the CEU reference panel to infer haplotype relationships between *C4* structural variants and classical alleles, and used the imputed data based on the much larger T1DGC reference panel for association analysis.

Association analysis

We performed association analyses in a logistic regression framework. To test association to copy number of *C4* structural features (*C4A*, *C4B*, *C4L, C4S*), we used the probabilistic dosages obtained from imputation. We tested association and calculated effect size for each *C4* structural allele (e.g., *AL-BL 1*, *AL-BL* 2, etc.) separately by including the dose of that allele per diploid genome (0, 1, or 2) as an independent variable in the regression. We obtained an estimate of effect size for each *C4* structure (e.g.*, AL-AL*) across all alleles that contained that given structure (e.g., *AL-AL 1* and *AL-AL 2*), by performing an inverse variance meta-analysis based on the effect size and standard error associated with each *C4* allele that contained the given structure. We included dosages of the risk allele of *HLA-DRB1*03:01* and *15:01* as covariates when testing association to *C4* conditional on these variants. In order to model a dominant effect of *03:01* and *15:01* on SLE risk, we assigned a dose of 0 for individuals lacking a risk allele and a dose of 1 for those who had at least 1 copy of the risk allele. We included DNA source (blood [n = 10,772] or buccal [n=325]) as a covariate in all of the association analyses.

Chapter 4

Structural variation in *C4*, a gene involved in synaptic pruning, affects gene expression in the brain and influences risk of schizophrenia

Introduction

The association of the *HLA* locus to schizophrenia has been a topic of investigation since 1974[62]. This association is the strongest, most well replicated, and most complex finding from genome-wide association studies (GWAS) of schizophrenia, as described in Chapter 1. In the latest GWAS meta-analysis of schizophrenia[22], the *HLA* association signal spanned several megabases and dozens to hundreds of genes, and no single variant within the locus could fully explain the association. It remains unclear which gene(s) and functional allele(s) underlie the association of the *HLA* locus to schizophrenia.

Schizophrenia has a typical age of onset in late adolescence to early adulthood[134], and the human brain undergoes significant morphological changes in the years leading up to this period[135]. Studies of *postmortem* human brain tissue have identified a reduction in dendritic spine density in multiple cortical regions in schizophrenia[136]. Genetic as well as epidemiological evidence have suggested a contribution of both neuronal and immunological processes to schizophrenia[22, 42, 67, 137, 138] but whether and how these processes may be mechanistically linked in the pathogenesis of schizophrenia is unknown.

We hypothesized that complex structural variation in the *C4* gene contributes to the *HLA* association signal in schizophrenia based on the following lines of reasoning. First, whereas the association of other phenotypes, such as rheumatoid arthritis, to the *HLA* locus has been ascribed to coding variants within classical *HLA* genes[139], these variants do not associate as strongly to schizophrenia as SNPs elsewhere in the extended *HLA* locus[35], suggesting that the causal variation could reside in other gene(s) within the locus. Second, *CSMD1* (*CUB and Sushi multiple domains 1*), which functions in promoting degradation of the activated fragment of C4[99] and inhibiting the classical complement pathway, was among the first loci to be identified from GWAS of schizophrenia[48]. Third, other genes that function in the complement pathway, *C1Q*, *C3*, and *CR3* (*complement receptor 3*), have been shown to play a role in the elimination of synapses during neurodevelopment[104, 105], a process that might contribute to the reduction in spine density in patients with schizophrenia[136].

We also reasoned that the complex form of structural variation in *C4* would present a powerful opportunity to investigate association to the *HLA* locus. The strong correlation between *HLA* SNPs has made it challenging to determine which variant(s) are responsible for driving the association to this locus. In contrast, the fact that the correlation of *C4* structural variants to individual SNPs is only partial (as described in Chapter 2) could provide additional power to assess independence of association to *C4* from association to other variants in the *HLA* locus.

C4A **expression in the brain increases with both** ***C4A*** **and** ***C4L*** **copy number**

We measured expression of *C4A* in *postmortem* brain samples from the Stanley Medical Research Institute (SMRI) collection, consisting of tissue from 5 brain regions (anterior cingulate cortex, orbital frontal cortex, parietal cortex, cerebellum, and corpus callosum) from 105 individuals. We used droplet-digital PCR (ddPCR, described in Chapter 2) to measure copy number of *C4* structural features in these samples and measured gene expression of *C4A* using reverse-transcriptase ddPCR (RT-ddPCR, described in Chapter 3).

C4A and *C4L* copy number both associated to *C4A* expression ($p < 10^{-10}$ from a linear model, **Figure 4.1a** and **4.1b**). Since *C4A* and *C4L* copy number were correlated ($r^2 = 0.50$ across the 105 samples), we next analyzed them conditionally upon each other. We regressed *C4A* expression against *C4A* copy number, and then fitted these residuals to *C4L* copy number. *C4L* copy number associated to *C4A* expression even after removing the effect of *C4A* copy number in this manner ($p < 0.01$, **Figure 4.1c**). In

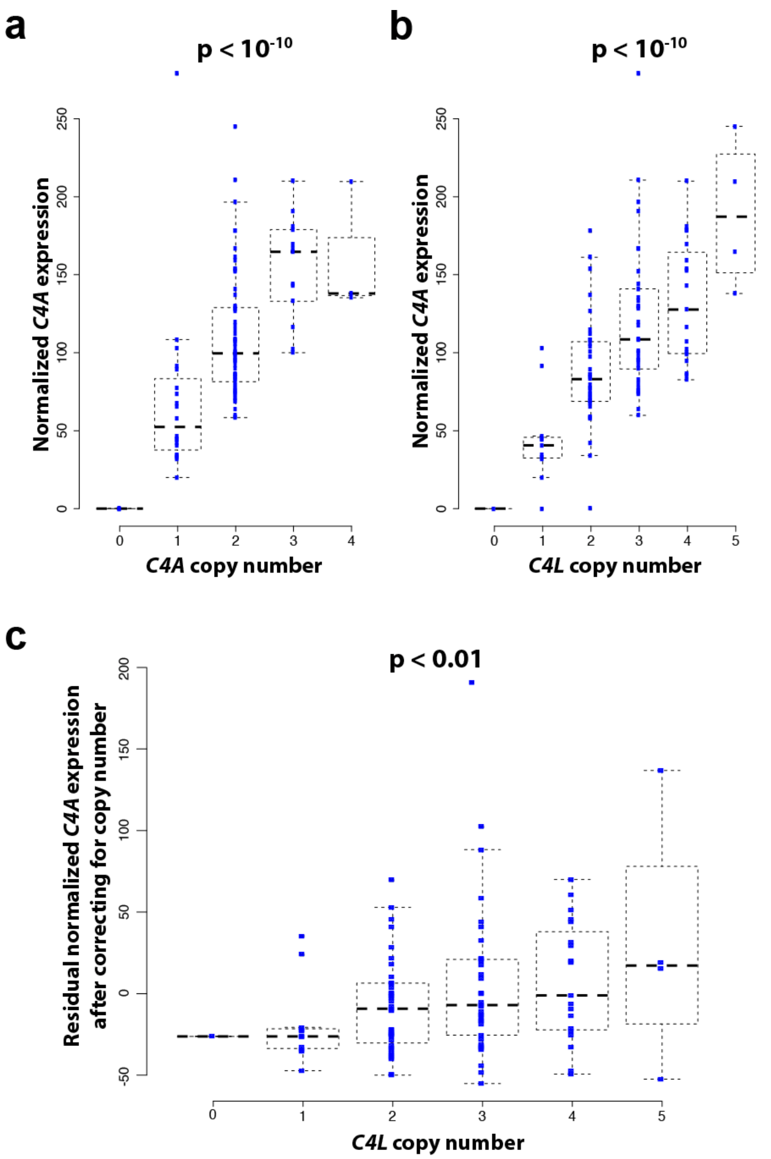


Figure 4.1. Relationship between *C4* structural variation and *C4A* expression in *postmortem* human brain tissue. (**a**) Association of *C4A* copy number to *C4A* expression. (**b**) Association of *C4L* copy number to *C4A* expression. (**c**) Association of *C4L* copy number to *C4A* expression after normalizing out the effect of *C4A* copy number. Each point in these plots represents a composite measurement of gene expression for an individual across five brain regions (details provided in the Methods section). P-values shown are from a linear model fitting expression to copy number.

a linear model that included both *C4A* and *C4L* copy number as explanatory variables, the effect size, β, per copy of *C4A* and *C4L* was nearly identical (β = 20.45 for *C4A*; 20.50 for *C4L*). *C4A* and *C4L* copy number, together, explained 42% of the variance in *C4A* expression.

We then related *C4A* expression to the four common *C4* structures. We calculated an effect size, β, associated with each of these four structures from fitting *C4A* expression to the number of chromosomes per diploid genome that carried that structure (0, 1, or 2). The resulting pattern of association was consistent with a contribution of *C4L* copy number in addition to *C4A* copy number to expression of *C4A* across multiple brain regions (**Figure 4.2a-f**). This result was in contrast to the finding in lymphoblastoid cell lines (Chapter 3, Figure 3.1b), in which only *C4A* copy number associated to *C4A* expression.

Variation in the expected expression of *C4A* and two additional, statistically independent effects contribute to the *HLA* association signal in schizophrenia

We analyzed the association of the *HLA* locus to schizophrenia in 40 cohorts (totaling 28,799 schizophrenia cases and 35,986 controls of European ancestry) from the Psychiatric Genomics Consortium[22]. We used the HapMap CEU reference panel that we created for imputation of *C4* structural variants (described in Chapter 2) for evaluating the contribution of *C4* to schizophrenia risk. We performed a mega-analysis that utilized individual-level genotype data from all of the samples, and tested association in a logistic regression framework that included study indicator variables to account for cohort-specific effects and principal components to control for population stratification.

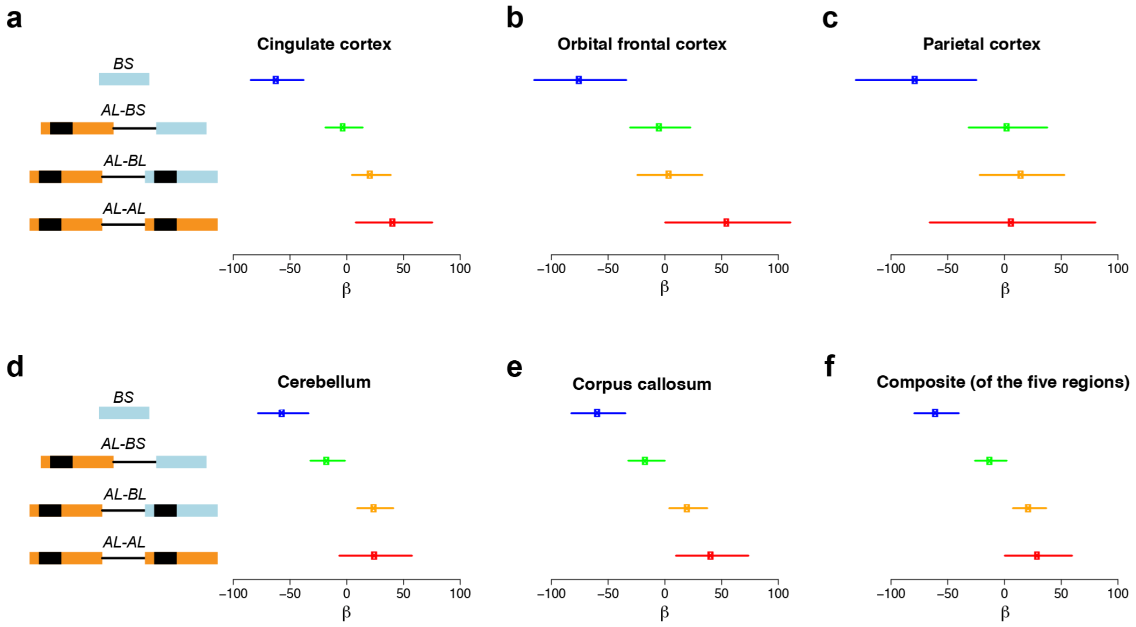


Figure 4.2. Association of the four common *C4* structures to *C4A* expression in (**a**) the cingulate cortex, (**b**) orbital frontal cortex, (**c**), parietal cortex, (**d**) cerebellum, and (**e**) corpus callosum. (**f**) Association to a composite measure of gene expression across the five brain regions. β measures the effect size calculated from fitting *C4A* expression to the number of chromosomes (0, 1, or 2) carrying that *C4* structure per diploid genome.

We tested association to 7,751 imputed SNPs and identified a strong, diffuse association of the extended *HLA* locus to schizophrenia (**Figure 4.3a**). The most strongly associated variant was rs13194504 (p = 5.5x10^{-28}), an intergenic SNP 4.6 kb downstream of the nearest gene, *tRNA_ala* (*transfer RNA Ala*). This SNP indexed an association to a long (approximately 2 Mb) haplotype on the telomeric end of the extended *HLA* locus (chromosome 6, 27.5-29.5 Mb).

Imputed copy number of *C4A* and *C4L* also associated strongly to schizophrenia (p = 6.3x10^{-19} for *C4A* and 7.2x10^{-23} for *C4L*). In a joint model, both *C4A* and *C4AL* copy number remained significant (p = 0.002 for *C4A* and 1.4x10^{-7} for *C4L*). Based on the influence of *C4A* and *C4L* copy number on expression of *C4A* in the brain, we hypothesized that the association of schizophrenia to expected *C4A* expression in the brain (E[*C4A* expression]) would be stronger than its association to *C4A* or *C4L* copy number alone. We calculated E[*C4A* expression] based on the linear model derived from fitting *C4A* expression in *postmortem* brain to *C4A* and *C4L* copy number. Consistent with our hypothesis, we identified a stronger association of schizophrenia to E[*C4A* expression], with higher expected expression associating to increased risk of schizophrenia ($p = 6.2 \times 10^{-25}$, **Figure 4.3a**).

The association to E[*C4A* expression] was largely independent of the association indexed by rs13194504 ($r^2 = 0.18$, **Figure 4.3b** and **4.3c**). In analyzing these effects in a joint model, E[*C4A* expression] associated to schizophrenia more strongly than any other variant in the *HLA* locus ($p = 2.4 \times 10^{-10}$, **Figure 4.3d**), and rs13194504 remained highly significant in a joint model with E[*C4A* expression] ($p = 1.5 \times 10^{-13}$, **Figure 4.3e**).

We identified a third, statistically independent effect in the centromeric end of the extended *HLA* locus. The variant most strongly associated in a joint model with rs13194504 and E[*C4A* expression] was rs210133 ($p = 1.8 \times 10^{-8}$, **Figure 4.3f**), an intergenic SNP 3.5 kb downstream of *BAK1* (*BCL2-antagonist/killer 1*). In addition, rs9461856, an intronic SNP in *SYNGAP1* (*synaptic Ras GTPase activating protein 1*) was also highly significant ($p = 1.0 \times 10^{-7}$) in a joint model with rs13194504 and E[*C4A* expression], but was only weakly correlated with rs210133 ($r^2 = 0.21$). No genome-wide significant associations remained in the *HLA* locus conditional on rs13194504, E[*C4A* expression] and rs210133 (**Figure 4.3g**), but all three effects remained highly significant in a joint model ($p = 1.5 \times 10^{-13}$, 8.9×10^{-9}, and 1.8×10^{-8}, respectively).

Figure 4.3. Association of the *HLA* locus to schizophrenia. (**a**) Association of schizophrenia to 7,751 imputed SNPs and imputed copy number of *C4* structural features based on analysis of 28,799 schizophrenia cases and 35,986 controls of European ancestry. E[*C4A* expression] indicates the expected expression of *C4A* in the brain, calculated based on imputed copy number of *C4A* and *C4L*, using the linear model built from relating *C4* structural variation to *C4A* expression in human *postmortem* brain samples. (**b**) Association of schizophrenia to SNPs in the *HLA* locus colored by correlation (r^2) to rs13194504, the SNP most strongly associated to schizophrenia in the *HLA* locus. (**c**) Association of schizophrenia to SNPs in the *HLA* locus, colored by correlation to E[*C4A* expression]. Association to SNPs in the *HLA* locus with (**d**) rs13194504, (**e**) E[*C4A* expression], (**f**) rs13194504 and E[*C4A* expression], (**g**) rs13194504, E[*C4A* expression], and rs210133 as covariates in the logistic regression model. Dashed red line indicates the threshold for genome-wide significance ($p = 5 \times 10^{-8}$).

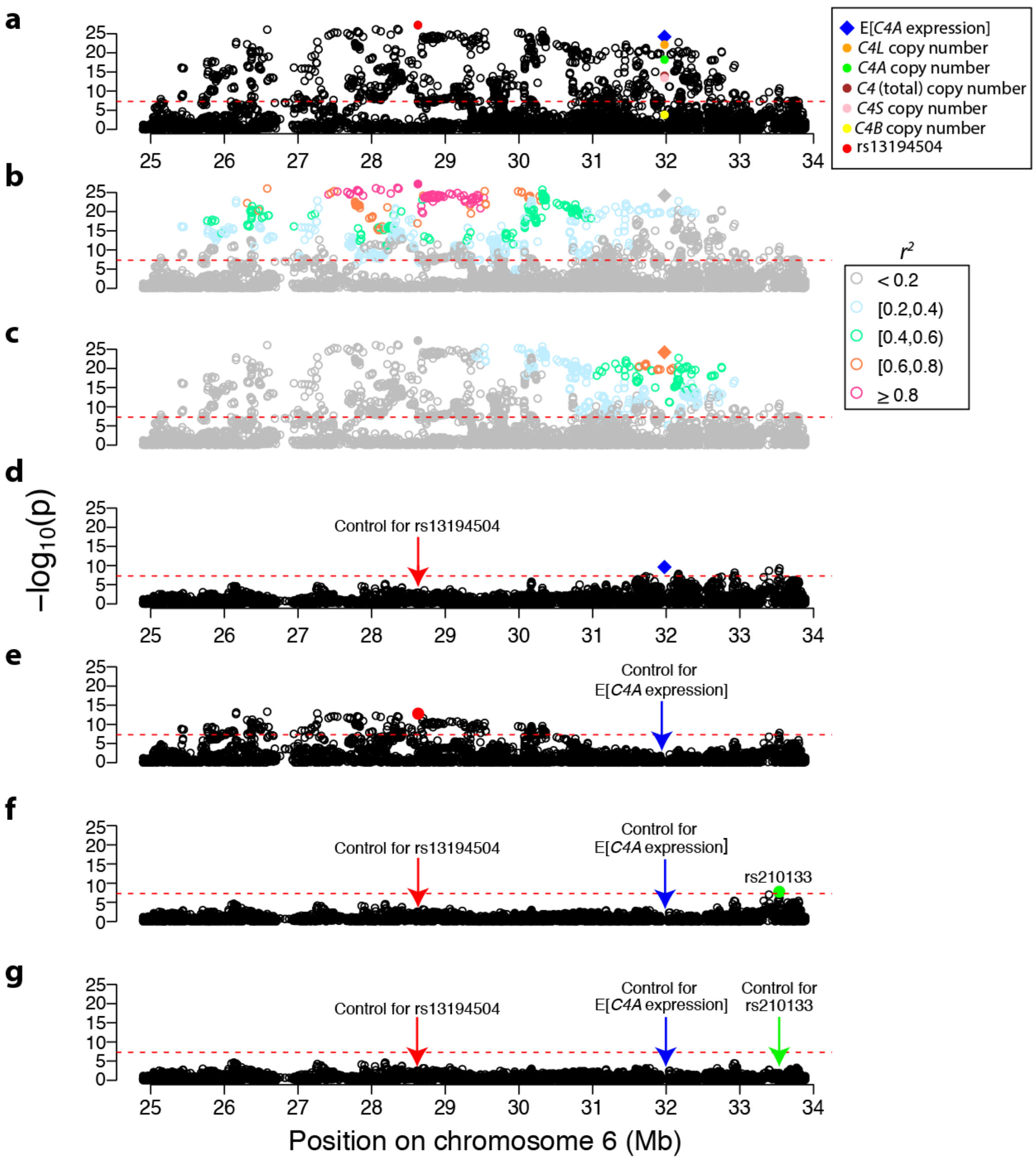


Figure 4.3 (continued)

***C4* structures associate to schizophrenia in proportion to their effect on *C4A* expression, with minimal influence of their *HLA* haplotype background**

We next analyzed association of schizophrenia to *C4* structures, taking into account variation in their *HLA* haplotype backgrounds, in order to distinguish between the effect of *C4* structure and effects elsewhere in the *HLA* locus. We encoded *C4* structures segregating on different *HLA* haplotypes (Chapter 2, Figure 2.10) as separate alleles in our reference panel for imputation. We then imputed these *C4* structural alleles into 28,799 schizophrenia cases and 35,986 controls analyzed above, using the surrounding SNP genotype data as the backbone for imputation, and evaluated association to each of these alleles.

Across 11 *C4* structural alleles (each present at a frequency of more than 3% in our reference panel), the odds ratio was defined by the *C4* structure that it contained, with minimal influence of the *HLA* haplotype on which it segregated (**Figure 4.4a**). This allelic series of effects on schizophrenia risk corresponded to the effect of *C4* structures on *C4A* expression in the brain (Figure 4.2) – across the four common *C4* structures, the greater the *C4A* expression in the brain, the greater the risk of schizophrenia. The effect sizes associated with the four common *C4* structures spanned an odds ratio of 0.84 (95% CI: 0.81- 0.87) associated with *BS* to 1.09 (95% CI: 1.05-1.13) associated with *AL-AL*. The correlation of effect sizes across the 11 *C4* alleles to the number of copies of *C4A* and *C4L* that each allele contained was statistically significant (p = 5.7x10^{-5}). In contrast, the heterogeneity of effect sizes across alleles that contained the same *C4* structure on different *HLA* haplotypes was not statistically significant (p = 0.55 for *AL-AL* alleles; p = 0.93 for *AL-BL* alleles; and p = 0.06 for *AL-BS* alleles).

We then analyzed the association of schizophrenia to *C4* in a joint model with rs13194504, the variant that indexed the association to the telomeric end of the extended *HLA* locus. Across the 11 *C4* structural alleles, the allelic series of effects

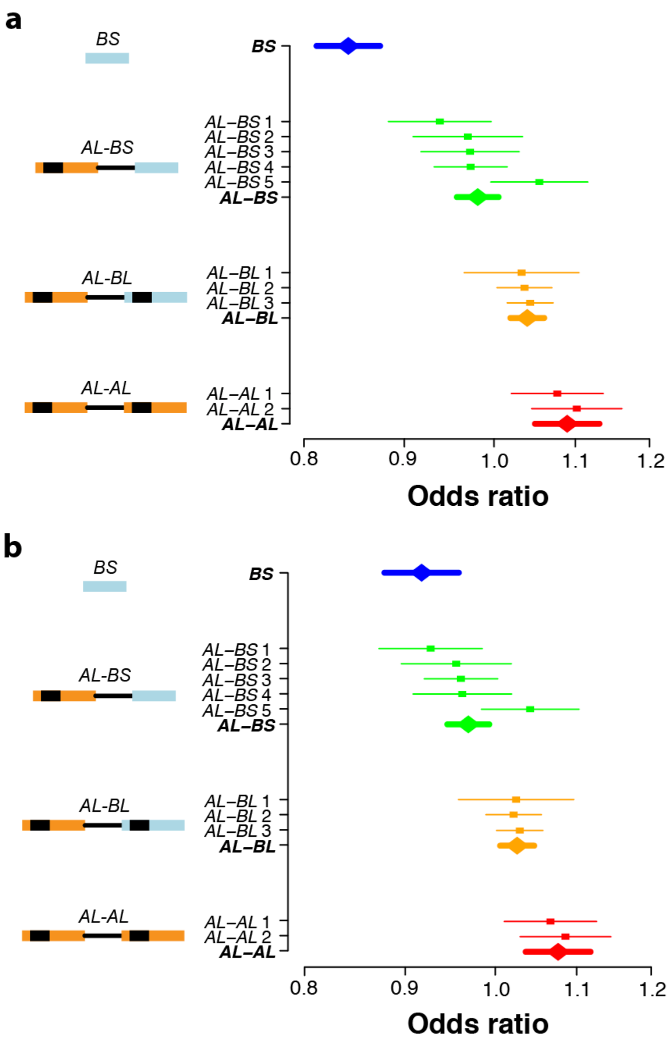


Figure 4.4. Association of *C4* structures on different *HLA* haplotype backgrounds to schizophrenia. (**a**) The effect size associated with each *C4* structural allele was predicted by its *C4A* and *C4L* copy number and the association of *C4* structures to *C4A* expression in the brain (Figure 4.2). (**b**) The pattern of association to *C4* was preserved when including rs13194504 in the model, with a moderation of the effect size associated with *BS*. *C4* structures suffixed with a different number (e.g., *AL-BL* 1 and *AL-BL* 2) represent alleles with the same structural content segregating on different *HLA* haplotype backgrounds, and they correspond to the annotation in Figure 2.10 in Chapter 2. The diamond symbols in bolder colors represent the effect size from a meta-analysis of the effect of each allele containing the same *C4* structural content.

remained intact in a joint model that included rs13194504 (**Figure 4.4b**). The main effect of including rs13194504 in the model was to moderate the outsized protective effect associated with the *BS* structure.

C4 structural variation, analyzed using ddPCR, associates strongly to schizophrenia

We next sought to evaluate the accuracy of our imputation-based results using molecular measurements of *C4* copy number from ddPCR. We measured copy number of *C4A*, *C4L*, and *C4S* in 2,195 cases and 2,485 controls of Swedish ancestry, one of the largest samples from the PGC data set. We were also able to infer copy number of *C4B* and total *C4* from these measurements, given the following relationship: copy number of *C4A* + *C4B* = *C4L* + *C4S* = total *C4*.

Across the samples that were genotyped using ddPCR, the imputed copy number for *C4A* and *C4L* correlated with the ddPCR-based copy number measurements with an r^2 of 0.59 and 0.81, respectively. These results were consistent with the predictions from leave-one-out trials described in Chapter 2 (Figure 2.12). E[*C4A* expression] based on imputed genotypes correlated with that based on ddPCR measurements with an r^2 of 0.78.

The association of schizophrenia to *C4L* copy number ($p = 8.4x10^{-8}$), based on ddPCR measurements, was at least as strong as to any individual SNP in the extended *HLA* locus (**Figure 4.5a**). E[*C4A* expression], which was calculated based on ddPCR-derived copy number measurements, also associated strongly ($p = 1.1x10^{-6}$). We determined the *C4* structures underlying the copy number of *C4* structural features measured in each individual, using ddPCR measurements together with SNP genotype data for these samples (as described in the Methods section). We tested association to the four most common *C4* structures, and the effect size associated with these structures spanned a wide range with an odds ratio of 0.74 (95% CI: 0.65-0.85) for *BS* to 1.17 (95% CI: 1.07-1.27) for *AL-BL* (**Figure 4.5b**). The pattern of association of schizophrenia to these four structures was in proportion to their effect on *C4A* expression in the brain (Figure 4.2).

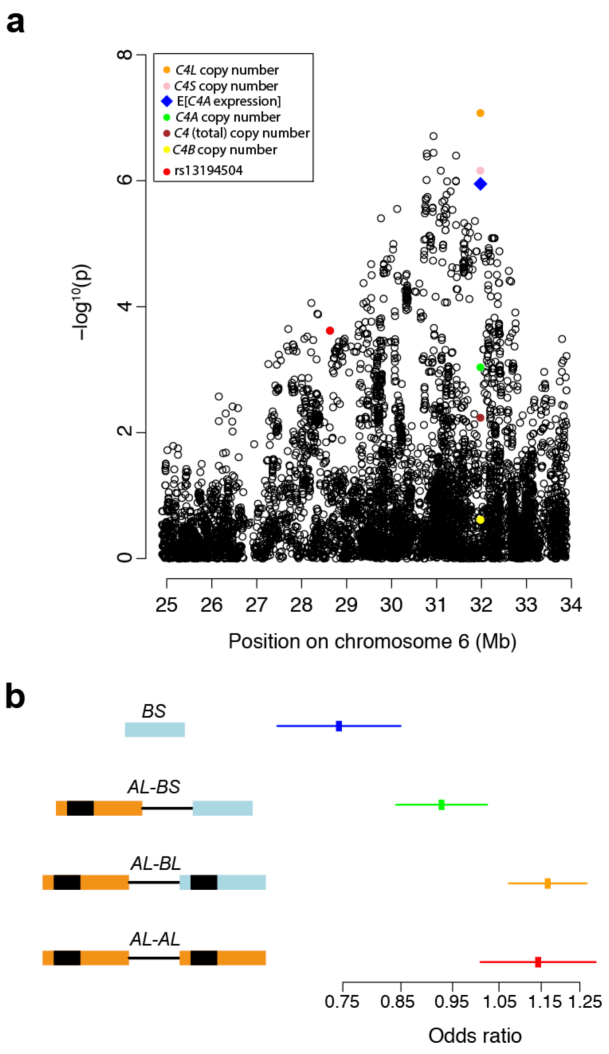


Figure 4.5. Association analysis in 2,195 cases and 2,485 controls of Swedish ancestry based on molecular measurements of *C4* copy number using ddPCR. (**a**) Association of schizophrenia to copy number of *C4* structural features, E[*C4A* expression], and to SNPs in the extended *HLA* locus. (**b**) Association to common *C4* structures that were determined using ddPCR-based copy number measurements together with SNP genotype data.

***C4A* expression is higher in *postmortem* brains samples from patients with schizophrenia than in unaffected controls**

Based on the finding that *C4* structures associating with increased expression of *C4A* also associated with increased risk of schizophrenia, we hypothesized that the expression of *C4A* would be higher in *postmortem* brain samples from patients with schizophrenia than in unaffected controls. To test this prediction, we related measurements of *C4A* expression in the SMRI cohort (35 unaffected controls, 35 patients with schizophrenia, 20 patients with bipolar disorder with psychotic features and 10 patients with bipolar disorder without psychotic features), to their clinical diagnosis. *C4A* expression across the five brain regions was higher in samples from patients with schizophrenia than in those from unaffected controls ($p = 0.002$, **Figure 4.6a**). The difference in expression of *C4A* between patients with schizophrenia and bipolar disorder patients with psychosis was not statistically significant ($p = 0.18$), but the difference in expression between bipolar disorder patients with and without psychosis trended toward significance ($p = 0.06$).

Since *C4A* and *C4L* copy number associated to both case-control status as well as expression, we next asked whether the difference in *C4A* expression between unaffected controls and patients with schizophrenia would still be significant after removing the effect of *C4* structural variation. We compared residual expression of *C4A*, after fitting it to *C4A* and *C4L* copy number, across the four diagnostic groups. The difference in expression between unaffected controls and patients with schizophrenia remained statistically significant ($p = 0.003$, **Figure 4.6b**) and the difference in expression between bipolar disorder patients with and without psychotic features was nominally significant ($p = 0.04$).

Figure 4.6. Expression of *C4A* in *postmortem* brain samples from control individuals (n= 35) and patients with schizophrenia (n = 35), bipolar disorder with psychotic features (n = 20) and bipolar disorder without psychotic features (n = 10), (**a**) before and (**b**) after removing the effect of *C4A* and *C4L* copy number on *C4A* expression. A composite measure of *C4A* expression across the five brain regions analyzed was used in these analyses. P-values are from a Mann-Whitney test with alternative hypotheses of lower expression in control individuals compared to patients with schizophrenia; unequal expression between patients with schizophrenia and bipolar disorder with psychosis; and higher expression in patients with bipolar disorder with psychotic features compared to those with bipolar disorder without psychotic features.

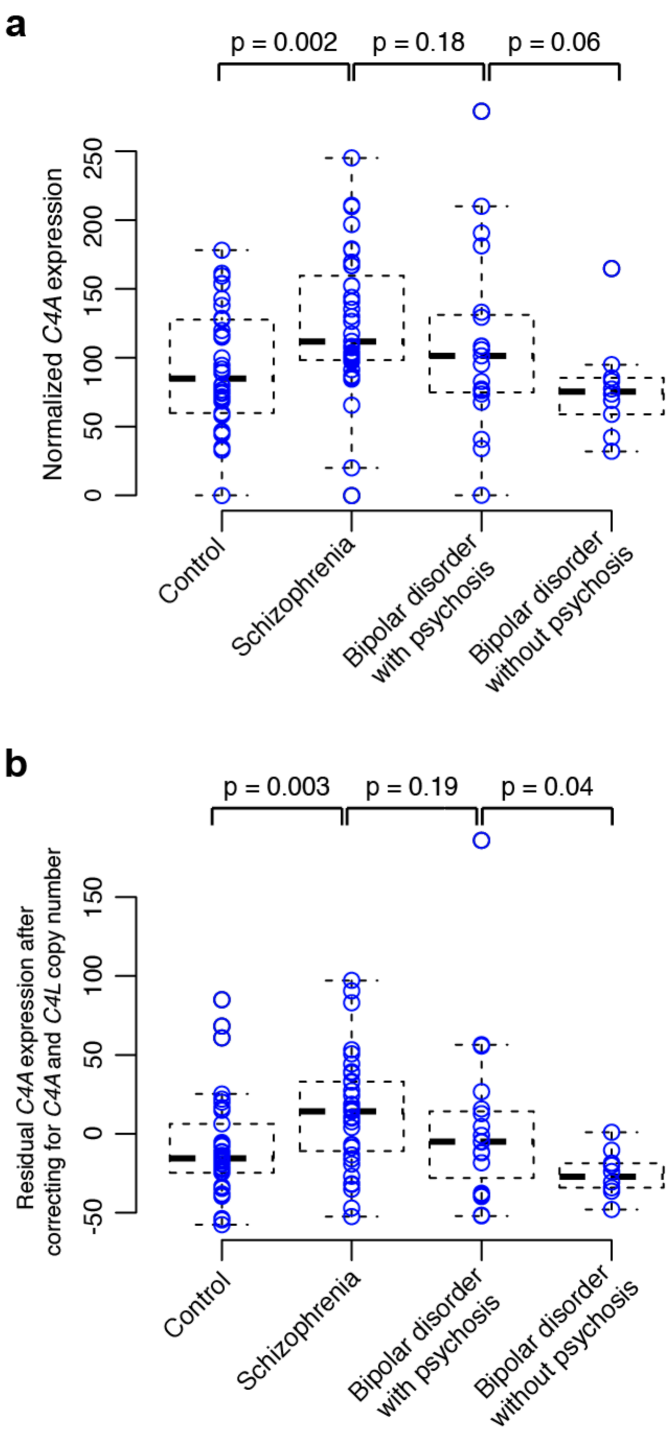


Figure 4.6 (continued)

We then evaluated whether the expression of other genes that function in the classical complement pathway was different between unaffected controls and patients with schizophrenia, and the findings are summarized in **Table 4.1**. The expression of *C4B*, a paralog of the *C4A* gene that is more than 99% identical at the nucleotide level, was not different between the two groups (p = 0.37, before correcting for copy number of *C4B*; p = 0.17 after correcting for copy number). Total expression of *C4* (the sum of *C4A* and *C4B* expression), however, was higher in patients with schizophrenia compared to unaffected controls, before as well as after correcting for *C4* copy number (p = 0.009 and 0.007, respectively). The expression of *C1QB*, which functions upstream of *C4* in the classical complement pathway (Chapter 1, Figure 1.4), showed a trend toward higher expression in patients with schizophrenia (p = 0.09). Similarly, *C2*, which also functions in the classical complement pathway, achieved a nominal level of significance (p = 0.04) for higher expression in patients with schizophrenia. We then asked whether the expression of *CSMD1*, which encodes an inhibitor of the classical complement pathway[98, 99], shows differential expression in the opposite direction as *C4A*, *C1QB*, and *C2* – with higher expression in controls compared to patients with schizophrenia. Of the two brain regions analyzed, the expression was indeed higher in unaffected controls in the cingulate cortex (p = 0.002), but not in the cerebellum (p = 0.76). Expression of *C3*, which functions in the alternative complement pathway in addition to the classical pathway, did not differ between cases and controls across the four brain regions tested (p = 0.34).

C4 expression increases in the brain as mice and humans enter adulthood

In order to understand how expression of *C4* varies with age – and particularly around the period corresponding to the typical age of onset of schizophrenia – we profiled its expression in brains from wild type mice across developmental age groups,

Table 4.1. Differences in expression of genes that function in the early steps of the complement pathway in *postmortem* brain samples from unaffected controls and patients with schizophrenia. P-values are from a Mann-Whitney (non-parametric) test of the specified alternative hypothesis, testing association across a composite measure of expression across the indicated brain regions unless otherwise stated. [a] Before removing the effect of variation in copy number; [b] after removing the effect of variation in copy number.

Gene	Function in the complement pathway	Brain regions tested	Hypothesis	p-value
C4A	Part of the classical pathway C3 convertase, an enzymatic complex that activates the central component of the complement pathway	Cingulate cortex Orbital frontal cortex Parietal cortex Cerebellum Corpus callosum	Higher expression in patients with schizophrenia	0.002[a] 0.003[b]
C4B	Part of the classical pathway C3 convertase	‘’	“	0.37[a] 0.17[b]
C4 (combined expression of *C4A* and *C4B*)		‘’	“	0.009[a] 0.007[b]
C1QB	Initial component of the classical complement pathway; part of the complex that activates C4	“	“	0.09
C2	Part of the classical pathway C3 convertase, together with C4	“	“	0.04
CSMD1	An inhibitor of the classical complement pathway	Cingulate cortex Cerebellum	Higher expression in unaffected controls	0.002 (cingulate cortex) 0.76 (cerebellum)
C3	Central component of the complement pathway; functions in both alternative and classical pathways of the complement system	Cingulate cortex Parietal cortex Cerebellum Corpus callosum	Higher expression in patients with schizophrenia	0.34

using RT-ddPCR. *C4* expression increased with age across the age groups analyzed, between postnatal day 3 (P3) to P60, in the mouse midbrain (**Figure 4.7a**)**.** We also measured *C4* expression in other brain regions from P28 and P60 animals and found a two-fold increase in expression between these two age groups in the hippocampus, but not in the cerebellum, temporal cortex, or visual cortex. We then analyzed human *C4* expression in the publicly available BrainSpan data set[140], which contains transcriptome-wide expression data from RNA-sequencing analysis of the developing human brain. Despite not having been corrected for variation in copy number (and genetic variation elsewhere in the genome) across individuals, a pattern of increasing *C4* expression into adulthood was identifiable in the dorsolateral prefrontal cortex (**Figure 4.7b**) and ventrolateral prefrontal cortex (**Figure 4.7c**) as well as in the striatum, anterior cingulate cortex and orbitofrontal cortex, but not in other regions (cerebellar cortex, and primary visual cortex).

Neurons and microglia express complementary components of the complement pathway

We next sought to identify which cell type(s) within the brain express *C4A* and other genes of the complement pathway. We profiled their expression using RT-ddPCR in cultured primary human cortical neurons, microglia, astrocytes, and oligodendrocytes (details about the source of these cells are provided in the Methods section). Given the role of microglia as the resident immune cell type of the brain, we expected complement genes to be expressed in these cells. Indeed, they expressed relatively high levels of *C1QB*, *C2*, and *C3*, but little-to-no *C4A*, *C4B*, or *CSMD1* (**Figure 4.8**). *C4A* was, however, expressed in neurons (approximately eight-fold higher than its expression in microglia) as well as in oligodendrocytes. Neurons and oligodendrocytes also expressed *CSMD1*.

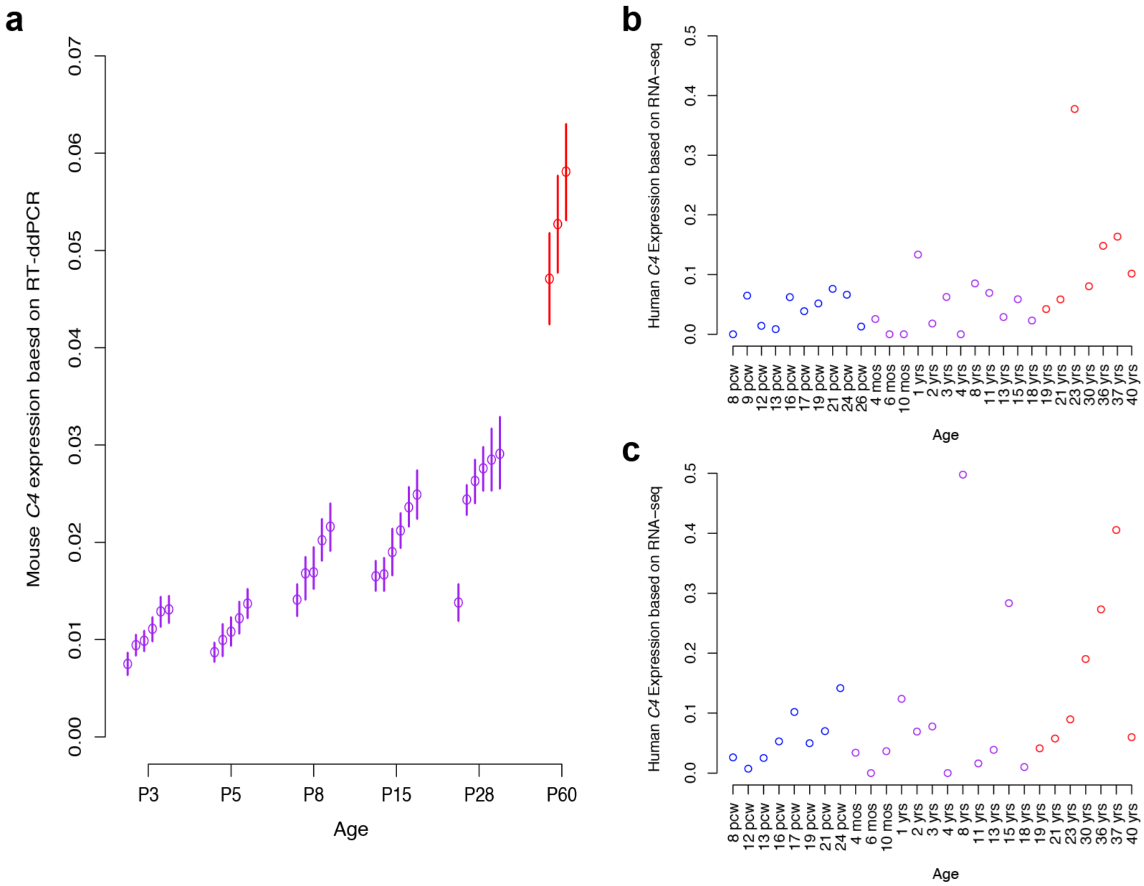


Figure 4.7. Expression of *C4* across developmental ages in mouse and human brains. (**a**) Expression of *C4* in the midbrain of C57BL/6 wild type mice, as measured by RT-ddPCR. Error bars indicate 95% Poisson confidence intervals. Age in postnatal days is shown on the x-axis. Each point represents a measurement from a different mouse. (**b**) Expression of human *C4* in the dorsolateral prefrontal cortex and (**c**) ventrolateral prefrontal cortex based on RNA-sequencing data from the BrainSpan data set[140]. Unlike the mouse data in (a), in which copy number of *C4* did not vary from animal-to-animal, the human data in (b) and (c) are not corrected for variation in *C4* copy number. Periods corresponding to fetal development are shown are blue, childhood and adolescence in purple, and adulthood (age > 18 years) in red. pcw, post-conception weeks; mos, months; yrs, years.

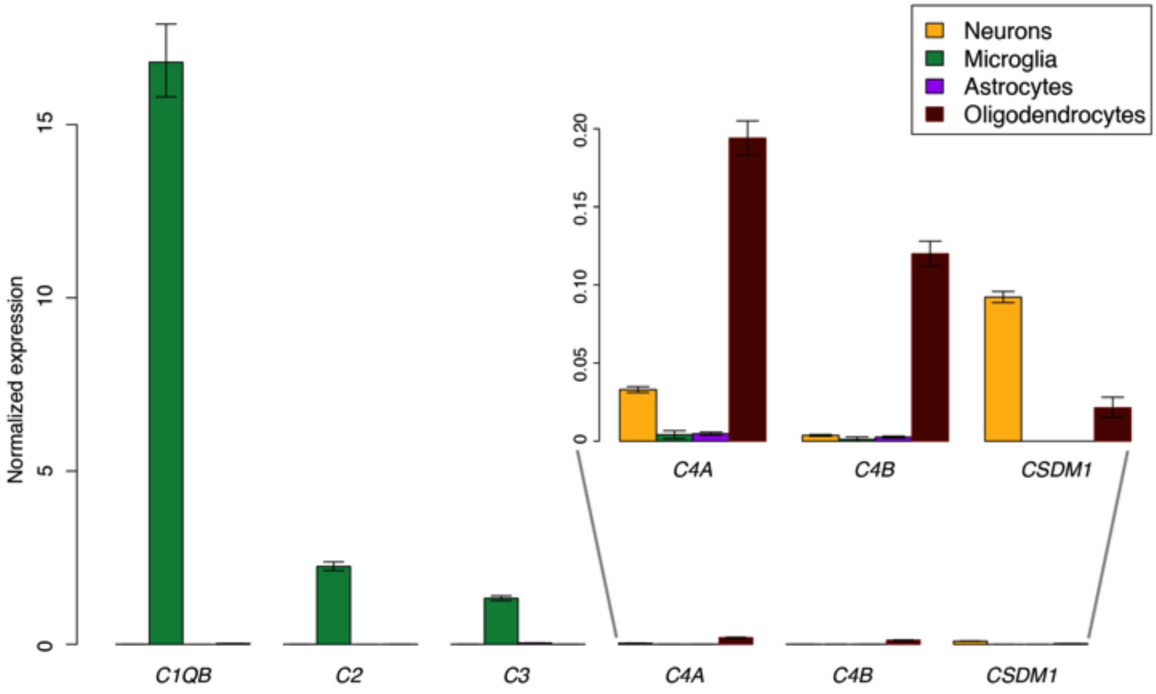


Figure 4.8. Expression of complement genes in cultured primary human neurons, microglia, astrocytes, and oligodendrocytes. Error bars represent 95% Poisson confidence intervals from RT-ddPCR.

Given the low baseline expression of *C4A* in microglia, we tested whether its expression could be upregulated upon stimulation of microglia with lipopolysaccharide (LPS), a component of the outer membrane of Gram-negative bacteria, or polyinosine-polycytidylic acid [poly(I:C)], a viral-like double-stranded RNA. Although the expression of *C1QB, C2,* and *C3* was detected at a 2-4 fold higher level in microglia treated with LPS (**Figure 4.9a**) or poly(I:C) (**Figure 4.9b**), the expression of *C4A* and *C4B* remained low.

The finding that *C4A* was expressed in different cell types than *C1QB*, *C2*, and *C3* led us to investigate whether the expression of these genes was still correlated across samples. To test this, we analyzed their expression in 105 *postmortem* brain

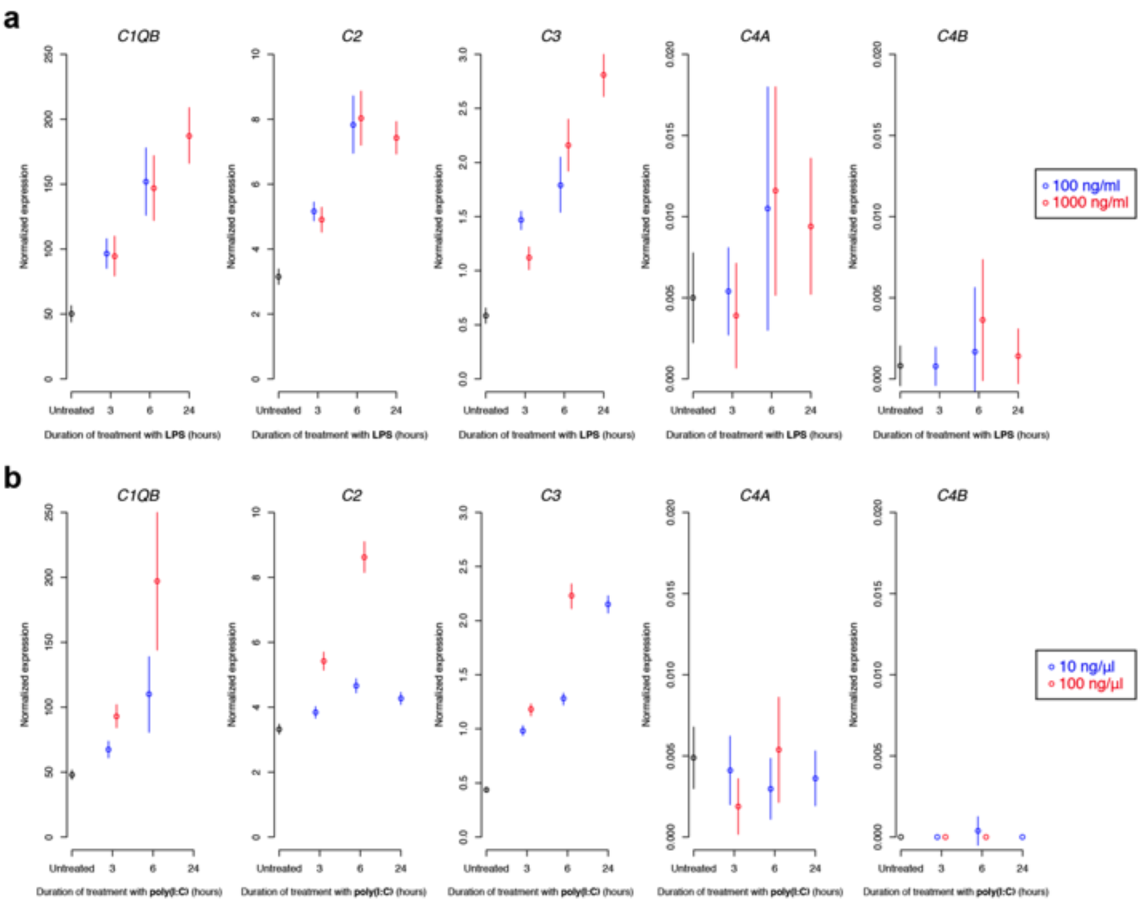


Figure 4.9. Expression of complement genes in primary human microglia treated with (**a**) lipopolysaccharide (LPS), and (**b**) polyinosine-polycytidylic acid [poly(I:C)] with concentrations and durations of treatment as indicated. Error bars are from 95% Poisson confidence intervals. Expression of *C1QB*, *C2*, and *C3* was detected at 2-4 fold higher levels with treatment, whereas expression of *C4A* and *C4B* remained low (note the variation in y-axis scales). The 24-hour time point from treatment with poly(I:C) is not shown for *C1QB* as the expression was too high to be reliably quantified using RT-ddPCR.

samples (from the SMRI collection). We first removed the effect of copy number variation on expression by calculating the residuals from fitting *C4A* expression to *C4A* and *C4L* copy number, and tested correlation between these residuals and the expression of other genes. As expected, residual *C4A* expression correlated strongly with residual *C4B* expression (**Figure 4.10a**). In addition, residual *C4A* expression also correlated to expression of *C1QB* (**Figure 4.10b**), *C2* (**Figure 4.10c**), and to a lesser extent, *C3* (**Figure 4.10d**).

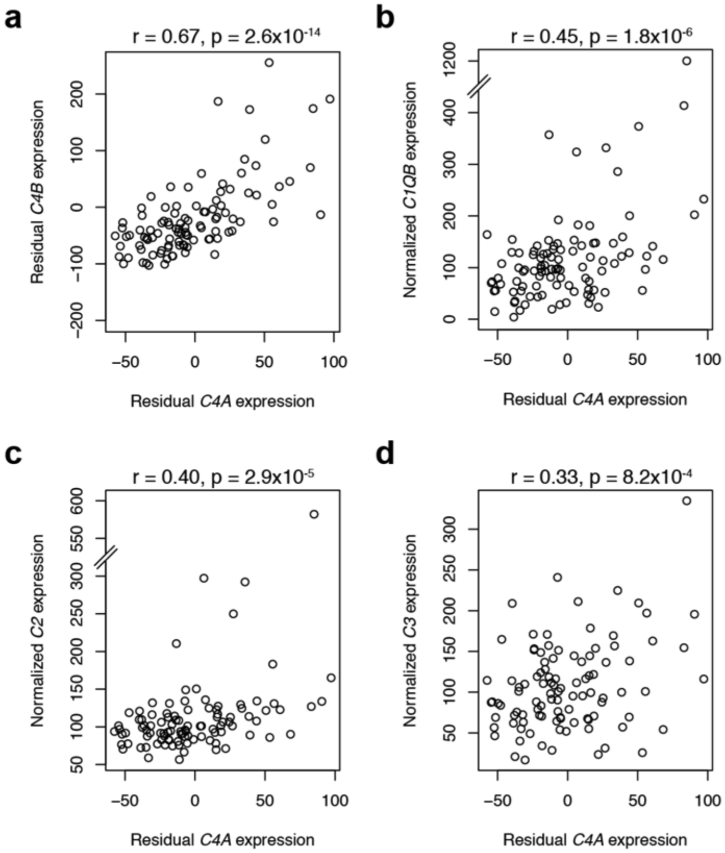


Figure 4.10. Correlation between expression of *C4A* and other genes of the complement pathway in *postmortem* brain samples. Expression of (**a**) *C4B*, (**b**) *C1QB*, (**c**) *C2*, and (**d**) *C3* are plotted against *C4A* expression, based on a composite measure of gene expression across multiple brain regions. Residual expression refers to the residuals values that remained after fitting gene expression to copy number in a linear model.

Mice lacking *C4* have a deficit in synaptic pruning that is rescued by human *C4*

Previous work has identified a role for *C1q*, *C3*, and *CR3* (*complement receptor 3*, a receptor for C3) in synaptic pruning[104, 105], the process by which the excessive number of synapses that are initially formed during development are eliminated, while a subset are maintained and strengthened. The position of C4 between C1q and C3 in the classical complement pathway (Chapter 1, Figure 1.4) suggests that *C4* is also likely to contribute to synaptic pruning. However, little is known about the expression of *C4* during the process of synaptic refinement, and whether *C4* plays a role in synapse elimination has not yet been investigated.

In order to test whether *C4* contributes to synaptic pruning, we used the well-studied retinogeniculate system[104, 105, 141, 142]. In this system, there is initially an overabundance of synapses formed between retinal ganglion cells (RGCs) and neurons within the dorsal lateral geniculate nucleus of the thalamus (dLGN), with dLGN neurons being innervated by up to ten RGC axons. Inputs from each eye compete for synapses in the dLGN, with some synapses weakening and becoming eliminated and others becoming mature. Within the first 10 days, axons from RGCs organize into eye-specific territories in the dLGN and by postnatal day 30, each dLGN relay neuron only receives input from 1-2 RGC axons.

We visualized the segregation of inputs from RGCs in the dLGN using an anterograde tracing technique in which mice receive an injection in each eye with a different colored dye conjugated to the cholera toxin-β subunit. We quantified the degree of overlap between inputs from each eye in the dLGN in P10 mice using an *R*-value analysis[143]. The *R*-value is the $\log_{10}$ ratio of fluorescence intensity from the ipsilateral and contralateral channel, $\log_{10}(F_{ipsi}/F_{contra})$. The higher the variance of this value, the greater the degree of segregation, and lower the overlap between the ipsilateral and contralateral inputs. Compared to wild type littermates, *C4* $^{-/-}$ mice had lower *R*-value variance (p = 0.004, Mann-Whitney test, **Figure 4.11a,b**), indicating greater overlap between RGC inputs from each eye. The degree of deficit in *C4* $^{-/-}$ was similar to that previously reported for *C1q* $^{-/-}$ and *C3* $^{-/-}$ mice[104, 141]. In addition, *C4* heterozygote mice, with one wild type copy of mouse *C4*, had a significantly higher *R*-value variance than *C4* $^{-/-}$ mice (p = 0.003, **Figure 4.11b**). Wild type mice had a higher mean *R*-value variance than *C4* heterozygote mice, but this difference did not reach statistical significance (p = 0.10)

We then tested whether human *C4*, with its endogenous regulatory sequence retained, could rescue this phenotype in *C4* $^{-/-}$mice*.* We generated transgenic mice that contained human *C4* in a background of mouse *C4* $^{-/-}$*,* using a human BAC clone that contains one copy each of *C4AL* and *C4BL* (*AL-BL*)*.* Screening of transgenic mice for copy number of human *C4* revealed the presence of two copies of *C4AL* and one copy of *C4BL* (potentially arising from a recombination event resulting in a partial duplication). Compared to *C4* $^{-/-}$ littermates, these transgenic mice had a significantly higher *R*-value variance (p = 0.004, **Figure 4.11 b**), similar to wild type mice.

Discussion

We have identified an allelic series of structural variants in the *C4* gene that contributes to the risk of schizophrenia and its association to the *HLA* locus. We found that *C4* structures with higher copy number of *C4A* and *C4L* associated with increased risk of schizophrenia, resembling their influence on *C4A* expression in *postmortem* human brain tissue. Furthermore, we developed a novel form of association testing that exploited the segregation of *C4* structures on different *HLA* haplotype backgrounds to show that their effect on schizophrenia risk is unlikely to be a result of variation elsewhere in the *HLA* locus. We have also shown that both mouse and human *C4* contribute to synaptic pruning in the mouse retinogeniculate system.

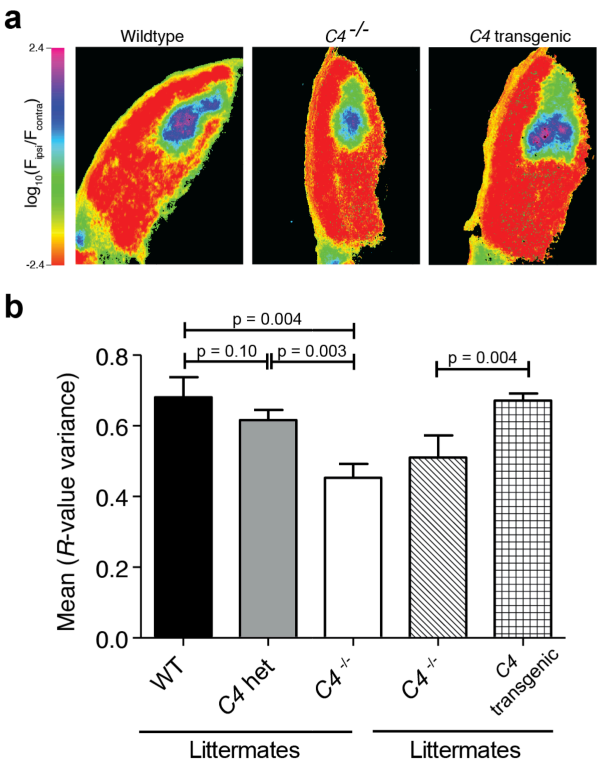


Figure 4.11. Mice lacking *C4* exhibit a synaptic pruning deficit that is rescued by human *C4*. (**a**) Representative images of the dorsal lateral geniculate nucleus (dLGN) from wild type*, C4* $^{-/-}$, and *C4* transgenic mice in which cholera toxin-conjugated dyes of two different colors were intraocularly injected. The *C4* transgenic mice used in these experiments had three copies of human *C4* (two *C4A* and one *C4B*; all three genes were of the long form, *C4L*) in a mouse *C4* $^{-/-}$ background. The images are pseudocolored to show the *R*-value for each pixel. The *R*-value is the $\log_{10}(F_{ipsi}/F_{contra})$, where F_{ipsi} is the fluorescence intensity in the ipsilateral channel and F_{contra} is the fluorescence in the contralateral channel. (**b**) Quantification of the variance of the *R*-value for each group. WT, two wild type copies of mouse *C4*; *C4* het, 1 wild type copy of mouse *C4*; *C4* $^{-/-}$, absence of wild type mouse *C4*. n = 5 WT, 7 *C4* heterozygotes and 5 *C4* $^{-/-}$ mice in the first set of littermates; 5 WT and 5 *C4* transgenic mice in the second set of littermates. P-values are from a Mann-Whitney test with an alternative hypothesis of higher *R*-value variance in WT compared to *C4* heterozygotes or *C4* $^{-/-}$ mice; higher in *C4* heterozygotes compared to *C4* $^{-/-}$; higher in *C4* transgenic mice compared to *C4* $^{-/-}$.

Genome-wide association studies of schizophrenia and other phenotypes typically identify loci that harbor multiple genes (3-4 on average in the latest GWAS meta-analysis of schizophrenia[22]). In most cases, neither the identity of the genes and variants within these loci that are driving the association nor their functional consequences are known. Similarly, studies of rare CNVs in schizophrenia have identified associations to large genomic segments, containing dozens of genes (e.g., deletions of the 22q11.2 locus in schizophrenia affect the dosage of up to 43 genes[37]). In contrast, our work has identified a series of alleles within a single gene that contribute both to its expression and to the risk of schizophrenia.

The finding that structural variation in *C4* influenced its expression in the brain enabled us to explicitly model the effect of these variants on *C4A* expression when testing their association to schizophrenia. Given that our analysis of the PGC data set was based on imputation, and that the imputed genotypes imperfectly capture the effect on *C4A* expression, it is possible that the association could be strengthened with more accurate genotyping of *C4* structural variation.

Our results revealed that copy number of *C4A* as well as *C4L* contributed to *C4A* expression in the brain and to schizophrenia risk. While it is expected that the dosage of the *C4A* gene would affect its expression in the brain (as it did in lymphoblastoid cell lines, Chapter 3), the finding that the number of copies of the Human Endogenous Retrovirus (HERV), the 6.4 kb insert that defines the *C4L* form, correlated with *C4A* expression is more surprising. However, this result is consistent with the recent finding that a HERV insertion in another gene, *PRODH*, is bound by the transcription factor SOX2 and serves as an enhancer in the brain[144]. The influence of HERV copy number on *C4A* expression also suggests a potential functional mechanism by which variation in *C4L* copy number could influence the risk of schizophrenia, but we do not rule out alternative mechanisms. For example, the HERV sequence, which is present in an antisense orientation in relation to the *C4* transcript, could mediate a gene-environment interaction in which variation in its copy number may influence the susceptibility to exogenous retroviruses[145, 146]. In this scenario, the inflammatory response to these viruses during susceptible periods of neurodevelopment may influence predisposition to later developing schizophrenia.

In addition to the association of schizophrenia to *C4* structural variation, we identified two effects within the *HLA* locus that were not explained by association to (imputed) *C4* structural variants. The presence of multiple effects in the *HLA* locus is consistent with its outsized strength of association to schizophrenia (Chapter 1, Figure 1.1b), though independently, the *C4* locus and rs13194504 were still among the genome's strongest signals ($p = 2.4 \times 10^{-10}$ and 1.5×10^{-13}, respectively, from a joint model).

The functional variants responsible for the association to rs13194504 and rs210133 remain to be identified. rs13194504 indexed a strong association to a long haplotype (27.5-29.5 Mb on chromosome 6), and it is possible that variation in class I *HLA* genes contributes to this association. Similar to complement genes, class I genes have also been implicated in synapse elimination and they colocalize with synaptic markers and with C1q[72, 147].

In a joint model with rs13194504 and E[*C4A* expression], we identified a third genome-wide significant association within the *HLA* locus, to rs210133 ($p = 1.8 \times 10^{-8}$) as well as a strong association ($p = 1.0 \times 10^{-7}$) to rs9461856, an intronic SNP in *SYNGAP1*. SYNGAP1 is a component of the postsynaptic density and mutations in *SYNGAP1* have been found to disrupt maturation of dendritic spine synapses[148] and cause an autosomal dominant form of mental retardation[149], making it a compelling functional candidate.

We found an increased expression of *C4A* in *postmortem* brain samples from patients with schizophrenia, compared to unaffected controls. This result was statistically significant even after removing the effect of *C4A* and *C4L* copy number on *C4A* expression. A potential explanation for this result is that variation elsewhere in the genome that influences the risk of schizophrenia could affect expression of *C4A*, independently of structural variation in *C4.* However, it is also possible that the increased expression of *C4A* in patients with schizophrenia could be reactive, in response to the disease process. We also found a trend toward increased *C4A* expression in bipolar patients with psychotic features compared to those without psychosis. Although this result was only nominally significant, it is consistent with findings from genetic studies of schizophrenia and bipolar disorder that suggest a shared etiology[35].

In addition to differences in *C4A* expression between patients with schizophrenia and unaffected controls, we also found nominal support for differential expression of other complement genes, in the direction expected based on the pattern observed with *C4A* – we observed a trend toward increased expression of *C1QB* and *C2* (both of which function together with *C4* in the early steps of the classical complement pathway) and decreased expression of *CSMD1* (which promotes degradation of an activated fragment of C4) in patients with schizophrenia. While these differences in expression were nominal and could be reactive as opposed to being causally involved in the pathogenesis of schizophrenia, the latest GWAS meta-analysis provides stronger evidence for the involvement of other complement genes in schizophrenia[22]. In addition to replicating the previously identified association to *CSMD1*, this study also identified associations to the *CD46* and *CLU* loci. *CD46* encodes an inhibitory complement receptor that functions in the degradation of the activated fragment of C4[150], similar to *CSMD1.* CLU also serves as an inhibitor of the complement pathway (among its other functions) and has been associated to Alzheimer's disease[151].

The expression of *C4* in the brain increased with age, into adulthood, in both mice and humans, similar to a recent finding of an increase in C1q protein with age[152].

This pattern was observed in some, but not all, brain regions, potentially due to cellular compositions and regulatory influences on gene expression differing across the brain structures analyzed. In addition, the pattern was less apparent in humans than in mice, likely owing to variation in copy number of *C4* across individuals as well as differences in the precision with which gene expression measurements were made in the mouse and human samples. The variation in expression of *C4* with age raises the question of whether the increase in expression of *C4* into adulthood contributes to synapse elimination that occurs during adolescence[135] and to the age of onset of schizophrenia, which is typically in late adolescence to early adulthood. However, this increase in expression of *C4* around the age of onset could be a coincidental finding as it is possible that the pathogenic phenomena that predispose to developing schizophrenia precede the age of onset by several years. It also remains to be investigated whether variation in expression of *C4* could be influenced by hormonal factors, such as hormones of puberty, and whether *C4A* and *C4B* show different temporal patterns of expression.

Our results from profiling expression of *C4* and other complement genes in primary cultures of human CNS cells suggest cell-type specificity in expression of these genes, with *C1QB*, *C2*, and *C3* being expressed by microglia, and *C4* by neurons and oligodendrocytes. However, there are several caveats to consider in interpreting these results. First, the pattern of expression in these cell types may be spatiotemporally dynamic in ways that analyzing these cells *in vitro* may not reveal. Signaling between neurons and microglia, for example, could influence the expression of these genes. Indeed, the correlation in expression of *C4* with *C1QB, C2,* and *C3* in *postmortem* brain tissue is consistent with this scenario and the possibility that there are shared influences in the expression of these genes across cell types. Second, profiling expression of these genes in different cell types (e.g., excitatory and inhibitory neurons) and from cells isolated from multiple brain regions could also reveal a different pattern of expression.

Third, the extent to which genetic variation in *C4* (as well as elsewhere in the genome) affects the pattern of expression in these cell types also remains to be explored. Despite these caveats, ongoing immunohistochemistry and *in situ* hybridization experiments in human and mouse brain tissue are providing corroborating evidence for our *in vitro* results.

The finding that mouse and human *C4* contribute to synaptic refinement in mice is consistent with the role of *C1q*, *C3,* and *CR3* in this process. The similar degree of deficit in synaptic pruning in the retinogeniculate system across mice deficient in *C1q*, *C3,* or *C4* suggests that the classical (as opposed to the alternative) pathway may be the main contributor to this process, given that *C1q* and *C4* function in the classical pathway whereas *C3* also functions in the alternative pathway. The difference in *R*-value variance was significant between *C4* heterozygous mice and *C4* $^{-/-}$ mice and the difference between wild type *C4* mice and *C4* heterozygous mice trended toward significance, suggesting that the dosage of the *C4* gene may influence the degree of synapse elimination in the retinogeniculate system, although increased sample sizes will be required to investigate this possibility. The role of *C4* in synaptic pruning also offers a potential biological mechanism that may link genetic variation in *C4* to its influence on the risk of schizophrenia. This possibility remains to be investigated further by examining whether the specific series of human *C4* alleles that associate to schizophrenia vary in the extent to which they influence this phenotype in mice.

Contributions

Steve McCarroll coordinated the study and provided guidance and input on all aspects of the work. Beth Stevens and Michael Carroll contributed expertise as well as reagents for the *in vivo* experiments in mice and provided significant feedback on other aspects of the work. Allison Bialas and Jessy Presumey performed the *in vivo* experiments in mice. Samuel Rose performed ddPCR analysis of copy number in the Swedish cohort. Vanessa Van Doren performed RT-ddPCR assays of gene expression in *postmortem* human brain tissue. Heather de Rivera performed the *in vitro* assays on human CNS cells, with guidance from Allison Bialas. Matthew Baum profiled gene expression in mouse brain samples across developmental ages. SNP genotype data from the extended *HLA* locus for the 40 cohorts analyzed was obtained with approval from the Psychiatric Genomics Consortium[22].

Methods

Gene expression analysis in postmortem *human brain tissue*

We analyzed expression of *C4* and other genes of the complement pathway in the Stanley Medical Research Institute Array Collection of *postmortem* human brain samples. This collection consists of 35 individuals in each of 3 diagnostic groups (schizophrenia, bipolar disorder, and unaffected controls) and samples from 5 brain regions (anterior cingulate cortex, orbital frontal cortex, parietal cortex, cerebellum, and corpus callosum). Of the 105 individuals, 102 were of European ancestry. We measured copy number of *C4* structural features in liver and cerebellar DNA samples from these individuals using ddPCR (details of this method are described in Chapter 2). We measured gene expression using RT-ddPCR, with reaction conditions described in Chapter 3 and primer and probe sequences in Supplemental Table 1).

We normalized the gene expression measurements to the expression of a control gene (*NUDCD3, NudC domain containing 3*) to account for variation in the amount of input RNA across samples. For each individual in the collection, we calculated a composite measure of expression across brain regions analyzed. In a *i* x *j* matrix of *i* individuals and *j* brain regions, we converted the normalized expression for i^{th} individual to a percentage of the median expression value across all the samples in *the* j^{th} column.

Then, we calculated a median value across *j* columns for each row, *i*, to obtain a summary value for expression in each individual across multiple brain regions.

We tested for association between *C4A* expression and *C4A* copy number using a linear model:

$$C4A\ \text{expression}_i = \beta \times (C4A\ \text{copy number}_i) + \theta$$

where *C4A* copy number$_i$ is the diploid copy number for the i^{th} individual, β is the additive effect per copy of *C4A* on *C4A* expression, and θ is a constant (intercept). In order to evaluate whether *C4L* copy number associated to expression of *C4A* independently of *C4A* copy number (with which it is correlated), we took the residuals from the model above and fit those to *C4L* copy number:

$$\text{Residual}(C4A\ \text{expression})_i = \beta \times (C4L\ \text{copy number}_i) + \theta$$

To calculate the effect of each of the four common *C4* structures on expression of *C4A*, we fit *C4A* expression to the dosage of that structure

$$C4A\ \text{expression}_i = \beta_j \times (\text{dose}_{i,j}) + \theta$$

where dose$_{i,j}$ is the number of chromosomes in each diploid genome *i* that carry the structure *j*. To determine the *C4* structural genotype for each individual in the SMRI collection, we integrated copy number data for each *C4* structural form (*C4A*, *C4B, C4L*, and *C4S*) from ddPCR together with SNP genotypes for these samples (from the Illumina Omni 2.5 SNP microarray). For each individual, we enumerated the list of structural genotypes consistent with the set of copy numbers of *C4* structural features, based on the 15 *C4* structures that we identified in the HapMap CEU population sample in Chapter 2 (Figure 2.8). For example, if the copy number in a diploid genome of *C4A*, *C4B*, *C4L*, and *C4S* were 1, 2, 2, and 1, respectively, in an individual, the only structural genotype consistent with this set of copy numbers is *AL-BS*/*BS*. In contrast, if the copy number of *C4A*, *C4B*, *C4L*, and *C4S* were 2, 1, 2, and 1, respectively, then two structural genotypes were possible: *AL/ AL-BS* and *AL-AL* / *BS*. Given the large number of structural genotypes possible (125 possible genotypes based on 15 structural haplotypes), in many individuals, more than 5 structural genotypes were consistent with a set of copy number data for *C4* structural features. We therefore used the backbone SNP genotype data to estimate the likelihood of observing each structural genotype given a set of copy number as well as SNP genotype data. We provided a vector of genotype likelihoods (of length 125) as input for phasing in Beagle[115] (version 4). Each structural genotype that was consistent with the copy number data was encoded as equally likely and those that were inconsistent were assigned a $\log_{10}$ likelihood of -1000 (i.e., to indicate that they are extremely unlikely). We phased these likelihoods together with SNP genotypes and obtained posterior genotype probabilities for each possible structural genotype, for each individual. These probability estimates readily identified the most likely genotype for each individual (with a mean probability of 0.99).

To test association between gene expression and clinical diagnosis, we used the Mann-Whitney (non-parametric) test. We specified the alternative hypothesis based on the direction of effect of *C4* structural variation on gene expression and on the risk of schizophrenia – given that *C4* structural variants associating to increased risk of schizophrenia also associated to higher expression, we hypothesized that the expression of *C4* (as well as other genes that function in the early steps of the complement pathway, *C1QB*, *C2*, and *C3*) would be higher in patients with schizophrenia compared to unaffected controls, and that the expression of *CSMD1* (which promotes degradation of C4), would be lower in patients with schizophrenia. In order to test whether the expression of *C4A* associated with clinical diagnosis independently of structural variation in *C4*, we took the residuals from fitting expression to *C4A* and *C4L* copy number in a linear model and performed a Mann-Whitney test to assess for differences in these residuals between patients with schizophrenia and unaffected controls.

Description of case-control samples analyzed from the Psychiatric Genomics Consortium (PGC)

A detailed description of the samples analyzed from the Psychiatric Genomics Consortium (PGC) data set is available elsewhere[22] and details of the cohorts, their sample sizes, and SNP genotyping arrays used are provided in Supplemental table 2. Of the 49 cohorts from the PGC, we analyzed 40 case-control cohorts of European ancestry for which we had data available. Relatedness among samples and population structure was analyzed by the PGC Statistical Analysis Working Group, using a set of 19,551 autosomal SNPs across all cohorts, removing one member of each pair with $\pi\hat{}> 0.2$. The first twenty principal components were tested for association to phenotype and ten of these were included as covariates in all of the association analyses (as described below).

The Swedish cohort analyzed for copy number of *C4A, C4L*, and *C4S* using ddPCR includes samples from 2 of the 40 PGC cohorts listed in Supplemental table 2 (scz_swe1_eur and scz_s234_eur). All subjects in this cohort were born in Sweden, and cases were defined based on a diagnosis of schizophrenia on two separate hospitalizations. Control subjects were matched by age, sex and county of residence and did not have a hospitalization with a psychiatric diagnosis.

Quality control

The SNPs and individuals retained for association analysis were subject to the following quality control (QC) parameters by the PGC Statistical Analysis Group included[22]: (i) SNP missingness < 0.05 (before sample removal); (ii) subject missingness < 0.02; (iii) autosomal heterozygosity deviation ($|F_{het}| < 0.2$); (iv) SNP missingness < 0.02

(after sample removal); difference in SNP missingness between cases and controls < 0.02; and SNP Hardy-Weinberg equilibrium ($p > 10^{-6}$ in controls or $P > 10^{-10}$ in cases).

In addition to the above parameters that were analyzed on a genome-wide scale, we applied the following QC filters to the SNP genotype data from the extended *HLA* locus in each of the 40 cohorts analyzed. We removed three additional sets of SNPs: (i) those that were within the duplicated *C4* locus (chromosome 6:31939608-32014384, hg 19); (ii) SNPs whose allele frequency differed by more than 0.15 from their frequency in our HapMap CEU reference panel for imputation; and (iii) transversion SNPs (A/T and G/C) whose minor allele frequency was greater than 0.35 (as it can be problematic to determine whether they have the same strand assignment as SNPs in the reference panel for imputation).

Imputation of C4 *structural variation and expected* C4A *expression in the brain*

We utilized the HapMap CEU reference panel created in Chapter 2 for imputation of *C4* structural variation, using Beagle[115]. We imputed separately into each of the 40 cohorts in the PGC data set. Our reference panel consisted of 111 unrelated individuals, with *C4* structural variants on haplotypes with HapMap phase III SNPs in the extended *HLA* locus (chromosome 6: 25 - 34 Mb). The encoding of *C4* structural variation in this reference panel was based on both the *C4* structure as well as its *HLA* haplotype background (Chapter 2, Figure 2.10). *C4* structures that segregated on different *HLA* SNP haplotypes were encoded as separate alleles in the reference panel – *AL-AL* structures were divided into two alleles, *AL-AL 1* and *AL-AL 2*, based on which of the two *HLA* haplotypes they segregated on; *AL-BL* structures into three alleles that were based on the three well-defined haplotypes backgrounds and a fourth allele to represent the remaining set of poorly-defined haplotypes; and *AL-BS* structures into six alleles (five of which had well-defined haplotype backgrounds).

This strategy enabled us to independently test association to each of these imputed alleles and evaluate the extent to which the *HLA* haplotype on which *C4* structures segregated influenced their association to schizophrenia. This strategy also allowed us to (i) infer copy number of *C4* structural features (*C4A, C4B*, *C4L*, and *C4S*) based on the *C4* alleles imputed in each individual (e.g., an individual with *C4* alleles *AL-AL 1* and *AL-BL 2* has a diploid copy number of 3 for *C4A*, 1 for *C4B*, 4 for *C4L* and 0 for *C4S*); and (ii) infer expected expression of *C4A* in the brain, which we estimated as the copy number of *C4A* + copy number of *C4L* in each individual. This was based on the nearly identical effect size per copy of *C4A* and *C4L* that we calculated from regressing *C4A* expression in *postmortem* human brain tissue against copy number of *C4A* and *C4L* in a joint linear model.

We imputed *C4* structures into the Swedish cohort by making use of the copy number data generated for *C4A*, *C4L*, and *C4S* from ddPCR, together with the SNP genotype data from the *HLA* locus for each subject in the data set, as described above for the imputation of *C4* structures into the SMRI samples.

Molecular analysis of C4 *copy number in the Swedish case-control cohort*

We measured copy number of *C4A*, *C4L*, and *C4S* in 2,195 cases and 2,485 controls of Swedish ancestry using ddPCR. Reaction conditions for ddPCR assays are provided in Chapter 2 and sequences of primers and probes used are available in Supplemental table 1. Quality control parameters for inclusion of ddPCR measurements in the analysis were as described in Chapter 3.

Testing association of C4 *and SNPs to schizophrenia*

We performed a mega-analysis that utilized individual-level genotype data from all 40 cohorts that we analyzed from the PGC data set. We tested association in a logistic regression framework that included study indicator variables to account for cohort-specific effects and principal components to control for population stratification:

$$\log(\text{odds}_i) = \beta_j \text{ x } (\text{dose}_{i,j}) + \sum_{c=1}^{39} \beta_c \text{ x } (\text{cohort}_{i,c}) + \sum_{p=1}^{10} \beta_p \text{ x } (\text{PC}_{i,p}) + \theta$$

where $\text{dose}_{i,j}$ is the number of chromosomes (0, 1, or 2) in each individual, *i*, that carried a *C4* structural allele, *j* and β_j is the additive effect per copy of the *C4* allele. 39 study indicator variables (the number of cohorts minus 1) were included, with $\text{cohort}_{i,c}$ equal to 1 if the i^{th} individual belonged to the c^{th} cohort and equal to 0 otherwise. In addition, ten principal components that associated to phenotype were included as covariates, with $\text{PC}_{i,p}$ being the p^{th} principal component for the i^{th} individual. We used the same framework for testing association to (i) individual SNPs, where $\text{dose}_{i,j}$ was the dosage of the minor allele*, j*, of the SNP in individual *i*; (ii) copy number of *C4* structural features, where $\text{dose}_{i,j}$ was the diploid copy number of the *C4* feature in individual *i*; and (iii) expected expression of *C4A,* where $\text{dose}_{i,j}$ was the number of copies of *C4A* + number of copies of *C4L* in individual *i*. To test association to *C4* conditional on rs13194504, we included the dosage of that SNP as an additional covariate in the model.

We obtained an estimate of effect size for a *C4* structure (e.g.*, AL-AL*) across all alleles that contained that given structure (e.g., *AL-AL 1* and *AL-AL 2*), by performing an inverse variance meta-analysis based on the effect size and standard error associated with each *C4* allele that contained the given *C4* structure.

In vitro *assays of human CNS cells*

Primary human cortical neurons, astrocytes, and oligodendrocyte precursor cells were obtained from Sciencell Research Laboratories (catalog no. 1520, 1800, and 1600, respectively). The neurons were characterized by Sciencell to be immunopositive for MAP2, neurafilament, and beta-tubulin III; astrocytes, for GFAP; and oligodendrocyte precursor cells for A2B5 and nestin. Primary human microglia, which expressed CD45, CD14, and CD68, were obtained from Clonexpress, Inc. (catalog no. HMG 030).

Neurons were cultured on poly-L-lysine (PLL) coated plates in neuronal medium, (Sciencell, catalog no. 1521) which contained basal medium, neuronal growth supplement, and penicillin/streptomycin solution. Astrocytes were cultured on PLL-coated plates in astrocyte medium (Sciencell, catalog no. 1801), which contained basal medium, fetal bovine serum, astrocyte growth supplement, and penicillin/streptomycin solution. Oligodendrocyte precursor cells were grown on PLL-coated plates in oligodendrocyte precursor cell medium (Sciencell, catalog no. 1601-b), and then differentiated into oligodendrocytes in oligodendrocyte precursor cell differentiation medium (Sciencell, catalog no. 1631). Microglia were cultured in 50:50 DMEM: F-12 supplemented with 5% FBS and 10 ng/ml of M-CSF. RNA was extracted from these cells using the RNeasy micro kit (Qiagen) and gene expression analysis was done using RT-ddPCR.

Copy number of *C4* structural features in these cells was determined by ddPCR. The neurons, microglia, and oligodendrocyte precursor cells used in these experiments each had the following *C4* copy number profile: 2 copies *of C4A*, 2 of *C4B*, 3 of *C4L*, and 1 of *C4S*, and the astrocytes had a copy number of 2 for each of the *C4* structural features.

LPS from *Escherichia coli* 026:B6 (Sigma, catalog no. L2654-1MG) and poly(I:C) (Sigma, catalog no. P9582) were added to primary microglia after the cells had been grown in culture for 4 days. To test the effect of duration of treatment with LPS or poly(I:C) on gene expression, the reagent was added at different time points to each well and RNA was harvested at the same time so as to control for the length of time the cells were in culture.

Analysis of C4 *expression in the BrainSpan data set*

We analyzed expression of *C4* across developmental ages using the publicly available BrainSpan data set[140]. This data set includes transcriptome-wide gene expression measurements using RNA-sequencing analysis from up to sixteen brain regions across the course of pre- and postnatal human brain development. We analyzed expression of the gene annotated as *C4A* in this data set, but given that the *C4A* and *C4B* paralogs share more than 99% nucleotide identity, the measurements from RNA-sequencing analysis reported for *C4A* would have included reads originating from either paralog. In addition, because the measurement at each developmental stage was from a different individual, the expression measurements are not corrected for variation in gene copy number across samples. In order to reduce the noise in the data, we calculated median measurements of expression for developmental ages that had data from multiple samples.

Mice

C57BL/6 wild type mice were used for profiling *C4* expression in the brain across developmental ages. The generation of the *C4* $^{-/-}$ mice that were used to investigate synapse elimination in the retinogeniculate system is described in detail elsewhere[153]. In these mice, the sequence spanning part of exon 23 through exon 29 has been replaced with a PGK-Neo gene. To generate transgenic mice with human *C4,* a BAC clone (CH501-116M5) containing the human *C4AL* and *C4BL* genes was obtained from the Children's Hospital Oakland Research Institute (CHORI). Circular BAC DNA was purified from 500-ml culture by cesium chloride gradient centrifugation. BAC DNA was linearized by digestion with PI-SceI and applied to a Sepharose CL-4B column in injection buffer (0.1 M NaCl, 10 mM Tris (pH 7.5), 0.1 mM EDTA). The DNA was diluted with injection buffer at 0.5 ng/µl and then microinjected into the pronuclei of fertilized eggs harvested from FVB mice. Transgenic founder mice were screened by PCR using the *C4A*-specific primers 5′- GCTCACAGCCTTTGTGTTG -3′ and 5′- CTGCATGCTCCTGTCTAAC -3′ and the *C4B*-specific primers 5′- GCTCACAGCCTTTGTGTTG -3′ and 5′- CTGCATGCTCCTATGTATCTC -3′. We used ddPCR to determine copy number of human *C4A*, *C4B*, *C4L* and *C4S* in these mice.

Analysis of dorsal lateral geniculate nucleus (dLGN) in mice

Visualization and analysis of RGC synaptic inputs in the mouse dLGN was performed as described[141]. Cholera toxin-β subunit (CTB) conjugated to Alexa 488 (green label) and CTB conjugated to Alexa 594 (red label) were intraocularly injected into the left and right eyes, respectively, of P9 mice, which were sacrificed the following day. Images were acquired using the Zeiss Axiocam and quantified blind to experimental conditions and compared to age-matched littermate controls. The degree of left and right eye axon overlap in dLGN was quantified using an *R*-value analysis as described[143]. Pseudocolored images representing the *R*-value distribution were generated in ImageJ. P-values were calculated from a Mann-Whitney test with an alternative hypothesis of higher *R*-value variance in wild type compared to *C4* heterozygotes or *C4* $^{-/-}$; higher in *C4* heterozygotes compared to *C4* $^{-/-}$; higher in *C4* transgenic mice compared to *C4* $^{-/-}$.

Chapter 5

Conclusions and future directions

In this thesis, we have developed approaches for characterizing complex forms of genome structural variation and for evaluating their contribution to phenotypes. We have applied these approaches toward investigating the genome's strongest influences on the risk of two common diseases that have been identified from genome-wide association studies. Specifically, we

(i) unraveled the various structural forms of the *C4* locus using novel molecular and statistical approaches;

(ii) enabled analysis of *C4* structural variation by imputation, and thereby, made it possible to investigate its contribution to phenotypes in large case-control cohorts using existing SNP genotype data;

(iii) conducted the first imputation-based association analysis of complex structural variation in investigating the contribution of *C4* to SLE, and found that increased copy number of *C4A* associated to lower disease risk;

(iv) related variation in *C4* structure to gene expression in lymphoblastoid cells lines, and found that the four common structural forms of *C4* associated to *C4A* expression in proportion to the number of copies of *C4A* they contained, and associated to SLE risk in a manner that mirrored their effect on gene expression;

(v) exploited the recurrence of *C4* structures on multiple *HLA* haplotype backgrounds to develop a novel form of association testing and showed that the association of SLE to *C4* is unlikely to be a result of correlation with genetic variation elsewhere in *HLA* locus;

(vi) investigated the contribution of *C4* structural variation to risk of schizophrenia in the largest case-control study to date of this disease (totaling > 60,000 individuals), and found that an allelic series of structural variants in *C4* influences the risk of schizophrenia in a manner predicted by its effect on *C4A* expression in the brain; and

(vii) identified a biological process, synaptic pruning, to which *C4* contributes in the nervous system and could potentially underlie its association to schizophrenia.

Previous studies investigating the association of *C4* to phenotypes[60, 154] have typically tested for differences in distribution of copy number between cases and controls, and/or evaluated the effect of deviation from a reference copy number (e.g., diploid copy number of 2). In contrast, the imputation and association strategy that we developed in this thesis enabled us to exploit the rich reservoir of genetic variation in the *C4* (and surrounding *HLA*) locus to evaluate association in multiple ways, taking into account variation in diploid copy number, variation in structure of the *C4* locus on a per-chromosome level, and variation in the *HLA* haplotype on which each *C4* structure segregates.

It has been suggested that large-scale deep-sequencing studies will be required to find allelic series that help to identify genes with a causal influence on disease risk and to explore their biology[11]. Our findings indicate that the investigation of loci that segregate in several structural forms can be an additional strategy for identifying such an allelic series.

A theme of *yin* and *yang* characterizes some of the main findings from this thesis. First, the segregation of common *C4* structures on multiple *HLA* haplotypes limited the accuracy with which they could be captured by imputation. However, this very property also enabled a novel form of association testing and provided additional power to evaluate independence of effects within the *HLA* locus. Second, we found that structural variation in *C4* contributed to the risk of SLE and schizophrenia with largely opposite directions of effect – the *C4* structure associated with the highest risk of SLE was protective in schizophrenia, and vice versa. The presence of multiple structures that segregate at a common frequency raises the question of whether balancing selection involving SLE and schizophrenia (and potentially other phenotypes) could have contributed to the high degree of polymorphism at the *C4* locus.

The apparent antagonistic influence that *C4* structural variation seems to have on the risk of SLE and schizophrenia does not necessitate that potential therapeutic options that could eventually be developed based on these findings should have a zero-sum impact with respect to these phenotypes. A greater opportunity for building a therapeutic armamentarium might exist in viewing *C4* and other genes implicated in disease not just as potential drug targets themselves, but as sentinels pointing to biological processes (e.g., synaptic pruning) or cell types (e.g., microglia, a key cell type involved in synaptic pruning) that may be relevant do a disease.

Future directions

The findings from this thesis motivate several additional studies, some of which are discussed here.

First, the *HLA* locus associates to dozens of phenotypes, for most of which the gene(s) and functional allele(s) driving the association have not been identified. The reference panels and imputation strategy developed in this thesis enable the investigation of whether *C4* contributes to these phenotypes in the context of association to *HLA* SNPs and classical alleles. In particular, phenotypes associating to *HLA-DRB1*03:01* may represent strong candidates to prioritize, given the high degree of linkage disequilibrium between this allele and the *BS* structure of *C4*.

Second, the results from Chapter 2 suggest that it is possible to analyze complex forms of structural variation using imputation. However, the extent to which this applies to structurally complex loci elsewhere in the genome remains to be investigated. To this end, a map of genomic loci harboring complex and multi-allelic forms of structural variation will be fruitful. Such a map may be generated using computational algorithms for discovering and genotyping structural variants from whole-genome sequencing data[85]. This will enable investigation of how such loci vary in structure across human populations, how this variation relates to flanking SNP haplotypes, and whether it can be captured using imputation.

Third, our findings in Chapter 3 revealed that increased copy number of *C4A* is associated with a lower risk of SLE. However, complement proteins have been implicated in causing tissue damage during the inflammatory phase of the disease[102]. Therefore, while *C4* may be protective against developing SLE, it may contribute to the disease pathology after its onset. This possibility can be investigated by examining whether *C4* structural variation associates to incidence or severity of specific clinical manifestations of the disease, such as lupus nephritis, within SLE patients. The findings from such studies could have important therapeutic implications – if *C4* is protective against the overall risk of developing SLE but contributes to disease pathology after its onset, potential pharmacologic interventions in SLE patients would be aimed at reducing activity of C4, rather than potentiating it.

Fourth, our results indicate a differential contribution of *C4A* and *C4B*, which are more than 99% identical at the nucleotide level, to SLE risk. While *C4A* copy number associated to SLE, *C4B* copy number did not independently contribute to disease risk. These findings may reflect differences in biochemical properties of these two paralogs[108], and are consistent with C4A having a higher affinity for immune complexes and C4B, for carbohydrate antigens. A better understanding of the role of these paralogs could come from studies in mice. For example, *C4* $^{-/-}$ mice that are reconstituted with either the human *C4A* or *C4B* gene could be analyzed for whether the two paralogs have a differential ability to rescue a lupus-like phenotype.

Fifth, a better understanding of the functional impact of the HERV sequence in *C4* may be gained through studying transgenic mice that carry a human *C4* gene with or without the HERV insertion (in a mouse *C4* $^{-/-}$ background). These mice can be analyzed qualitatively as well as quantitatively, to investigate whether there is (i) variation in the cell types and brain structures in which *C4* is expressed; (ii) variation in the temporal pattern of *C4* expression in the brain across developmental stages; and (iii) an influence of the HERV sequence on the quantitative levels of *C4* expression (as we found in *postmortem* human brain tissue).

Sixth, our studies of the retinogeniculate system in *C4* $^{-/-}$ mice and transgenic mice with human *C4* revealed that *C4* contributes to synaptic pruning in this system. A recapitulation of the allelic series of human *C4* structural variants in mice would enable investigation of whether the specific structural alleles of human *C4* that associate to schizophrenia differ with respect to their ability to rescue the pruning phenotype in *C4* $^{-/-}$ mice. In addition, these mice can be analyzed for pathological changes seen in *postmortem* brains from patients with schizophrenia, such as thinning of the cortex and reduction in dendritic spine density[136]. Furthermore, these mice could also be subject to behavioral analyses to investigate whether structural variation in *C4* contributes to such phenotypes.

We hope that the approaches presented in this thesis will enable investigation of other structurally complex loci in the human genome and their contributions to phenotypes, and that our findings implicating *C4* in SLE and schizophrenia will motivate additional functional studies to gain biological insights into these diseases.

Appendix

Supplemental tables

Supplemental Table 1. Primer and probe sequences. All sequences are provided in the 5’ to 3’ orientation.

*, Assays were based on Wu et al.[109]

Assay	Forward Primer	Reverse Primer	Probe
Copy number of human *C4A**	CCTTTGTGTTGAAGGTCCTGAGTT	TCCTGTCTAACACTGGACAGGGGT	VIC-CCAGGAGCAGGTAGGAGGCTCGC-MGB
Copy number of human *C4B**	TGCAGGAGACATCTAACTGGCTTCT	CATGCTCCTATGTATCACTGGAGAGA	VIC-AGCAGGCTGACGGC-MGB
Copy number of human *C4L**	TTGCTCGTTCTGCTCATTCCTT	GTTGAGGCTGGTCCCCAACA	VIC-CTCCTCCAGTGGACATG-MGB
Copy number of human *C4S**	TTGCTCGTTCTGCTCATTCCTT	GGCGCAGGCTGCTGTATT	VIC-CTCCTCCAGTGGACATG-MGB
Control for copy number assays of human DNA (*RP1*)*	GACCAAATGACACAGACCTTTGG	GACTTTGGTTGGTTCCACAAGTC	FAM-AGGGACTCAGAAATCACGT
Control for copy number assays of human DNA (*RPP30*)	GATTTGGACCTGCGAGCG	GCGGCTGTCTCCACAAGT	FAM-CTGACCTGAAGGCTCT-MGB
Control for copy number assays of human DNA (Ultraconserved element)	GCTATAATAGAAGGGGGAAGTCG	ATTGTGGGCCTGTTTTGAAT	FAM-ATTGCGTCGCTCTGAGCCCC
Control for copy number assays of mouse DNA (*Rpp30*)	TGACCCTATCAGAGGACTGC	CTCTGCAATTTGTGGACACG	FAM-TGGGCTTTCTGAAAATGATGGCAA
Long-range PCR assay spanning intron 9 to exon 26 (to amplify sequences defining the *L/S* and *A/B* structural features from all *C4* genes)	TTGCTCGTTCTGCTCATTCCTT	CAATGGCTCTGCACCCTCAT	
Long-range PCR assay spanning intron 9 to exon 26 (to amplify sequences defining the *L/S* and *A/B* structural features from *C4A* genes)	TTGCTCGTTCTGCTCATTCCTT	TCCTGTCTAACACTGGACAGGGGT	
Expression of human *C4A*	CCTGAGAAACTGCAGGAGACAT	GTGAGTGCCACAGTCTCATCAT	FAM-CAGGACCCCTGTCCAGTGTTAGAC
Expression of human *C4B*	CCTGAGAAACTGCAGGAGACAT	GTGAGTGCCACAGTCTCATCAT	FAM-CTATGTATCACTGGAGAGAGGTCCTGGAAC
Expression of human *C1QB*	CTCACTCTACCCCCAACACC	TTCACTCAGCAGCATTCACC	FAM-ACCCCTTGCCCAACCAATGC
Expression of human *C2*	CTCCATCTTCTACCTCTGAATGG	AGAGACCCGGATTTACTCAGC	FAM-CACCCTTAGACCCTGTGATCCATCCT
Expression of human *C3*	GTTGGAGGGACACATCAAGG	CTCTACCCAGGCCACCTTC	FAM-TGGTGTTCCAAGCCTTGGCTCA
Expression of human *CSMD1*	GCAAGTCTGGCTTCTCCATC	ATTTTGGGGCATACCTGGAT	FAM-CCACCTCAATTGCAGCCACCTG
Expression of mouse *C4*	AGCCTGTTTCCAGCTCAAAG	GTCCTAAGGCCTCACACCTG	FAM-CCCCGGCTGCTGAACTCCAT
Control for expression assays of human RNA (*NUDCD3*)	AGTCCTGTGACCAGGTGTAGTTC	CAGGAGCAGTTCCAGAAAAATC	HEX-TCGGACAGCACCATTGTAACTGTCG
Control for expression assays of mouse RNA (*Eif4h*)	GTGCAGCTTGCTTGGTAGC	GTAAATTGCCGAGACCTTGC	VIC-AGCCTACCCCTTGGCTCGGG

Supplemental Table 2. Case-control cohorts analyzed from the Psychiatric Genomics Consortium (PGC) data set. The details presented in this table are from the latest PGC manuscript,[22] where additional information about these cohorts can be found. Cohort name indicates the name of the cohort in the PGC databases. PMID is the PubMed ID number for the initial report on that study, where available.

Cohort name	PMID	Site	Genotyping array	Number of cases	Number of controls
scz_aarh_eur	19571808	Denmark	Illumina 650K	876	871
scz_aber_eur	19571811	Aberdeen, UK	Affymetrix 6.0	719	697
scz_ajsz_eur	24253340	Israel	Illumina 1M	894	1594
scz_asrb_eur	21034186	Australia	Illumina 650K	456	287
scz_boco_eur	19571808	Bonn/Mannheim, Germany	Illumina 550K	1773	2161
scz_buls_eur		Bulgaria	Affymetrix 6.0	195	608
scz_cati_eur	18347602	US (CATIE)	Affymetrix 500K	397	203
scz_caws_eur	19571811	Cardiff, UK	Affymetrix 500K	396	284
scz_cims_eur		Boston, US (CIDAR)	Illumina OmniExpress	67	65
scz_clm2_eur	22614287	UK (CLOZUK)	Illumina 1M	3426	4085
scz_clo3_eur	22614287	UK (CLOZUK)	Illumina OmniExpress	2105	1975
scz_cou3_eur	21850710	Cardiff, UK (CogUK)	Illumina OmniExpress	530	678
scz_denm_eur	19571808	Denmark	Illumina 650K	471	456
scz_dubl_eur	19571811	Ireland	Affymetrix 6.0	264	839
scz_edin_eur	19571811	Edinburgh, UK	Affymetrix 6.0	367	284
scz_egcu_eur	15133739	Estonia (EGCUT)	Illumina OmniExpress	234	1152
scz_ersw_eur	19571808	Sweden (Hubin)	Illumina OmniExpress	265	319
scz_fi3m_eur	19571808	Finland	Illumina 317K	186	929
scz_fii6_eur		Finnish	Illumina 550K	360	1082
scz_gras_eur	20819981	Germany (GRAS)	Affymetrix Axiom	1067	1169
scz_irwt_eur	22883433	Ireland (WTCCC2)	Affymetrix 6.0	1291	1006
scz_lacw_eur	22885689	Six countries, WTCCC controls	Illumina 550K	157	245
scz_lie2_eur	11381111	NIMH CBDB	Illumina Omni 2.5M	133	269

Supplemental Table 2 (continued)

Cohort name	PMID	Site	Genotyping array	Number of cases	Number of controls
scz_lie5_eur	11381111	NIMH CBDB	Illumina 550K	497	389
scz_mgs2_eur	19571809	US, Australia (MGS)	Affymetrix 6.0	2638	2482
scz_msaf_eur	20489179	New York, US & Israel	Affymetrix 6.0	325	139
scz_munc_eur	19571808	Munich, Germany	Illumina 317K	421	312
scz_pewb_eur	23871474	Seven countries (PEIC, WTCCC2)	Illumina 1M	574	1812
scz_pews_eur	23871474	Spain (PEIC, WTCCC2)	Illumina 1M	150	236
scz_port_eur	19571811	Portugal	Affymetrix 6.0	346	215
scz_s234_eur	23974872	Sweden (sw234)	Affymetrix 6.0	1980	2274
scz_swe1_eur	23974872	Sweden (sw1)	Affymetrix 5.0	215	210
scz_swe5_eur	23974872	Sweden (sw5)	Illumina OmniExpress	1764	2581
scz_swe6_eur	23974872	Sweden (sw6)	Illumina OmniExpress	975	1145
scz_top8_eur	19571808	Norway (TOP)	Affymetrix 6.0	377	403
scz_ucla_eur	19571808	Netherlands	Illumina 550K	700	607
scz_uclo_eur	19571811	London, UK	Affymetrix 6.0	509	485
scz_umeb_eur		Umeå, Sweden	Illumina OmniExpress	341	577
scz_umes_eur		Umeå, Sweden	Illumina OmniExpress	193	704
scz_zhh1_eur	17522711	New York, US	Affymetrix 500K	190	190

References

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

2. Lawrence, J. S., Martins, C. L. & Drake, G. L. A family survey of lupus erythematosus. 1. Heritability. *J. Rheumatol.* **14**, 913-921 (1987).

3. Tsokos, G. C. Systemic lupus erythematosus. *N. Engl. J. Med.* **365**, 2110-2121 (2011).

4. Urowitz, M. B., Feletar, M., Bruce, I. N., Ibanez, D. & Gladman, D. D. Prolonged remission in systemic lupus erythematosus. *J. Rheumatol.* **32**, 1467-1472 (2005).

5. Sullivan, P. F., Kendler, K. S. & Neale, M. C. Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. *Arch. Gen. Psychiatry* **60**, 1187-1192 (2003).

6. World Health Organization. The Global Burden of Disease. 2004 Update. (2008).

7. Saha, S., Chant, D. & McGrath, J. A systematic review of mortality in schizophrenia: is the differential mortality gap worsening over time? *Arch. Gen. Psychiatry* **64**, 1123-1131 (2007).

8. Hyman, S. E. Revolution stalled. *Sci. Transl. Med.* **4**, 155cm11 (2012).

9. Picchioni, M. M. & Murray, R. M. Schizophrenia. *BMJ* **335**, 91-95 (2007).

10. Kola, I. & Landis, J. Can the pharmaceutical industry reduce attrition rates? *Nat. Rev. Drug Discov.* **3**, 711-715 (2004).

11. Plenge, R. M., Scolnick, E. M. & Altshuler, D. Validating therapeutic targets through human genetics. *Nat. Rev. Drug Discov.* **12**, 581-594 (2013).

12. Hirschhorn, J. N. Genomewide association studies--illuminating biologic pathways. *N. Engl. J. Med.* **360**, 1699-1701 (2009).

13. Klein, R. J. *et al*. Complement factor H polymorphism in age-related macular degeneration. *Science* **308**, 385-389 (2005).

14. Rioux, J. D. *et al*. Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. *Nat. Genet.* **39**, 596-604 (2007).

15. Abifadel, M. *et al*. Mutations in PCSK9 cause autosomal dominant hypercholesterolemia. *Nat. Genet.* **34**, 154-156 (2003).

16. Cohen, J. *et al*. Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. *Nat. Genet.* **37**, 161-165 (2005).

17. Kotowski, I. K. *et al*. A spectrum of PCSK9 alleles contributes to plasma levels of low-density lipoprotein cholesterol. *Am. J. Hum. Genet.* **78**, 410-422 (2006).

18. Stein, E. A. *et al*. Effect of a monoclonal antibody to PCSK9 on LDL cholesterol. *N. Engl. J. Med.* **366**, 1108-1118 (2012).

19. Yang, Y. *et al*. Mutations in SCN9A, encoding a sodium channel alpha subunit, in patients with primary erythermalgia. *J. Med. Genet.* **41**, 171-174 (2004).

20. Cox, J. J. *et al*. An SCN9A channelopathy causes congenital inability to experience pain. *Nature* **444**, 894-898 (2006).

21. Schmalhofer, W. A. *et al*. ProTx-II, a selective inhibitor of NaV1.7 sodium channels, blocks action potential propagation in nociceptors. *Mol. Pharmacol.* **74**, 1476-1484 (2008).

22. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, forthcoming (2014).

23. Keebler, M. E. *et al*. Association of blood lipids with common DNA sequence variants at 19 genetic loci in the multiethnic United States National Health and Nutrition Examination Survey III. *Circ. Cardiovasc. Genet.* **2**, 238-243 (2009).

24. International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299-1320 (2005).

25. Stranger, B. E., Stahl, E. A. & Raj, T. Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics* **187**, 367-383 (2011).

26. Guerra, S. G., Vyse, T. J. & Cunninghame Graham, D. S. The genetics of lupus: a functional perspective. *Arthritis Res. Ther.* **14**, 211 (2012).

27. Han, S. *et al*. Evaluation of imputation-based association in and around the integrin-alpha-M (ITGAM) gene and replication of robust association between a non-synonymous functional variant within ITGAM and systemic lupus erythematosus (SLE). *Hum. Mol. Genet.* **18**, 1171-1180 (2009).

28. MacPherson, M., Lek, H. S., Prescott, A. & Fagerholm, S. C. A systemic lupus erythematosus-associated R77H substitution in the CD11b chain of the Mac-1 integrin compromises leukocyte adhesion and phagocytosis. *J. Biol. Chem.* **286**, 17303-17310 (2011).

29. Jostins, L. *et al*. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119-124 (2012).

30. Li, Y. *et al*. A genome-wide association study in Han Chinese identifies a susceptibility locus for primary Sjogren's syndrome at 7q11.23. *Nat. Genet.* **45**, 1361-1365 (2013).

31. Orozco, G. *et al*. Novel rheumatoid arthritis susceptibility locus at 22q12 identified in an extended UK genome-wide association study. *Arthritis Rheumatol.* **66**, 24-30 (2014).

32. Dubois, P. C. *et al*. Multiple common variants for celiac disease influencing immune gene expression. *Nat. Genet.* **42**, 295-302 (2010).

33. Nair, R. P. *et al*. Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways. *Nat. Genet.* **41**, 199-204 (2009).

34. Ramos, P. S. *et al*. A comprehensive analysis of shared loci between systemic lupus erythematosus (SLE) and sixteen autoimmune diseases reveals limited genetic overlap. *PLoS Genet.* **7**, e1002406 (2011).

35. International Schizophrenia Consortium *et al*. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748-752 (2009).

36. Purcell, S. M. *et al*. A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* **506**, 185-190 (2014).

37. International Schizophrenia Consortium. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* **455**, 237-241 (2008).

38. Cross-Disorder Group of the Psychiatric Genomics Consortium *et al*. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat. Genet.* **45**, 984-994 (2013).

39. Cross-Disorder Group of the Psychiatric Genomics Consortium & Genetic Risk Outcome of Psychosis (GROUP) Consortium. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet* **381**, 1371-1379 (2013).

40. McCarthy, S. E. *et al*. Microduplications of 16p11.2 are associated with schizophrenia. *Nat. Genet.* **41**, 1223-1227 (2009).

41. Autism Genome Project Consortium *et al*. Mapping autism risk loci using genetic linkage and chromosomal rearrangements. *Nat. Genet.* **39**, 319-328 (2007).

42. Kirov, G. *et al*. De novo CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia. *Mol. Psychiatry* **17**, 142-153 (2012).

43. Gateva, V. *et al*. A large-scale replication study identifies TNIP1, PRDM1, JAZF1, UHRF1BP1 and IL10 as risk loci for systemic lupus erythematosus. *Nat. Genet.* **41**, 1228-1233 (2009).

44. International Consortium for Systemic Lupus Erythematosus Genetics (SLEGEN) *et al*. Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in ITGAM, PXK, KIAA1542 and other loci. *Nat. Genet.* **40**, 204-210 (2008).

45. Yang, W. *et al*. Meta-analysis followed by replication identifies loci in or near CDKN1B, TET3, CD80, DRAM1, and ARID5B as associated with systemic lupus erythematosus in Asians. *Am. J. Hum. Genet.* **92**, 41-51 (2013).

46. Hom, G. *et al*. Association of systemic lupus erythematosus with C8orf13-BLK and ITGAM-ITGAX. *N. Engl. J. Med.* **358**, 900-909 (2008).

47. Shi, J. *et al*. Common variants on chromosome 6p22.1 are associated with schizophrenia. *Nature* **460**, 753-757 (2009).

48. Ripke, S. *et al*. Genome-wide association study identifies five new schizophrenia loci. *Nat. Genet.* **43**, 969-976 (2011).

49. Ripke, S. *et al*. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat. Genet.* **45**, 1150-1159 (2013).

50. Horton, R. *et al*. Gene map of the extended human MHC. *Nat. Rev. Genet.* **5**, 889-899 (2004).

51. de Bakker, P. I. *et al*. A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nat. Genet.* **38**, 1166-1172 (2006).

52. Traherne, J. A. *et al*. Genetic analysis of completely sequenced disease-associated MHC haplotypes identifies shuffling of segments in recent human history. *PLoS Genet.* **2**, e9 (2006).

53. Horton, R. *et al*. Variation analysis and gene annotation of eight MHC haplotypes: the MHC Haplotype Project. *Immunogenetics* **60**, 1-18 (2008).

54. Morris, D. L. *et al*. Unraveling multiple MHC gene associations with systemic lupus erythematosus: model choice indicates a role for HLA alleles and non-HLA genes in Europeans. *Am. J. Hum. Genet.* **91**, 778-793 (2012).

55. Fernando, M. M. *et al*. Identification of two independent risk factors for lupus within the MHC in United Kingdom families. *PLoS Genet.* **3**, e192 (2007).

56. Graham, R. R. *et al*. Specific combinations of HLA-DR2 and DR3 class II haplotypes contribute graded risk for disease susceptibility and autoantibodies in human SLE. *Eur. J. Hum. Genet.* **15**, 823-830 (2007).

57. Wu, Y. L., Hauptmann, G., Viguier, M. & Yu, C. Y. Molecular basis of complete complement C4 deficiency in two North-African families with systemic lupus erythematosus. *Genes Immun.* **10**, 433-445 (2009).

58. Wu, Y. L. *et al*. Phenotypes, genotypes and disease susceptibility associated with gene copy number variations: complement C4 CNVs in European American healthy subjects and those with systemic lupus erythematosus. *Cytogenet. Genome Res.* **123**, 131-141 (2008).

59. Yang, Y. *et al*. Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European Americans. *Am. J. Hum. Genet.* **80**, 1037-1054 (2007).

60. Boteva, L. *et al*. Genetically determined partial complement C4 deficiency states are not independent risk factors for SLE in UK and Spanish populations. *Am. J. Hum. Genet.* **90**, 445-456 (2012).

61. Stefansson, H. *et al*. Common variants conferring risk of schizophrenia. *Nature* **460**, 744-747 (2009).

62. Cazzullo, C. L., Smeraldi, E. & Penati, G. The leucocyte antigenic system HL-A as a possible genetic marker of schizophrenia. *Br. J. Psychiatry* **125**, 25-27 (1974).

63. Wright, P., Nimgaonkar, V. L., Donaldson, P. T. & Murray, R. M. Schizophrenia and HLA: a review. *Schizophr. Res.* **47**, 1-12 (2001).

64. Corvin, A. & Morris, D. W. Genome-wide association studies: findings at the major histocompatibility complex locus in psychosis. *Biol. Psychiatry* **75**, 276-283 (2014).

65. Mortensen, P. B. *et al*. Effects of family history and place and season of birth on the risk of schizophrenia. *N. Engl. J. Med.* **340**, 603-608 (1999).

66. Khandaker, G. M., Zimbron, J., Lewis, G. & Jones, P. B. Prenatal maternal infection, neurodevelopment and adult schizophrenia: a systematic review of population-based studies. *Psychol. Med.* **43**, 239-257 (2013).

67. Eaton, W. W. *et al*. Association of schizophrenia and autoimmune diseases: linkage of Danish national registers. *Am. J. Psychiatry* **163**, 521-528 (2006).

68. Mukherjee, S. *et al*. Serum antibodies to nicotinic acetylcholine receptors in schizophrenic patients. *Schizophr. Res.* **12**, 131-136 (1994).

69. Borda, T., Gomez, R., Berria, M. I. & Sterin-Borda, L. Antibodies against astrocyte M1 and M2 muscarinic cholinoceptor from schizophrenic patients' sera. *Glia* **45**, 144-154 (2004).

70. Miller, B. J., Buckley, P., Seabolt, W., Mellor, A. & Kirkpatrick, B. Meta-analysis of cytokine alterations in schizophrenia: clinical status and antipsychotic effects. *Biol. Psychiatry* **70**, 663-671 (2011).

71. Goddard, C. A., Butts, D. A. & Shatz, C. J. Regulation of CNS synapses by neuronal MHC class I. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 6828-6833 (2007).

72. Datwani, A. *et al*. Classical MHCI molecules regulate retinogeniculate refinement and limit ocular dominance plasticity. *Neuron* **64**, 463-470 (2009).

73. Fourgeaud, L. *et al*. MHC class I modulates NMDA receptor function and AMPA receptor trafficking. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 22278-22283 (2010).

74. McAllister, A. K. Major histocompatibility complex I in brain development and schizophrenia. *Biol. Psychiatry* **75**, 262-268 (2014).

75. Rudduck, C., Beckman, L., Franzen, G., Jacobsson, L. & Lindstrom, L. Complement factor C4 in schizophrenia. *Hum. Hered.* **35**, 223-226 (1985).

76. Schroers, R. *et al*. Investigation of complement C4B deficiency in schizophrenia. *Hum. Hered.* **47**, 279-282 (1997).

77. Mayilyan, K. R., Dodds, A. W., Boyajyan, A. S., Soghoyan, A. F. & Sim, R. B. Complement C4B protein in schizophrenia. *World J. Biol. Psychiatry.* **9**, 225-230 (2008).

78. Iafrate, A. J. *et al*. Detection of large-scale variation in the human genome. *Nat. Genet.* **36**, 949-951 (2004).

79. Sebat, J. *et al*. Large-scale copy number polymorphism in the human genome. *Science* **305**, 525-528 (2004).

80. Conrad, D. F. *et al*. Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704-712 (2010).

81. Locke, D. P. *et al*. Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. *Am. J. Hum. Genet.* **79**, 275-290 (2006).

82. McCarroll, S. A. *et al*. Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nat. Genet.* **40**, 1107-1112 (2008).

83. Willer, C. J. *et al*. Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat. Genet.* **41**, 25-34 (2009).

84. McCarroll, S. A. *et al*. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat. Genet.* **40**, 1166-1174 (2008).

85. Handsaker, R. E., Korn, J. M., Nemesh, J. & McCarroll, S. A. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat. Genet.* **43**, 269-276 (2011).

86. Mills, R. E. *et al*. Mapping copy number variation by population-scale genome sequencing. *Nature* **470**, 59-65 (2011).

87. McCarroll, S. A. *et al*. Common deletion polymorphisms in the human genome. *Nat. Genet.* **38**, 86-92 (2006).

88. McCarroll, S. A. Copy-number analysis goes more than skin deep. *Nat. Genet.* **40**, 5-6 (2008).

89. Carpenter, D., Walker, S., Prescott, N., Schalkwijk, J. & Armour, J. A. Accuracy and differential bias in copy number measurement of CCL3L1 in association studies with three auto-immune disorders. *BMC Genomics* **12**, 418-2164-12-418 (2011).

90. Field, S. F. *et al*. Experimental aspects of copy number variant assays at CCL3L1. *Nat. Med.* **15**, 1115-1117 (2009).

91. He, W. *et al*. Reply to: "Experimental aspects of copy number variant assays at CCL3L1". *Nat. Med.* **15**, 1117-1120 (2009).

92. Zhao, A. *et al*. Lack of support for association between the copy number variants in the FCGR locus and schizophrenia: A case control study. *Neurosci. Lett.* (2012).

93. Rodriguez-Santiago, B. *et al*. Association of common copy number variants at the glutathione S-transferase genes and rare novel genomic changes with schizophrenia. *Mol. Psychiatry* **15**, 1023-1033 (2010).

94. Sudmant, P. H. *et al*. Diversity of human copy number variation and multicopy genes. *Science* **330**, 641-646 (2010).

95. Conrad, D. F. *et al*. Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704-712 (2010).

96. Abbas, A. & Lichtman, A. in *Basic immunology: functions and disorders of the immune system* (Elsevier Health Sciences, 2008).

97. Carroll, M. C. The role of complement and complement receptors in induction and regulation of immunity. *Annu. Rev. Immunol.* **16**, 545-568 (1998).

98. Kraus, D. M. *et al*. CSMD1 is a novel multiple domain complement-regulatory protein highly expressed in the central nervous system and epithelial tissues. *J. Immunol.* **176**, 4419-4430 (2006).

99. Escudero-Esparza, A., Kalchishkova, N., Kurbasic, E., Jiang, W. G. & Blom, A. M. The novel complement inhibitor human CUB and Sushi multiple domains 1 (CSMD1) protein promotes factor I-mediated degradation of C4b and C3b and inhibits the membrane attack complex assembly. *FASEB J.* **27**, 5083-5093 (2013).

100. Scharfstein, J., Ferreira, A., Gigli, I. & Nussenzweig, V. Human C4-binding protein. I. Isolation and characterization. *J. Exp. Med.* **148**, 207-222 (1978).

101. Manderson, A. P., Botto, M. & Walport, M. J. The role of complement in the development of systemic lupus erythematosus. *Annu. Rev. Immunol.* **22**, 431-456 (2004).

102. Walport, M. J. Complement and systemic lupus erythematosus. *Arthritis Res.* **4 Suppl 3**, S279-93 (2002).

103. Carroll, M. C. The role of complement in B cell activation and tolerance. *Adv. Immunol.* **74**, 61-88 (2000).

104. Stevens, B. *et al*. The classical complement cascade mediates CNS synapse elimination. *Cell* **131**, 1164-1178 (2007).

105. Schafer, D. P. *et al*. Microglia sculpt postnatal neural circuits in an activity and complement-dependent manner. *Neuron* **74**, 691-705 (2012).

106. McDonald, J. F. & Nelsestuen, G. L. Potent inhibition of terminal complement assembly by clusterin: characterization of its impact on C9 polymerization. *Biochemistry* **36**, 7464-7473 (1997).

107. Dangel, A. W. *et al*. The dichotomous size variation of human complement C4 genes is mediated by a novel family of endogenous retroviruses, which also establishes species-specific genomic patterns among Old World primates. *Immunogenetics* **40**, 425-436 (1994).

108. Law, S. K., Dodds, A. W. & Porter, R. R. A comparison of the properties of two classes, C4A and C4B, of the human complement component C4. *EMBO J.* **3**, 1819-1823 (1984).

109. Wu, Y. L. *et al*. Sensitive and specific real-time polymerase chain reaction assays to accurately determine copy number variations (CNVs) of human complement C4A, C4B, C4-long, C4-short, and RCCX modules: elucidation of C4 CNVs in 50 consanguineous subjects with defined HLA genotypes. *J. Immunol.* **179**, 3012-3025 (2007).

110. Fernando, M. M. *et al*. Assessment of complement C4 gene copy number using the paralog ratio test. *Hum. Mutat.* **31**, 866-874 (2010).

111. Hindson, B. J. *et al*. High-throughput droplet digital PCR system for absolute quantitation of DNA copy number. *Anal. Chem.* **83**, 8604-8610 (2011).

112. 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).

113. Browning, B. L. & Browning, S. R. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* **84**, 210-223 (2009).

114. Jia, X. *et al*. Imputing amino acid polymorphisms in human leukocyte antigens. *PLoS One* **8**, e64683 (2013).

115. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**, 1084-1097 (2007).

116. Banlaki, Z., Doleschall, M., Rajczy, K., Fust, G. & Szilagyi, A. Fine-tuned characterization of RCCX copy number variants and their relationship with extended MHC haplotypes. *Genes Immun.* **13**, 530-535 (2012).

117. Browning, B. L. & Browning, S. R. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* **84**, 210-223 (2009).

118. Hauptmann, G., Grosshans, E. & Heid, E. Lupus erythematosus syndrome and complete deficiency of the fourth component of complement. *Boll. Ist. Sieroter. Milan.* **53**, suppl:228 (1974).

119. Fielder, A. H. *et al*. Family study of the major histocompatibility complex in patients with systemic lupus erythematosus: importance of null alleles of C4A and C4B in determining disease susceptibility. *Br. Med. J. (Clin. Res. Ed)* **286**, 425-428 (1983).

120. Moulds, J. M., Warner, N. B. & Arnett, F. C. Complement component C4A and C4B levels in systemic lupus erythematosus: quantitation in relation to C4 null status and disease activity. *J. Rheumatol.* **20**, 443-447 (1993).

121. Yang, Y. *et al*. The intricate role of complement component C4 in human systemic lupus erythematosus. *Curr. Dir. Autoimmun.* **7**, 98-132 (2004).

122. Slingsby, J. H. *et al*. Homozygous hereditary C1q deficiency and systemic lupus erythematosus. A new family and the molecular basis of C1q deficiency in three families. *Arthritis Rheum.* **39**, 663-670 (1996).

123. Morgan, B. P. & Walport, M. J. Complement deficiency and disease. *Immunol. Today* **12**, 301-306 (1991).

124. Einav, S., Pozdnyakova, O. O., Ma, M. & Carroll, M. C. Complement C4 is protective for lupus disease independent of C3. *J. Immunol.* **168**, 1036-1041 (2002).

125. Nauta, A. J., Daha, M. R., van Kooten, C. & Roos, A. Recognition and clearance of apoptotic cells: a role for complement and pentraxins. *Trends Immunol.* **24**, 148-154 (2003).

126. Munoz, L. E., Lauber, K., Schiller, M., Manfredi, A. A. & Herrmann, M. The role of defective clearance of apoptotic cells in systemic autoimmunity. *Nat. Rev. Rheumatol.* **6**, 280-289 (2010).

127. Suber, T. & Rosen, A. Apoptotic cell blebs: repositories of autoantigens and contributors to immune context. *Arthritis Rheum.* **60**, 2216-2219 (2009).

128. Munoz, L. E. *et al*. Remnants of secondarily necrotic cells fuel inflammation in systemic lupus erythematosus. *Arthritis Rheum.* **60**, 1733-1742 (2009).

129. Merrill, J. *et al*. Assessment of flares in lupus patients enrolled in a phase II/III study of rituximab (EXPLORER). *Lupus* **20**, 709-716 (2011).

130. Kalunian, K. C. *et al*. Measurement of cell-bound complement activation products enhances diagnostic performance in systemic lupus erythematosus. *Arthritis Rheum.* **64**, 4040-4047 (2012).

131. Rovin, B. H. *et al*. Efficacy and safety of rituximab in patients with active proliferative lupus nephritis: the Lupus Nephritis Assessment with Rituximab study. *Arthritis Rheum.* **64**, 1215-1226 (2012).

132. Sham, P. C. & Purcell, S. M. Statistical power and significance testing in large-scale genetic studies. *Nat. Rev. Genet.* **15**, 335-346 (2014).

133. Purcell, S. *et al*. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559-575 (2007).

134. DeLisi, L. E. The significance of age of onset for schizophrenia. *Schizophr. Bull.* **18**, 209-215 (1992).

135. Feinberg, I. Schizophrenia: caused by a fault in programmed synaptic elimination during adolescence? *J. Psychiatr. Res.* **17**, 319-334 (1982).

136. Glausier, J. R. & Lewis, D. A. Dendritic spine pathology in schizophrenia. *Neuroscience* **251**, 90-107 (2013).

137. Fromer, M. *et al*. De novo mutations in schizophrenia implicate synaptic networks. *Nature* **506**, 179-184 (2014).

138. Benros, M. E., Eaton, W. W. & Mortensen, P. B. The epidemiologic evidence linking autoimmune diseases and psychosis. *Biol. Psychiatry* **75**, 300-306 (2014).

139. Raychaudhuri, S. *et al*. Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. *Nat. Genet.* **44**, 291-296 (2012).

140. BrainSpan: Atlas of the Developing Human Brain [Internet]. Funded by ARRA Awards 1RC2MH089921-01, 1RC2MH090047-01, and 1RC2MH089929-01. © 2011. Available from: http://developinghumanbrain.org.

141. Bialas, A. R. & Stevens, B. TGF-beta signaling regulates neuronal C1q expression and developmental synaptic refinement. *Nat. Neurosci.* **16**, 1773-1782 (2013).

142. Shatz, C. J. & Kirkwood, P. A. Prenatal development of functional connections in the cat's retinogeniculate pathway. *J. Neurosci.* **4**, 1378-1397 (1984).

143. Torborg, C. L. & Feller, M. B. Unbiased analysis of bulk axonal segregation patterns. *J. Neurosci. Methods* **135**, 17-26 (2004).

144. Suntsova, M. *et al*. Human-specific endogenous retroviral insert serves as an enhancer for the schizophrenia-linked gene PRODH. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 19472-19477 (2013).

145. Debnath, M., Cannon, D. M. & Venkatasubramanian, G. Variation in the major histocompatibility complex [MHC] gene family in schizophrenia: associations and functional implications. *Prog. Neuropsychopharmacol. Biol. Psychiatry* **42**, 49-62 (2013).

146. Leboyer, M., Tamouza, R., Charron, D., Faucard, R. & Perron, H. Human endogenous retrovirus type W (HERV-W) in schizophrenia: a new avenue of research at the gene-environment interface. *World J. Biol. Psychiatry.* **14**, 80-90 (2013).

147. Lee, H. *et al*. Synapse elimination and learning rules co-regulated by MHC class I H2-Db. *Nature* **509**, 195-200 (2014).

148. Clement, J. P. *et al*. Pathogenic SYNGAP1 mutations impair cognitive development by disrupting maturation of dendritic spine synapses. *Cell* **151**, 709-723 (2012).

149. Hamdan, F. F. *et al*. Mutations in SYNGAP1 in autosomal nonsyndromic mental retardation. *N. Engl. J. Med.* **360**, 599-605 (2009).

150. Barilla-LaBarca, M. L., Liszewski, M. K., Lambris, J. D., Hourcade, D. & Atkinson, J. P. Role of membrane cofactor protein (CD46) in regulation of C4b and C3b deposited on cells. *J. Immunol.* **168**, 6298-6304 (2002).

151. Lambert, J. C. *et al*. Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. *Nat. Genet.* **41**, 1094-1099 (2009).

152. Stephan, A. H. *et al*. A dramatic increase of C1q protein in the CNS during normal aging. *J. Neurosci.* **33**, 13460-13474 (2013).

153. Fischer, M. B. *et al*. Regulation of the B cell response to T-dependent antigens by classical pathway complement. *J. Immunol.* **157**, 549-556 (1996).

154. Liu, Y. H. *et al*. Association between copy number variation of complement component C4 and Graves' disease. *J. Biomed. Sci.* **18**, 71-0127-18-71 (2011).